Preface

In 1991 two of us, Luc Massart and Bernard Vandeginste, discussed, during one of our many meetings, the possibility and necessity of updating the book *Chemometrics: a textbook*. Some of the newer techniques, such as partial least squares and expert systems, were not included in that book which was written some 15 years ago. Initially, we thought that we could bring it up to date with relatively minor revision. We could not have been more wrong. Even during the planning of the book we witnessed a rapid development in the application of natural computing methods, multivariate calibration, method validation, etc.

When approaching colleagues to join the team of authors, it was clear from the outset that the book would not be an overhaul of the previous one, but an almost completely new book. When forming the team, we were particularly happy to be joined by two industrial chemometricians, Dr. Paul Lewi from Janssen Pharmaceutica and Dr. Sijmen de Jong from Unilever Research Laboratorium Vlaardingen, each having a wealth of practical experience. We are grateful to Janssen Pharmaceutica and Unilever Research Vlaardingen that they allowed Paul, Sijmen and Bernard to spend some of their time on this project. The three other authors belong to the Vrije Universiteit Brussel (Prof. An Smeyers-Verbeke and Prof. D. Luc Massart) and the Katholieke Universiteit Nijmegen (Professor Lutgarde Buydens), thus creating a team in which university and industry are equally well represented. We hope that this has led to an equally good mix of theory and application in the new book.

Much of the material presented in this book is based on the direct experience of the authors. This would not have been possible without the hard work and input of our colleagues, students and post-doctoral fellows. We sincerely want to acknowledge each of them for their good research and contributions without which we would not have been able to treat such a broad range of subjects. Some of them read chapters or helped in other ways. We also owe thanks to the chemometrics community and at the same time we have to offer apologies. We have had the opportunity of collaborating with many colleagues and we have profited from the research and publications of many others. Their ideas and work have made this book possible and necessary. The size of the book shows that they have been very productive. Even so, we have cited only a fraction of the literature and we have not included the more sophisticated work. Our wish was to consolidate and therefore to explain those methods that have become more or less accepted, also to newcomers to chemometrics. Our apologies, therefore, to those we did not cite or not extensively: it is not a reflection on the quality of their work.

Each chapter saw many versions which needed to be entered and re-entered in the computer. Without the help of our secretaries, we would not have been able to complete this work successfully. All versions were read and commented on by all authors in a long series of team meetings. We will certainly retain special memories of many of our two-day meetings, for instance the one organized by Paul in the famous abbey of the regular canons of Prémontré at Tongerlo, where we could work in peace and quiet as so many before us have done.

Much of this work also had to be done at home, which took away precious time from our families. Their love, understanding, patience and support was indispensable for us to carry on with the seemingly endless series of chapters to be drafted, read or revised.

September 1997

Chapter 1

Introduction

1.1 The aims of chemometrics

1.1.1 Chemometrics and the "arch of knowledge"

Scientific methodology follows a two-fold pathway for the establishment of knowledge. As explained by Oldroyd [1], these pathways lead through an examination of observable phenomena to general rational "first principles" (analysis); and from such "first principles" back again to observables, which are thereby explained in terms of the principles from which they are held to be deducible (synthesis). The shape of this methodological project led Oldroyd to the concept of the "arch of knowledge". *Chemometrics* conforms to this general pattern. This will become apparent from its definition. For the purposes of this book, we define chemometrics as follows: "Chemometrics is a chemical discipline that uses mathematics, statistics and formal logic (a) to design or select optimal experimental procedures; (b) to provide maximum relevant chemical information by analyzing chemical data; and (c) to obtain knowledge about chemical systems".

This definition is derived and adapted from the one given in an earlier version of this book [2].

Starting with a certain chemical knowledge (the "first principles") (Fig. 1.1), chemists define a hypothesis. To be able to test this hypothesis and thereby verify its validity, they need experimental data (the "observables"). They therefore first decide which experiments to carry out (point (a) of the definition). The chemometrician's approach will be to do this with the help of mathematical and statistical techniques, such as the use of experimental design methodology. The experiments generate data and the chemometrician uses them to extract information (point (b) of the definition), for instance to derive a model by computing a regression equation that describes how the result of the measurement (the response) is related to the experimental variables. A chemist can use this information and chemical intelligence to generate more knowledge about the system (point (c) of the definition). If, for example, the chemical domain investigated is the study of chemical reactions, the chemist may conclude that the reaction kinetics are second order (analysis). With the increased knowledge about the system the chemist can formulate

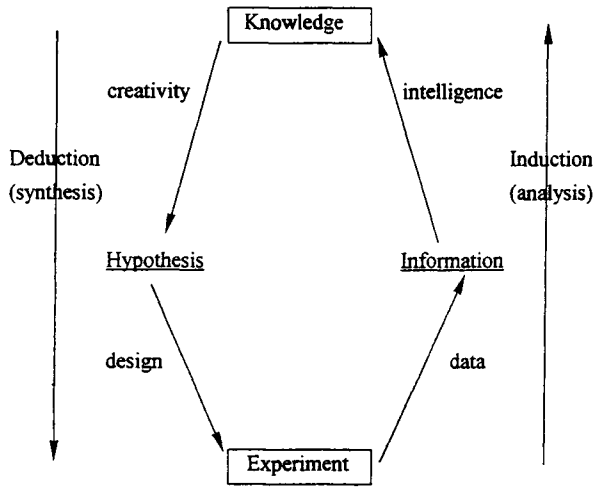


Fig. 1.1. The arch of knowledge (adapted from Ref. [1]).

an experimental design (synthesis) to obtain still more data and therefore more information and insight and eventually build a solid arch of knowledge.

The chemometrics relating to points (a) and (b) have hitherto been more extensively investigated than those relating to (c). Many of the 44 chapters of this book will refer to (a) and (b), and some also to (c). However, most chapters stress one of the three points to a greater extent than the others. In Sections 1.2.1 to 1.2.3 we will give for each a short overview of the main subjects to be discussed.

1.1.2 Chemometrics and quality

Chemometrics is not always involved in obtaining new knowledge and this is particularly so in industrial applications. Chemometrics is involved in the process of producing data and in the extraction of the information from these data. If the quality of the measurement processes and therefore the quality of the data is not good enough, the information may be uncertain or even wrong.

Quality is an essential preoccupation of chemometrics and this is also the case for industry. It is, therefore, not surprising that chemometrics has been recognized in recent years as an important subject. Indeed, many of the techniques that chemometricians apply to obtain better measurement processes are also used to obtain better processes in general or better products. The measurement processes themselves often have the aim of assisting the development of better products or of controlling processes.

Very often, therefore, the ultimate aim of chemometrics is to improve or optimize or to monitor and control the quality of a product or process. Several chapters are devoted to such domains or to quality aspects. An overview is given in Section 1.2.4.

1.2 An overview of chemometrics

1.2.1 Experiments and experimental design

Whenever experimentation is considered, one should first decide which experiments should be carried out (point (a) of the definition). This is discussed in many chapters. For instance, in Chapter 4 one of the questions is how many experiments must be carried out to be able to accept or reject a hypothesis with sufficient confidence that the decision is correct; in Chapter 5 we explain that to compare two means one can opt for a paired or an unpaired design; and in Chapter 6 we describe when an analysis of variance should be carried out according to a crossed or a hierarchical design. The chapters that discuss regression for calibration or modelling purposes such as Chapter 8 (linear regression), Chapter 10 (multivariate regression) and Chapter 36 (multivariate calibration) insist on the importance of the selection of the calibration design to obtain the best-fitting models or the best predictions.

The design of experiments is the more important element in Chapters 21 to 27. These chapters describe how to design experiments to decide in a cost-effective way which variables are important for the quality of a product or a process and then how to find the optimal combination of variables, i.e. the one that yields the best result. An overview is given in Table 1.1.

TABLE 1.1

Brief contents of the chapters on experimental design

Chapter 21:	General introduction into experimental design.
Chapter 22:	Two-level factorial designs to decide which variables are important and which variables interact, and to describe the effects of variables on responses with first-order models.
Chapter 23:	Two-level fractional factorial designs to achieve the aims of Chapter 22 but with fewer experiments and with the lowest loss of information possible.
Chapter 24:	More than two-level designs to obtain second- (or higher-) order models for the responses in function of the variables affecting the process (response surfaces) and to obtain optimum responses for these process variables.
Chapter 25:	Mixture designs to model mixture variables and to optimise mixtures.
Chapter 26:	Sequential approaches to optimization, selection of evaluation criteria, including Taguchi designs.
Chapter 27:	Numerical optimization through the use of genetic algorithms and related techniques.

1.2.2 Extraction of information from data

1.2.2.1 Displaying data

The extraction of information from data or *data analysis* (point (b) of the definition) usually starts by *describing* or *displaying* the experimental data obtained. In general, these data constitute a data table or tables. Let us first consider

the situation where a single table is obtained (Figs. 1.2a–d). This consists of columns (and in the simplest case, one single column) each giving the value of one variable for a set of objects. Very often these objects are samples of a population of objects and the reason for making the measurements is to infer from the results obtained some characteristic of the population. An example is the estimation of the mean and standard deviation (Chapters 2 and 3). Often we want to know whether the data follow a certain distribution, such as the normal distribution (Chapter 3) or other distributions (Chapter 15), or whether outliers are present (Chapter 5).

It is very important to look at the data whenever possible. To be able to do this we need methods to display them. Histograms (Chapter 2) and box plots (Chapter 12) are among the methods that best allow visual evaluation for univariate data (i.e., data for a single variable x; Fig. 1.2a). Plots for special purposes, such as normal distribution plots, that allow us to evaluate visually whether a data set is normally distributed are also available (Chapter 3).

In many cases the objects are described by many variables (multivariate data; Fig 1.2b). To plot them we need as many dimensions as there are variables. Of

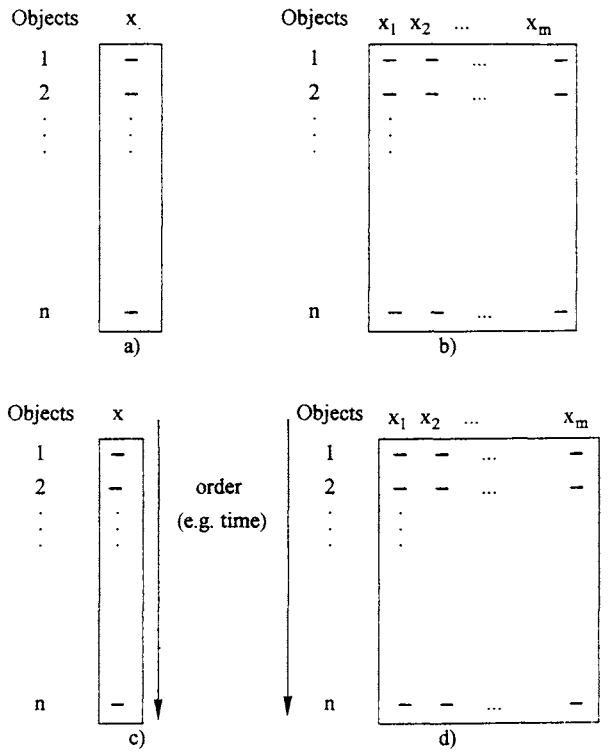


Fig. 1.2. Data structures for display: (a) univariate; (b) multivariate; (c) univariate, with objects subjected to some kind of order, e.g. ordered in time; (d) multivariate, with ordered objects.

course, as soon as there are more than three dimensions a straightforward plot of the data is no longer possible so that displaying them for visual evaluation becomes more difficult. Chemometrics offers methods that allow the display of such multi-dimensional data by reducing the dimensionality to only a few dimensions. The most important technique in this respect is principal component analysis, which is first treated in a relatively non-mathematical way in Chapter 17. Principal component analysis is also the starting point for studying many other multivariate techniques and so-called latent variable techniques such as factor analysis (Chapter 34) and partial least squares (Chapter 35). For this reason, principal components analysis is considered again in a more formal and mathematical way in Chapter 31.

In some instances the objects in a table are ordered, for example, in time when samples are taken from a (measurement) process. In Chapter 7 measurements of a single variable are displayed in control charts to find out whether a process is under control (Fig. 1.2c). In Chapter 20 the same is done for multivariate data (multivariate quality control) and we investigate how to sample processes in time or space to be able to predict the value at other times or locations. In data sets ordered in time (data taken from a continuous process) or according to wavelength (spectra) the individual responses that are measured are subject to noise, as is the case for all measurements. The signal processing techniques described in Chapter 40 allow a better and more informative description of the process by reducing the noise.

Many of the techniques described in Chapter 34 are also applied to data that are ordered in some way. However, the data are now multivariate (Fig. 1.2d). A typical situation is that of spectra obtained for samples ordered in time. In such situations one suspects that there are several compounds involved, i.e. the measured spectra are the sum of spectra of different compounds. However, one does not know how many compounds, or their spectra, or their individual concentration profiles in the time direction. The techniques described attempt to extract this type of information.

1.2.2.2 Hypothesis testing

The description of the data often leads to the formulation of hypotheses (which are then verified by *hypothesis testing*), to describe quantitatively the value of one or more variables as a function of the value of some other variables (*modelling*) or to try to classify the objects according to the values of the variables obtained (*classification*). Hypothesis testing, modelling and classification are the main operations required when we want to extract information from data in a more formal way than by visual evaluation. They are related: classification can be considered as a special kind of modelling and a model is often validated through the use of hypothesis tests.

Hypothesis testing is the main subject of the chapters listed in Table 1.2. In these chapters the characteristics of two sets of data are often compared, for instance

TABLE 1.2

Brief contents of the chapters in which the emphasis is on hypothesis testing

Chapter 4:	General principles; comparing an experimentally obtained mean with a given value.
Chapter 5:	Comparison of two means or variances, detection of extreme values, comparison of an experimental distribution with the normal distribution.
Chapter 6:	Comparison of more than two means and/or variances (analysis of variance).
Chapter 13:	Applications to method validation.
Chapter 16:	Tests of hypotheses about frequency (contingency) tables involving only two variables.

their means or their standard deviations and the hypothesis tested is their equality (or sometimes their inequality). However, other hypotheses can also be tested, such as, e.g., whether a certain result belongs to a set of results or not (outlier testing).

The subject of hypothesis testing is so essential for statistics and chemometrics that it is applied in most chapters. It is for instance important in the chapters on modelling, e.g. Chapters 8 and 10. In these chapters a model is proposed and a hypothesis test is required to show that the model can indeed be accepted. In the chapters on experimental design (Section 1.2.1) techniques are described that allow us to detect factors that may have an effect on the response under study. A hypothesis test is then applied to decide whether indeed the effect is significant.

Certain tests are specific to certain application domains. This is the case for Chapter 38, for instance, where tests are applied to sensory analysis that are not applied in other domains.

1.2.2.3 Modelling

Modelling is the main emphasis of the chapters described in Table 1.3. These chapters describe regression techniques of different complexity. Modelling is also an important aspect in: Chapter 44 on neural networks; Chapters 12 and 19, where methods are explained for robust and fuzzy regression, respectively; and in Chapter 41, where Kalman filters are applied to the modelling of dynamic processes. It is an important tool in many other chapters, e.g. Chapter 13 (method validation), Chapter 24 (response surface methodology), Chapter 37 (quantitative structure–activity relationships), and Chapter 39 (pharmacokinetic models).

TABLE 1.3

Brief contents of the chapters in which the emphasis is on modelling

Chapter 8:	Univariate regression and calibration.
Chapter 10:	Multivariate and polynomial regression.
Chapter 11:	Non-linear regression.
Chapter 35:	Latent variable-based methods for relating two data tables.
Chapter 36:	Multivariate calibration.

Modelling is applied when two or more characteristics of the same objects are measured, for example when one tries to relate instrumental responses to sensory characteristics (Chapter 38), chemical structure of a drug to its activity (Chapter 37), or the performance of two analytical methods by analysing the same objects with the two methods (Chapter 13). This is also the case when one verifies whether there is a (linear) relationship between objects of two populations by measuring the correlation coefficient (Chapter 8). The purpose of the modelling is to find relationships that explain the data and/or to allow us to make predictions.

The data structure is shown in Figs. 1.3a–c. Two sets of data are related: the **y** or **Y** data have to be explained or predicted from the **x** or **X** data. In Fig. 1.3a a single column of y values are related to a single column of x values. In a classical univariate calibration experiment (Chapter 8) the x values would be concentrations and the y values some response such as absorbance. Both Figs. 1.3b and c are

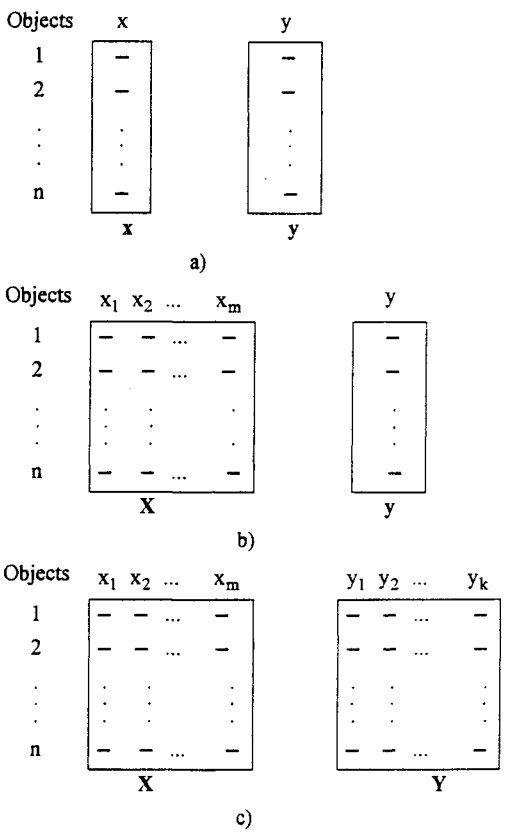


Fig. 1.3. Data structures for modelling: (a) relationship between two variables x and y for a set of objects, i.e. between vectors **x** and **y**; (b) relationship between a set of variables x (matrix **X**) and a single variable y (vector **y**); (c) relationship between two sets of variables (**X** and **Y**).

multivariate situations with one y and several x values, respectively (Chapter 10), or even several x and several y values per object (Chapter 35). Techniques that allow us to work with such data structures are extremely important topics in chemometrics. However, we do not share the tendency of some chemometricians to consider this as the only topic of importance.

The modelling element can be important without being explicit. This is the case with the neural networks of Chapter 44. One of its main uses is to model complex phenomena. Very good results can be obtained, but the model as such is usually not derived.

Modelling and hypothesis testing are related. In many cases they are either alternatives or complementary. When the methods that emphasize hypothesis testing are applied, the question is often: does this process (or measurement) yield the same result (or response) at pH = 6 and at pH = 7 or pH = 8? It does not give an immediate answer to what will happen at pH = 7.5 and this can be answered by modelling techniques. When such a technique is applied, the question will be: how does the result or response depend on pH? When simplifying both questions, one eventually is led to ask: does the pH influence the result or response? It is, therefore, not surprising that the same question can be treated both with hypothesis tests and modelling approaches, as will be the case in Chapters 13 on method validation and Chapter 22 on two-level factorial designs.

1.2.2.4 Classification

In classification one tries to decide whether the objects can be classified into certain classes, based on the values they show for certain variables. The data structures are shown in Figs. 1.4a–c. Chapters 30 and 33 are devoted entirely to this aspect and it is an important topic in Chapter 44.

Basically, there are three types of question:

- Can the objects be classified into certain classes (see also Fig. 1.4a)? The classes are not known *a priori*. This is called unsupervised pattern recognition or learning or also clustering; it is discussed in Chapter 30. The Kohonen and fuzzy adaptive resonance theory networks of Chapter 43 have the same purpose.
- Can a new object be classified in one of a number of given classes (see also Fig. 1.4b) described by a set of objects of known classification? This is called supervised pattern recognition or discriminant analysis and is described in Chapter 33. Most of the neural networks described in Chapter 44 can be applied for the same purpose and so can the inductive expert systems of Chapter 18.
- Does the object belong to a given class described by a set of objects known to belong to that class (see also Fig 1.4c)? This can be studied with the disjoint class modelling by supervised pattern recognition methods such as SIMCA or UNEQ described in Chapter 33.

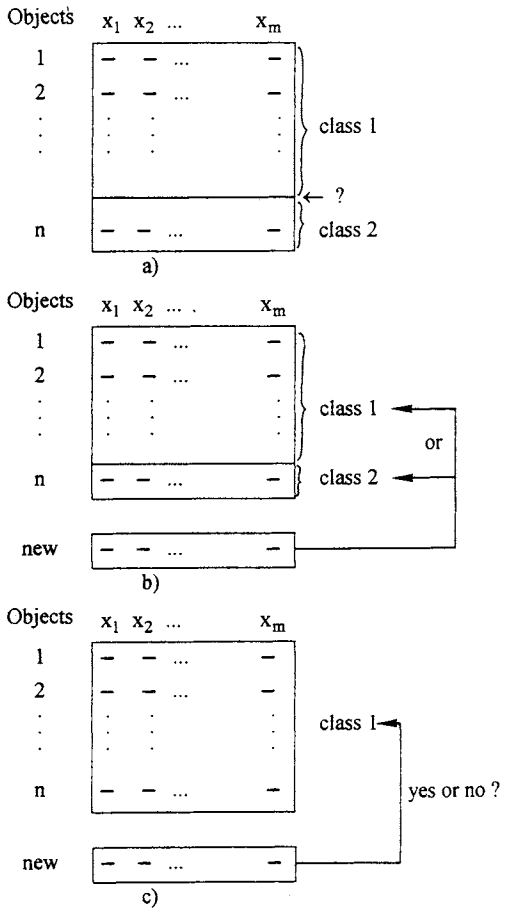


Fig. 1.4. Data structures for classification: (a) classification of a set of objects, characterized by several variables, classes not known *a priori*; (b) classification of a new object into one of a number of given classes, each class being described by a set of objects for which several variables were measured; (c) does a new object belong to a given class, described by a set of objects for which several variables were measured?

Additional aspects are discussed in Chapter 16 (quality of attributes in relation to classification in one of a few classes) and Chapter 19 (fuzzy search).

We should stress here again the relationship between classification on the one hand and modelling or hypothesis testing on the other. For instance, supervised pattern recognition methods can, in certain cases, be replaced by modelling methods such as PLS (see Chapter 33) when the y-variables are class indicator variables (e.g. 0 for class A and 1 for class B), while SIMCA can be reformulated as a multivariate outlier test.

1.2.3 Chemical knowledge and (artificial) intelligence

Deductive reasoning capacity in chemistry, as in all science branches, is the basic and major source of chemical information. This reasoning capacity is generally associated with the concept of intelligence. With the development of numerical methods and computer technology it became possible to extract chemical information from data in a way that had not previously been possible. Chemometric methods were developed that incorporated and adapted these numerical techniques to solve chemical problems. It became clear, however, that numerical chemometric techniques did not replace deductive reasoning, but rather were complementary. Problems that can be solved by numerical techniques, e.g. pattern recognition, cannot easily be solved by deductive reasoning. However, selection of the best chemical analysis conditions, for example, is a deductive reasoning process and cannot be solved in a straightforward way by mathematical methods.

To increase further the efficiency and power of chemometric methods the deductive reasoning process must be incorporated. This has resulted in the development of the so-called "expert systems" (Chapter 43). These are computer programs which incorporate a small part of the formalized reasoning process of an expert. In the 1980s they were very popular but their performance to solve difficult problems was clearly overestimated. In the early 1990s there was a dip in their application and development and the phrase "expert system" became almost taboo. Recently, however, they have reappeared under the name of decision support systems, incorporated among others in chemical instruments. In combination with the numerical chemometric methods they can be very useful.

The inductive reasoning process (learning from examples) is implemented in the inductive expert systems (see Chapter 18). Neural networks can also be considered as an implementation of the inductive reasoning process (Chapter 44).

1.2.4 Chemical domains and quality aspects

Quality is an important point in many chapters of this book. In Chapter 2, we introduce the first elements of statistical process control (SPC), and Chapter 7 is entirely devoted to quality control. Chapters 13 and 14 describe an important element of quality assurance in the laboratory, namely, how to validate measurement methods, i.e. how to make sure that they are able to achieve sufficient precision, accuracy, etc. Chapter 14 also describes how to measure proficiency of analytical laboratories. It makes no sense to carry out excellent analysis on samples that are not representative of the product or the process: the statistics of sampling are described in Chapter 20.

The treatment of sensory data is described in Chapter 38. Their importance for certain products is evident. However, sensory characteristics are not easy to

measure and require expert statistical and chemometrical attention. Chapter 40 is devoted to the analysis of signals and their improvement. It is also evident that in many cases the experimental design in Chapters 21–26 has as its final objective to achieve better quality measurements or products.

1.2.5 Mathematical and statistical tools

Statistics are important in this book, so we have decided to give a rather full account of it. However, the book is not intended to be an introduction to statistics, and therefore we have not tried to be complete. In certain cases, where we consider that chemometricians do not need that knowledge, we have provided less material than statistics books usually do. For instance, we have attached relatively little importance to the description of statistical distributions, and, while we need of course to use degrees of freedom in many calculations, we have not tried to explain the concept, but have restricted ourselves to operational and context-dependent definitions.

Most chapters describe techniques that often can only be applied to data that are continuous and measured on so-called ratio or interval scales (lengths, concentrations, temperatures, etc.). The use of other types of data often requires different techniques or leads to other results. Chapters 12, 15, 16, 18 and 19 are devoted to such data. Chapter 12 describes how to carry out hypothesis tests and regression on ranked data or on continuous data that violate the common assumption of normal distribution of measurement errors; Chapter 15 describes distributions that are obtained when the data are counts or binary data (i.e., can only be 0 or 1); Chapter 16 concerns hypothesis tests for attributes, i.e. variables that can take only two or a few values. Chapter 18 describes information theory and how this is used mainly to characterize the performance of qualitative measurements, and Chapter 19 discusses techniques that can be used with fuzzy data.

Figure 1.2 shows that in all cases the structure of the data is that of a table or tables, sometimes reduced to a single column. Mathematically, the columns are vectors and the tables are matrices. It is therefore important to be able to work with vectors and matrices; an introduction is given first in Chapter 9 and a fuller account later in Chapter 29.

1.2.6 Organization of the book

The book consists of two volumes. In the first volume (Part A) the emphasis is on the classical statistical methods for hypothesis testing and regression and the methods for experimental design. In the second volume (Part B) more attention is given to multivariate methods, often based on latent variables, to signal processing and to some of the more recent methods that are considered to belong to the artificial intelligence area.

One can certainly not state that the methods described in Volume I are all simpler, older, or used more generally than those described in Volume II. For instance, techniques such as non-linear regression using ACE or cubic splines (Chapter 11), robust regression (Chapter 12), fuzzy regression (Chapter 19) or genetic algorithms (Chapter 27) are certainly not commonplace. However, the general level of mathematics is higher in Part B than in Part A. For that reason certain subjects are discussed twice, once at a more introductory level in Part A, and once at a higher level of abstraction in Part B. This is the case for matrix algebra (Chapters 9 and 29) and principal component analysis (Chapters 17 and 31).

1.3 Some historical considerations

The roots of chemometrics go back to 1969 when Jurs, Kowalski and Isenhour published a series of papers in *Analytical Chemistry* [3–5] on the application of a linear learning machine to classify low resolution mass spectra. These papers introduced an innovative way of thinking to transform large amounts of analytical data into meaningful information. The incentive for this new kind of research in analytical chemistry was that "for years experimental scientists have filled laboratory notebooks which often has been disregarded because lack of proper data interpretation techniques" [6]. How true this statement still is today, more than 25 years later! This new way of thinking was developed further by Wold into what he called "soft modelling" [7] when he introduced the SIMCA algorithm for modelling multivariate data. These new techniques did not pass unnoticed by other academic groups, who became actively involved in the application of 'modern' algorithms as well. The common interest of these groups was to take advantage of the increasing calculation power offered by computers to extract information from large data-sets or to solve difficult optimization problems. Dijkstra applied information theory to compress libraries of mass spectra [8]. Compression was necessary to store spectra in the limited computer memory available at that time and to speed up the retrieval process. At the same time Massart became active in this field. His interest was to optimize the process of developing new chromatographic methods by the application of principles from operations research [9]. These developments coincided with a fundamental discussion about the scientific basis of analytical chemistry. In Germany this led to the foundation of the 'Arbeitskreis Automation in der Analyse', which published a series of more or less philosophical papers on the systems approach of analytical chemistry [10,11]. All this coincided with the growing belief of analytical chemists that "some of the newer mathematical methods or theories, such as pattern recognition, information theory, operations research, etc. are relevant to some of the basic aims of analytical chemistry, such

as the evaluation, optimization, selection, classification, combination and assignment of procedures, in short all those processes involved in determining exactly which analytical procedure or programme should be used" [12]. Also in other fields outside analytical chemistry, the application of pattern recognition received a great deal of attention. An important area was the study of relationships between chemical structures and their biological activity, e.g. in drug design, where several papers began to appear in the early 1970s [13].

It took until June 1972 before this research was called "chemometrics". This name was mentioned for the first time by Wold in a paper published in a Swedish journal [14] on the application of splines to fit data. He christened his group "Forskningsgruppen for Kemometri", an example which would be followed by Kowalski, who named his group the "Laboratory of Chemometrics".

The collaboration between Wold and Kowalski resulted in the foundation of the Chemometrics Society in 1974. A year later, the society defined chemometrics as follows: "it is the chemical discipline that uses mathematical and statistical methods to design or select optimal measurement procedures and experiments and to provide maximum chemical information by analysing chemical data" [15]. As one may notice, in this book we have adapted this definition by including a third objective "to obtain knowledge about chemical systems" and we have specified that the chemical information should be "relevant". With the distribution of the software packages ARTHUR [16] by Kowalski and SIMCA [7] by Wold, many interested analytical chemists were able to explore the potentials of pattern recognition and multivariate statistics in their work. In 1976 a symposium was organized entitled "Chemometrics: Theory and Application" sponsored by the Division of Computers in Chemistry of the American Chemical Society, which published the first book on chemometrics [17] with contributions from Deming (optimization), Harper (ARTHUR), Malinowski (factor analysis), Howery (target-transformation factor analysis), Wold (SIMCA) and others. This book already indicated some of the main directions chemometrics would follow: design of experiments, optimization and multivariate data analysis.

Two years later, in 1978, three European chemometricians, Kateman, Massart and Smit organized the international "Computers in Analytical Chemistry (CAC)" conference in Amsterdam — the first of what was to be a long series. This coincided with the launching of Elsevier's series "Computer Techniques and Optimization" in *Analytica Chimica Acta* under the editorship of Clerc and Ziegler [18]. On this occasion more than 100 analytical chemists from all over the world gathered to hear about a new and exciting discipline. After the first book on chemometrics was published by the ACS, other textbooks rapidly followed in 1978 by Massart et al. [12], in 1981 by Kateman and Pijpers [19], in 1988 again by Massart et al. [2], in 1982 by Lewi [20] and in 1986 by Sharaf et al. [21]. Other, no less important, textbooks are mentioned in the suggested reading list.

A milestone in the short history of chemometrics was certainly the introduction of Partial Least Squares [22] by S. Wold and coworkers in 1983, based on the early work of H. Wold [23]. Since 1972, *Analytical Chemistry*, an ACS publication, included a section on Statistical and Mathematical Methods in Analytical Chemistry in their biannual reviews. In 1978 Shoenfeld and DeVoe [24] provided the editor of *Analytical Chemistry* with the new title "Chemometrics". This was a formal recognition of the appearance of a new discipline in analytical chemistry, which was emphasized by the special attention on chemometrics at a symposium organized on the occasion of the celebration of the 50th anniversary of *Analytical Chemistry* [25]. Since 1980, the field of research expanded rapidly and several new centres in Europe and the USA emerged which became actively involved in chemometrics. Norway, Italy and Spain, for instance, are three of the centres of chemometrics in Europe. In 1974 two teams became active in chemometrics in Italy, those of Forina in Genova and of Clementi in Perugia. Around this time, other chemists began to pay more attention to the statistical side of analytical chemistry, such as Dondi in Ferrara. In 1978 Forina and Armanino published "Elements of Chemometrics", the first book for second-cycle students of Chemometrics in Italy. However, a milestone in the chemometrics history in Italy is certainly April 13, 1981, when the first Italian seminar on chemometrics was organized in Genova. In 1983 all major centres were represented at the most inspiring NATO Advanced Study Institute on Chemometrics [26] in Cosenza, hosted by Forina. Experts presented and discussed in a relaxed ambience recent developments in experimental design, multivariate calibration and factor analysis. The 1980s witnessed further growth and diversification. In Seattle Kowalski focused on Process Analytical Chemistry at his Centre for Process Analytical Chemistry (CPAC), a successful consortium between the University of Washington and a number of industrial partners, which would expand to the impressive number of 50 partners. In Europe there was a growing belief that much analytical knowledge could not be caught in either hard or soft models. Therefore, an EEC-funded research project "Expert systems for chemical analysis" was initiated in 1986 with Vandeginste, Massart and Kateman and three industrial partners. At the same time the EEC funded a large chemometrics teaching network, Eurochemometrics, which would organize an impressive series of short courses all over Europe. In the US leading chemometricians gather annually at the prestigious Gordon Research Conference "Statistics in Chemistry and Chemical Engineering" where advanced research topics are discussed and commented upon. The launching of two specialized chemometrics journals in 1986, the *Journal of Chemometrics* (Wiley) and *Chemometrics and Intelligent Laboratory Systems* (Elsevier), confirmed that chemometrics had evolved into an established science.

By the end of the 1980s, industry had become increasingly interested in this new and promising field and was offering positions to young chemometricians. A real

challenge for them is to prove that chemometric algorithms are robust enough to cope with the dirty data measured in practice, and powerful enough to derive the requested information from the data. For instance, curve resolution methods should not only be able to resolve the spectra of major co-eluting compounds by HPLC but also compounds at a level below 1% detected with a diode array detector.

Although the study and introduction of novel multivariate statistical methods remained in the mainstream of chemometric research, a distinct interest arose in computationally intensive methods such as neural networks, genetic algorithms, more sophisticated regression techniques and the analysis of multiway tables. Chemometricians rapidly discovered the wealth of new opportunities offered by modern communication tools such as e-mail and the Internet. An on-line discussion group ("ICS-L") on chemometrics has been set up for the International Chemometrics Society, using the LISTSERV facility. More recently, the first worldwide electronic conferences (InCINC94 and InCINC96) have been organized by Wise and Hopke, who have set an example which will certainly be followed by many others. Also, many research groups in chemometrics have started to build home pages with information about their current research activities.

Despite the enormous progress in our capability of analyzing two- and three-way tables, a fundamental issue remains the poor precision reported in collaborative analytical studies. Apparently, analytical methods which are essential in process control and in research and development lack robustness and are not suitable for their purpose. The application of multicriteria decision-making and Taguchi designs as suggested by several chemometricians should lead to some improvements. Such a study is the objective of an EU-funded project within the Standards, Measurement and Testing Programme, carried out by a group of Italian, Spanish and Belgian academic chemometrics centres in collaboration with major European industries.

The 1990s appear to be the age of quality and quality improvement. Most industrial and governmental laboratories are opting for compliance with one of the quality systems: GLP, ISO 9000, etc. To demonstrate and maintain quality requires a skilful application of statistics. This is the area of Qualimetrics which is the synergy between chemometrics and quality assurance — an area of great importance for the industrial chemist.

1.4 Chemometrics in industry and academia

Many excellent researchers in academia as well as in industry have contributed to the successful development of chemometrics in recent years. The interested reader is referred to the *Journal of Chemometrics* [27,28], which devotes a column on academic chemometric research, in which centres present their work and

philosophies on chemometrics and its future. It shows that many distinguished academic chemometricians either started their careers in industry, or (Windig, Lewi, de Jong, Berridge, Vandeginste, etc.) still occupy a research position in industry. We ought to realize that long before statistics and chemistry found their synergy in chemometrics applied statistics was an indispensable tool in industrial research and production. The first industrial chemometrician was probably W.S. Gossett (Student), who developed the *t*-test, while working for the Guinness breweries [29]. In 1947 Box from ICI published a book on "Statistical Methods in Research and Production" [30], followed in 1956 by "The Design and Analysis of Industrial Experiments" [31], which culminated in the book by Box, Hunter and Hunter [32]. In his introduction Box writes "Imperial Chemical Industries Ltd has long recognized that statistical methods have an important part to play in industrial research and production". This book, illustrated with many real-life examples from ICI and written for the chemist is still recommended basic reading on experimental design. Lloyd Currie at the National Bureau of Standards (now the National Institute of Standards and Technology) has been an early promoter of the application of statistics in analytical chemistry. Certainly the work of Malmstadt, Enke and Crouch, while not directly chemometrics in the early days, led to its development by getting more people involved in the details of data collection. Very rapidly, chemometrics started to cover the whole range from fundamental research to development. For obvious reasons industry became active in developing chemometrics applications and tools. Optimization in HPLC is a very good illustration. For many years sequential and simultaneous optimization strategies by Simplex and experimental design were the subject of research for many academic chemometricians. The importance of optimization for productivity improvement was quickly recognized by industrial researchers who demonstrated its practical applicability — Berridge at Pfizer, and Glajch and Kirkland at DuPont de Nemours, to name but a few. Instrument manufacturers took over the idea and developed this technique further to systems integrated with HPLC equipment. As a result, optimization strategies are now widely available and routinely applied in industrial laboratories. The same happened to PLS which is included in the software of NIRA instruments and in molecular modelling software.

Chemometrics is fairly well disseminated in industry. Many applications are found in the pharmaceutical industry, e.g. the study of structure–activity correlations is of great importance to guide the synthesis process in the search for new active drugs. In the experimental phase synthesis conditions are optimized for maximal yield. Once a compound is synthesized, a long and tedious route follows in determining whether the compound is really active or not. In this area, Lewi developed in 1976 a Biplot technique, called Spectral Map Analysis (SMA) [33]. Originally, SMA was developed for the visualization (or mapping) of activity spectra of chemical compounds that had been tested in a battery of pharmacological

assays. The problem of classifying compounds with respect to their biological activity spectra is a multidimensional problem which can be solved by factor analytical methods. In the food industry the analysis of sensory data in its various facets requires a multivariate approach, as does, for example, the prediction of taste keepability of vegetable oils from a few indicators, the classification of oils according to their origin and the evaluation of nutrition trials. Complex modelling techniques are applied to relate data blocks from several origins, e.g. the chewing pattern of test persons to the intensity of their taste perception as a function of time for various products. The petrochemical industry wants to predict the oil content in rocks, e.g. by variable temperature FTIR.

Coinciding with early academic chemometrics activities, analytical chemists of the Dutch States Mines (a large producer of bulk chemicals and fertilizers) realized that the analytical laboratory is not simply a producer of numbers but forms an integral part of a process control chain. The quality of the analytical results, expressed in terms of speed, precision and sampling rate, defines the effectiveness of process control. Van der Grinten [34] and Leemans [35] in their pioneering work on process analytical chemistry derived a relationship between the properties of the analytical method and the capability of process control. In-line analysis offers speed often at the expense of precision, specificity and selectivity. This is the area of sensors and NIRA in combination with multivariate calibration, where several successful industrial applications have been reported [36].

Another important line of industrial applications is the retrieval of spectroscopic data from libraries, and the interpretation of combined IR, NMR and MS spectra, including the resolution of mixture spectra by Factor Analysis [37].

In the area of chromatography, the optimization of the mobile phase in HPLC has received much attention from both instrument manufacturers and industrial analytical chemists [38]. LC linked to full scan spectroscopic detectors (UV-Vis, IR) is a common technique in many analytical laboratories, specifically for the assessment of the purity of pharmaceutical compounds. In this area Windig at Kodak developed the Variogram [39] and SIMPLISMA [40] to decompose bi-linear data produced by hyphenated techniques in its pure factors (spectra).

The ultimate added value of chemometrics — better chemistry, more efficient experiments and better information from data — is highly relevant and appealing to any industry. This is also the incentive for collaboration with academic centres of expertise, usually in a regional network, as there are in the Benelux, Spain, Scandinavia and the USA. These centres fulfil a twofold function: the distribution of chemometric principles to the bench of the industrial chemical researcher, and a source of inspiration for improvements and new challenges to be picked up by fundamental chemometrics researchers.

Which challenges can we expect for the future? A number of clear trends are showing up. First, there is the still growing mass of data. An example is an image

where each pixel is no longer characterized by a grey scale or by a colour but instead by a complete spectrum. Supposing that the spectrum is measured at 1024 wavenumbers, this represents a stack of 1024 images of a typical size of 512×512 pixels, which should all be treated together by multivariate procedures! Secondly, data have become increasingly complex. In high resolution NMR, for example, spectra are measured in two and higher dimensions and contain an enormous amount of information on the secondary and tertiary structure of large molecules. This is an area in which advanced chemometric methods should provide real added value. Thirdly, relationships or models being studied are increasingly complex, e.g. complex multivariate models are necessary to relate product quality to all relevant manufacturing conditions for process control, or to relate three-dimensional structures of macromolecules to pharmaceutical or biological activity. Computer-intensive methods will become increasingly important in the development of robust models and non-parametric inference methods based on randomization tests. By using genetic algorithms the model itself is becoming part of the modelling process. Many industries realize that the preservation and accessibility of corporate knowledge stored in (electronic) lab notebooks or laboratory information management systems is becoming the Achilles' heel of success. Scientists should be able to interrogate and visualize data available in various formats: spectra, structures, texts, tables, images, catalogues, databases, etc. and their interrelationships in a network of information. One can imagine that by using hypermedia [41] a researcher will be able to navigate through a giant web of data, which are instantaneously processed into information by local modelling or artificial intelligence and are displayed in a directly interpretable way, e.g. by virtual reality. This enormous task will be one of the challenges for chemometricians and information scientists.

References

1. D. Oldroyd, The Arch of Knowledge. Methuen, New York, 1986.
2. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, Chemometrics: A Textbook. Elsevier, Amsterdam, 1988.
3. P.C. Jurs, B.R. Kowalski, T.L. Isenhour and C.N. Reilly, Computerized learning machines applied to chemical problems. Anal. Chem., 41 (1969) 690–695.
4. P.C. Jurs, B.R. Kowalski, T.L. Isenhour and C.N. Reilly, Computerized learning machines applied to chemical problems: Interpretation of infrared spectrometry data. Anal. Chem., 41 (1969) 1949–1953.
5. P.C. Jurs, B.R. Kowalski, T.L. Isenhour and C.N. Reilly, Computerized learning machines applied to chemical problems: Multicategory pattern classification by least squares. Anal. Chem., 41 (1969) 695–700.
6. C.F. Bender, Pattern recognition. New approach interpreting chemical information. Comput. Chem. Res. Educ., Proc. Int. Conf., 2 (1973) 3/75–3/80.
7. S. Wold and M. Sjöström, SIMCA: A method for analyzing chemical data in terms of similarity and analogy, in: B.R. Kowalski (Ed.) Chemometrics: Theory and Application. ACS Symp. Ser., 52, Analytical Chemical Society, Washington, DC, 1977.

8. G. van Marlen and A. Dijkstra, Information theory applied to selection of peaks for retrieval of mass spectra. Anal. Chem., 48 (1976) 595–598.
9. D.L. Massart and L. Kaufman, Operations research in analytical chemistry. Anal. Chem., 47 (1975) 1244A–1253A.
10. Arbeitskreis Automation in der Analyse, System Theorie in der Analytik. I. Definitionen und Interpretationen Systemtheoretischer Grundbegriffe. Zeit. Anal. Chemie, 256 (1971) 257–270.
11. Arbeitskreis Automation in der Analyse, System Theorie in der Analytik. II. System der Analytischen Mengenbereiche. Zeit. Anal. Chemie, 261 (1972) 1–10.
12. D.L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures. Elsevier, Amsterdam, 1978.
13. A.J. Stuper, W.E. Brugger and P.C. Jurs, A computer system for structure–activity studies using chemical structure information handling and pattern recognition techniques, in: B.R. Kowalski (Ed.) Chemometrics: Theory and Application. ACS Symp. Ser., Analytical Chemical Society, Washington, DC, 1977.
14. S. Wold, Spline-funktioner-ett nytt verktyg i data-analysen. Kemisk Tidskr., 3 (1972) 34–37.
15. B.R. Kowalski, Chemometrics. Chem. Ind., 22 (1978) 882.
16. A.M. Harper, D.L. Duewer, B.R. Kowalski and J.L. Fasching, ARTHUR, an experimental data analysis: the heuristic use of a polyalgorithm, in: B.R. Kowalski (Ed.), Chemometrics: Theory and Application. ACS Symp. Ser. 52, American Chemical Society, Washington, DC, 1977.
17. B.R. Kowalski (Ed.), Chemometrics: Theory and Application. ACS Symp. Ser. 52, American Chemical Society, Washington, DC, 1977.
18. J. Clerc and E. Ziegler, Editorial. Anal. Chim. Acta, 95 (1977) 1.
19. G. Kateman and F. Pijpers, Quality Control in Analytical Chemistry. Wiley, New York, 1981.
20. P.J. Lewi, Multivariate Data Analysis in Industrial Practice. Wiley, Chichester, 1982.
21. M.A. Sharaf, D.L. Illman and B.R. Kowalski, Chemometrics. Wiley, New York, 1986.
22. W. Lindberg, J.A. Person and S. Wold, Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate. Anal. Chem., 55 (1983) 643–648.
23. H. Wold in P.R. Krishnaiah (Ed.), Multivariate Analysis. Academic Press, New York, 1966.
24. P.S. Schoenfeld and J.R. DeVoe, Statistical and mathematical methods in analytical chemistry. Anal. Chem., 48 (1976) 403R–411R.
25. B.R. Kowalski, Analytical chemistry: the journal and the science, the 1970's and beyond. Anal. Chem., 50 (1978) 1309A–1313A.
26. B.R. Kowalski (Ed.), Chemometrics: Mathematics and Statistics in Chemistry. Reidel, Dordrecht, 1984.
27. P. Geladi and K. Esbensen, The start and early history of chemometrics: selected interviews, part 1. J. Chemometrics, 4 (1990) 337–354.
28. K. Esbensen and P. Geladi, The start and early history of chemometrics: selected interviews, part 2. J. Chemometrics, 4 (1990) 389–412.
29. Student, The probable error of a mean. Biometrika, 6 (1908) 1–25.
30. O.L. Davies (Ed.), Statistical Methods in Research and Production Design and Analysis of Industrial Experiments. Oliver and Boyd, Edinburgh, 1947.
31. O.L. Davies (Ed.), The Design and Analysis of Industrial Experiments. Oliver and Boyd, Edinburgh, 1954.
32. G.E.P. Box, W.G. Hunter and J.S. Hunter, Statistics for Experimenters, an Introduction to Design, Data Analysis and Model Building. Wiley, New York, 1978.
33. P.J. Lewi, Spectral map analysis: factorial analysis of contrast, especially from log ratios. Chemom. Intell. Lab. Syst., 5 (1989) 105–116.
34. P.M.E.M. van der Grinten, Regeltechniek en Automatisering in de procesindustrie. Het Spectrum,

Amsterdam, 1970.
35. F.A. Leemans, Selection of an optimum analytical technique for process control. Anal. Chem., 43 (1971) 36A–49A.
36. H. Martens and T. Naes, Multivariate Calibration. Wiley, New York, 1989.
37. M. Maeder and A.D. Zuberbuehler, The resolution of overlapping chromatographic peaks by evolving factor analysis. Anal. Chim. Acta, 181 (1986) 287–291.
38. J.C. Berridge, Techniques for the Automated Optimization of HPLC Separations. Wiley, New York, 1985.
39. W. Windig and H.L.C. Meuzelaar, Nonsupervised numerical component extraction from pyrolysis mass spectra of complex mixtures. Anal. Chem., 56 (1984) 2297–2303.
40. W. Windig and J. Guilment, Interactive self-modeling mixture analysis. Anal. Chem., 63 (1991) 1425–1432.
41. C.L. Macher, M. Cadish, J.-T. Clerc and E. Pretsch, Hypermedia — a new concept for information management. Chemom. Intell. Lab. Syst., 28 (1995) 213–228.

Further suggested reading on historical and general chemometrics

P. Geladi, The history of chemometrics. Chemometrics in Belgium Newsletter, 2 (1995).
P. Geladi and A. Smilde, The future of Chemometrics. J. Chemometrics, 9 (1995) 1.
B.G.M. Vandeginste, Chemometrics — general introduction and historical development. Topics Curr. Chem., 141 (1987) 1.
B.G.M. Vandeginste, Chemometrics in the Benelux. Chemom. Intell. Lab. Syst., 25 (1994) 147.

Other books on chemometrics

R.G. Brereton, Chemometrics: Application of Mathematics and Statistics to Laboratory Systems. Ellis Horwood, Chichester, 1990.
R.G. Brereton, Multivariate Pattern Recognition in Chemometrics. Elsevier, Amsterdam, 1992.
R. Cela, Avances en Quimiometría Práctica, Universidade de Santiago de Compostela, 1994.
S.N. Deming and S.L. Morgan, Experimental Design: A Chemometric Approach. Elsevier, Amsterdam, 1987.
I.E. Frank and R. Todeschini, The Data Analysis Handbook, Elsevier. Amsterdam, 1994.
A. Höskuldsson, Prediction Methods in Science and Technology. Vol. 1: Basic Theory. Thor Publishing, Denmark, 1996.
G. Kateman and L. Buydens, Quality Control in Analytical Chemistry, 2nd ed. Wiley, New York, 1993.
E.R. Malinowski, Factor Analysis in Chemistry, 2nd ed. Wiley, New York, 1991.
T. Naes and E. Risvik, Multivariate Analysis of Data in Sensory Science. Elsevier, Amsterdam, 1996.
R. Nortvedt, F. Brakstad, D.M. Kvalheim and T. Lundstedt, Application of Chemometrics within Research and Industry (Anvendelse av Kjemometri innem Forsking og Industri), Tidsskreift-forlaget KJEMI, 1996.

Chapter 2

Statistical Description of the Quality of Processes and Measurements

2.1 Introductory concepts about chemical data

Measurements generate data and these data are used to describe or evaluate the quality of processes and measurement procedures. A first question we must answer is how to utilize the data so that they give us more insight into the performance characteristics that will be used in the evaluation of the processes and measurements. This description and the performance characteristics are the subject of this chapter.

2.1.1 Populations and samples

Let us suppose that we have determined the concentration of sodium in five randomly selected bottles of water of a certain brand. These five bottles then constitute a *sample* in the statistical sense. They are a sample of the *population* of all existing bottles of water of that certain brand. In the same way, if we carry out six replicate determinations of sodium in a certain material, then the six individual observations constitute a sample — in this case from a population of all determinations of sodium that could have been made with that measurement technique on that specific matrix if its supply were unlimited.

The population of measurements consists of all the possible measurements that can be made and a set of experiments is considered to be a sample of the population of all the experiments that can be made, given unlimited resources.

We observe that populations are often very large (the number of bottles) or infinite (the number of determinations). Although the number of existing bottles may be considered finite, it will be treated as infinite. There are, however, cases where populations are clearly finite. For instance, if we were to measure some characteristic for the 50 states of the USA, then that population (of states) is finite and small enough to be completely measured. In a few instances, statistical texts make distinctions between finite and infinite populations, but in almost all cases the population will be considered to consist of an infinite number of individuals, objects, measurements and we would investigate a finite sample of these to make conclusions about the whole population.

There is clearly a problem in terminology due to the use of the term "sample" and derivations such as "sampling" by analytical chemists, where the word means any material or test portion to be analyzed without necessarily supposing that that material is a sample of a larger population. For instance, if a forensic toxicologist is asked to analyze a tablet collected on the scene of a crime, he will call that his sample although the object is unique. To avoid confusion between statistical and chemical usage of the word, IUPAC [1] has proposed that in chemistry "sample" should only be used when it is a portion of a material selected from a larger quantity of material. This is consistent with statistical terminology. It also implies the existence of a sampling error, since the sample may not reflect accurately the content of the larger quantity. Sampling and sampling errors are discussed in Chapter 20. When sampling errors are negligible, such as when the parent material is a liquid and a small portion of it is analyzed, then the IUPAC guideline suggests the use of terms such as test portion, aliquot or specimen.

2.1.2 Variables and attributes

Variables can be defined as properties with respect to which individual elements in a sample differ in some ascertainable way [2].

Variables can be measured on three types of statistical scales:

- The *nominal scale* is used when the individuals or objects can only be described in words. An object may be black, white or red, an individual manufactured object may be defective or acceptable, etc. The terms black, white and red or defective and acceptable constitute the nominal scale. Variables measured in this way are often called *qualitative* or *categorical* variables or *attributes*.
- The *ordinal scale* consists of giving ranked values to a variable. An individual object's quality may be rated as "very poor", "poor", "average", "good", "excellent". There is a clear gradation in these terms. Variables measured in this scale are often called *ranked variables*.
- The *interval* and *ratio scales* are measured on a scale in which the distance along that scale can be measured as a number. We sometimes distinguish between the two scales (the ratio scale has a zero point with an absolute value, e.g. temperature in degrees Kelvin; the interval scale has an arbitrary zero point, e.g. temperature in degrees Celsius), but this distinction is often not important to us. Variables measured in this scale are often called *measurement variables* or *quantitative* variables.

Of equal importance is the difference between *continuous variables*, such as temperature or concentration, and *discrete variables*, which can take only certain values. The latter are often the result of counting (bacterial counts, number of defects on an object) and the only possible values are then integer numbers.

We should be cautious about confusion between discrete variables and ranked variables. We could code the terms "very poor", ... "excellent" used above as 1, ..., 5. This does not make it a variable on an interval or ratio scale because the distance between "very poor" and "poor" is not necessarily equal to that between "good" and "excellent".

The type of scale of a variable determines the statistical tests that can be carried out and the distributions with which they are described. This is discussed further, for instance in Chapters 12 and 15.

While we are discussing the meaning of the term variables, it is useful to make the distinction between *univariate* and *multivariate*. More precise definitions of terms such as univariate and multivariate distribution or space will be required later, but for the moment it is sufficient to state that a data set is univariate when the individual elements are described by only one variable and multivariate when the same individual element is described by two (sometimes also called bivariate) or more variables. In the next few sections and chapters only univariate data sets will be considered, but multivariate data sets will be discussed at length later in this book.

2.1.3 Histograms and distributions

When one has many data available and wants to describe them, it is useful to group them into classes and visualize their distribution with a *histogram*. This is demonstrated with the data of Table 2.1 concerning fluoride in the enamel of young children as obtained by Cleymaet and Coomans [3]. In this case the *range* of the data is 3754–722 = 3032. A convenient class interval is 200. This yields 16 classes and leads to Table 2.2. The number of classes is chosen so that there is neither too much nor too little detail. This may require some trials, but in general one should not make fewer than 5 classes (for small numbers of data) and not more than 25 (for large data sets). Another rule of thumb is that the number of classes should be equal to the square root of the number of data.

TABLE 2.1

Fluoride concentrations in μg/g in the enamel of teeth of n = 63 young children in Antwerp, Belgium (from Cleymaet and Coomans [3])

1506	3063	2657	1964	2220	2730	3754	1128
1946	1186	1375	2196	2284	1654	1631	3081
2150	1898	2452	2187	2443	2154	3292	2162
1418	2360	2897	3208	2260	722	2495	2382
1130	2357	1890	1622	1738	2332	1399	2234
2041	1358	2733	2225	1195	2237	1975	1811
2842	1288	1862	2212	1194	1813	2189	2726
1628	1909	2239	2154	2116	2509	2004	

TABLE 2.2

Relative and cumulative frequency distribution for the data of Table 2.1

Class interval	Class mark	Frequency	Relative frequency	Cumulative frequency	Cumulative rel. frequency
700– 900	800	1	0.016	1	0.016
900–1100	1000	0	0	1	0.016
1100–1300	1200	6	0.095	7	0.111
1300–1500	1400	4	0.063	11	0.174
1500–1700	1600	5	0.079	16	0.254
1700–1900	1800	6	0.095	22	0.349
1900–2100	2000	6	0.095	28	0.444
2100–2300	2200	16	0.254	44	0.698
2300–2500	2400	7	0.111	51	0.810
2500–2700	2600	2	0.032	53	0.841
2700–2900	2800	5	0.079	58	0.921
2900–3100	3000	2	0.032	60	0.952
3100–3300	3200	2	0.032	62	0.984
3300–3500	3400	0	0	62	0.984
3500–3700	3600	0	0	62	0.984
3700–3900	3800	1	0.016	63	1.000

ISO [4] has defined the term *class* in the case of quantitative characteristics as each of the consecutive intervals into which the total interval of variation is divided. The *class limits* are the values defining the upper and lower bounds of a class. The *mid-point of a class* is then the arithmetic mean of the upper and lower limits and the *class interval* the difference between upper and lower limits of a class. The mid-point is sometimes also called *class mark* (although this is not recommended by ISO).

By counting the number of individuals in each class and dividing by the total number of all individuals, one obtains the *relative frequency* of a class and the table of these values is the *relative frequency distribution.* This can be plotted in function of the class mid-point and yields then Fig. 2.1.

By summing all frequencies up to a certain class, we obtain the *cumulative frequency*. For instance, the cumulative frequency up to and including class 1300–1500 is 1 + 0 + 6 + 4 = 11. The relative cumulative frequency is then 11/63 = 0.174 or 17.4%. Again, it is possible to plot the *cumulative relative frequency distribution* to obtain Fig. 2.2. It should be noted that, in practice, we often drop the word relative.

All these distributions are *discrete*, because the frequencies are given for discrete classes or discrete values of x (the class midpoint). When the x-values can assume continuous values, *continuous distributions* result. If the data of Table 2.2

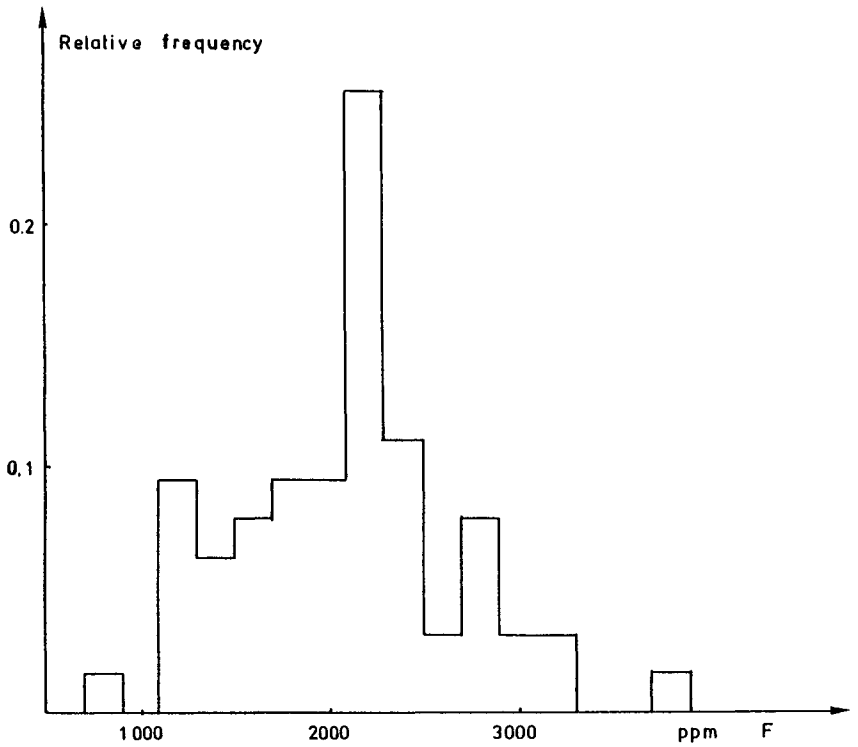


Fig. 2.1. Relative frequency distribution of the data of Table 2.1.

are truly representative of the population (i.e. the population of fluoride concentrations of the enamel of young Belgian children), then the frequencies can also be considered to be probabilities to encounter certain fluoride-values in that population. We could then state that the *probability* of obtaining fluoride values between 1300 and 1500 μg/g is 0.063 and that the *cumulative probability* of encountering values up to 1500 is 0.174. The plots of Figs. 2.1 and 2.2 could then be considered as the *probability distribution* (also called the *probability density function*) and *cumulative probability distribution*, respectively. Although, at first sight, the frequency and probability distributions are really the same, we often make a distinction between them. The frequency distribution describes the actual data, that is the data of a sample of the population. The probability distribution describes the population as such, i.e. the distribution that would be obtained for an infinite number of data. In Section 3.8 we will show that the fluoride data can be considered to be normally distributed. We can then state that the data for the $n = 63$ children yield the frequency distribution, while the probability distribution is really the normal distribution.

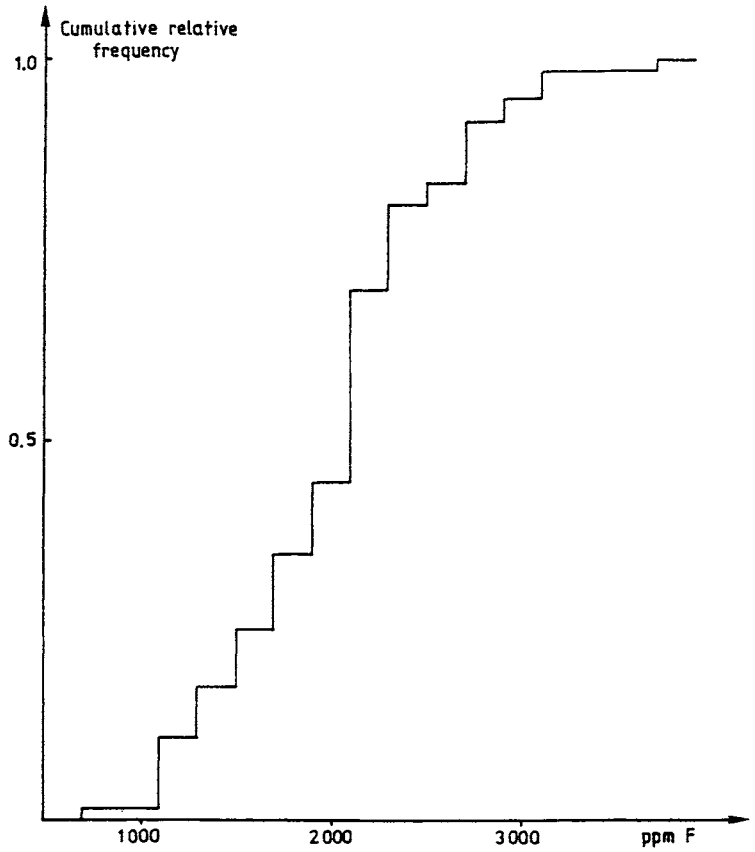


Fig. 2.2. Cumulative relative frequency distribution of the data of Table 2.1.

2.1.4 Descriptive statistics

2.1.4.1 Population parameters and their estimators

The essential statistical information for describing a simple data set consists of:

- the *number* of observations or individuals in the set, n;
- a parameter for *central tendency* or *location*, such as the (arithmetic) mean (or average);
- a parameter for *dispersion*, such as the standard deviation.

Probability distributions are characterized by *population parameters*, such as the mean and the standard deviation. To determine them would require an exhaustive number of determinations. For instance, suppose we need to determine the pH of a certain solution, then an infinite number of measurements would yield the probability distribution of the outcome of the pH measurement of that solution with a *population mean* of the measurements, μ, and a *population standard deviation* σ.

In practice, we would make a limited number of measurements, n. This is called the *sample size* because the measurements are viewed as a random sample of n measurements taken from all possible measurements. The mean obtained in this way is called the *sample mean*. The sample mean is an *estimator* of the true population mean. The concept of estimators will be discussed further in Section 3.1.

2.1.4.2 Mean and other parameters for central location

The *mean*, $\bar{x}$, is given by:

$$\bar{x} = \left(\sum_{i=1}^{n} x_i \right) / n \tag{2.1}$$

where x_i is the ith individual observation or measurement. To avoid too cumbersome a notation, this type of equation will in future usually be written as:

$$\bar{x} = \Sigma x_i / n$$

In Chapter 12 a non-parametric measure of central tendency, the median, will be described. In that Chapter we will also explain the term "non-parametric". For the moment, it is sufficient to state that non-parametric measures are preferred when the distribution characteristics are not known.

2.1.4.3 Standard deviation and variance

The *standard deviation*, s, is given by

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} \tag{2.2}$$

and the *variance* by the square of the standard deviation, s^2. They estimate respectively σ, the population standard deviation and σ^2, the population variance. The term $(n - 1)$ gives the number of *degrees of freedom* (df). In some cases, one does not divide by $(n - 1)$ but by n. When to do this is described in Chapter 3. However, eq. (2.2) can be used without problems in the rest of this chapter.

The *relative standard deviation* is given by:

$$s_r = s / \bar{x} \qquad (\bar{x} > 0) \tag{2.3}$$

and when it is expressed as a percentage, by:

$$s_r\,(\%) = 100\, s_r \tag{2.4}$$

The latter is sometimes called the *coefficient of variation*. IUPAC prefers not to use this term.

As will be described further in Chapter 3, experimentally obtained means are also subject to variation. The standard deviation of the means is called *standard*

error of the mean (SEM) and is given by

$$s_{\bar{x}} = s_x/\sqrt{n} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)}} \tag{2.5}$$

$s_{\bar{x}}$ should not be confounded with s. The term standard error is further defined in Chapter 3.

Here again we will introduce in Chapter 12 a non-parametric measure of dispersion.

The mean is also called the first (statistical) moment, and the variance the second moment of a distribution. Moments, including higher order moments such as skewness, are discussed further in Chapter 3.

2.1.4.4 Pooled standard deviation and standard deviation from paired data

In many cases groups of data have been obtained at different times or on different (but similar) samples and one wants to obtain a standard deviation from these grouped data. Let us suppose, for instance, that we want to determine the standard deviation for a determination of water in cheese [5]. Replicate determinations on several types of cheese have been carried out (see Table 2.3). Consider first only the 7 first types of cheese ($k = 7$). We would be able to determine a standard deviation for each of the 7 types of cheese separately. However, we would prefer to determine one single standard deviation for cheese as a whole. This can be done by pooling the variances according to:

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots (n_k - 1)s_k^2}{(n_1 - 1) + (n_2 - 1) + \ldots (n_k - 1)} \tag{2.6}$$

TABLE 2.3

Example of calculation of pooled standard deviation. The data are results of moisture determinations in cheese products with the Karl Fischer method (adapted from ref. [5]). The result for type 8 is artificial.

No (j)	Type of cheese product	$\bar{x}_j$(%)	s_j	n_j	df$_j$	df$_j$ s_j^2
1	Processed cheese food	43.36	0.29	10	9	0.7569
2	Processed cheese food	43.45	0.31	10	9	0.8649
3	Monterey jack	41.20	0.35	8	7	0.8575
4	Cheddar	34.96	0.24	8	7	0.4032
5	Processed American	40.41	0.30	8	7	0.6300
6	Swiss	38.52	0.31	8	7	0.6727
7	Mozzarella	52.68	0.24	9	8	0.4608
	Σ =			61	54	4.6460
	$s_p^2 = 4.6460/54 = 0.0860$	$s_p = 0.29$				
8	Type 8	51.00	1.12	8		

where n_1, ... n_k are the number of replicates in the first, ..., kth type of cheese and s_1, ... s_k are the corresponding standard deviations. The notation can be simplified, since $n_1 - 1$, ..., $n_k - 1$ are the degrees of freedom df_1, ..., df_k for each type. We can then write:

$$s^2_{\text{pooled}} = \frac{df_1\, s_1^2 + df_2\, s_2^2 + \ldots\, df_k\, s_k^2}{\sum df_j} = \frac{\sum df_j\, s_j^2}{\sum df_j} \tag{2.7}$$

for $j = 1, ..., k$.

The computations are performed in Table 2.3. The pooled standard deviation is 0.29. Suppose now that we add an 8th type of cheese as in the last line of the table. Would we be able to make the computations in the same way? At first sight, we could use eq. (2.7) with the 8 categories instead of 7. However, the standard deviation from type 8 is clearly very different from that of the 7 others. The pooled standard deviation would be (too) heavily influenced by the type 8 cheese, so that the resulting value would not be representative. This example shows that only similar variances should be pooled. Stated in a more scientific way, we can pool variances provided that they are homogeneous.

For the special case of paired replicates (i.e. each determination was carried out in duplicate, all $n_j = 2$), we can also use eq. (2.7). However, we often find the following equation:

$$s_d = \sqrt{\left(\frac{\sum d_j^2}{2}\right)/k} \tag{2.8}$$

where $d_j = x_{j1} - x_{j2}$, i.e. the difference between the two replicate values for the jth sample and k the number of pairs. An example is given in Table 2.4.

This useful result can be derived as follows. Let us first consider the standard deviation computed for a single pair of results, x_1 and x_2, with mean $\bar{x}$. It is equal to

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2}{2 - 1}}$$

As in this special case $|\bar{x} - x_1| = |\bar{x} - x_2|$, it follows that

$$s^2 = 2(\bar{x} - x_1)^2$$

or

$$s^2 = 2\left(\left(\frac{x_1 + x_2}{2}\right)^2 - 2\,\frac{x_1 + x_2}{2}\,x_1 + x_1^2\right)$$

$$= \frac{(x_1 - x_2)^2}{2}$$

TABLE 2.4

Example of calculation of s using paired data. The data concern moisture in American Cheese measured on 20 successive days (adapted from ref. [6]).

Day j	x_{j1}	x_{j2}	d_j	d_j^2
1	42.68	42.77	–0.09	0.0081
2	42.08	42.38	–0.30	0.0900
3	43.39	43.33	0.06	0.0036
4	42.87	42.98	–0.11	0.0121
5	42.70	42.95	–0.25	0.0625
6	42.93	42.95	–0.02	0.0004
7	42.78	42.97	–0.19	0.0361
8	42.92	43.20	–0.28	0.0784
9	43.34	42.89	0.45	0.2025
10	43.12	43.26	–0.14	0.0196
11	42.43	42.54	–0.11	0.0121
12	43.05	43.15	–0.10	0.0100
13	42.99	42.86	0.13	0.0169
14	43.04	42.78	0.26	0.0676
15	43.41	43.14	0.27	0.0729
16	43.12	43.23	–0.11	0.0121
17	42.25	42.53	–0.28	0.0784
18	42.96	42.78	0.18	0.0324
19	42.83	42.72	0.11	0.0121
20	42.83	43.04	–0.21	0.0441
			$\Sigma =$	0.8719

$s_d^2 = 0.8719/(2 \times 20) = 0.0218$.

$s_d = 0.148$.

Using eq. (2.7) for pooling the variances of the k pairs, we then obtain eq. (2.8). Since this is a pooled standard deviation it is subject to the same assumptions, namely that the variance is the same for all samples analyzed. It should be noted that the number of degrees of freedom on which s_d is based is not equal to $2k$ as might be thought at first, but to k.

2.1.4.5 Range and its relation to the standard deviation

The *range*, R, of a set of measurements is the difference between the highest and the lowest value. Taking the lowest and highest value means that one implicitly orders the data in order of numerical value. Statistics applied to ordered data are called *order statistics* and the range is therefore such an order statistic. Another application of order statistics is shown in Chapter 3 where data are ordered to apply the graphical test for a normal distribution. A good approximation of s and therefore an estimation of σ can be obtained by dividing the range by a constant d_n or d_2 (the habitual symbols in the statistical process and the quality control literature).

TABLE 2.5

The coefficient d_2 (or d_n) (adapted from [9])

No. of sets (k)	Number of replicates in a set (n)							
	2	3	4	5	6	7	8	10
1	1.41	1.91	2.24	2.48	2.67	2.82	2.95	3.16
3	1.23	1.77	2.12	2.38	2.58	2.75	2.89	3.11
5	1.19	1.74	2.10	2.36	2.56	2.73	2.87	3.10
10	1.16	1.72	2.08	2.34	2.55	2.72	2.86	3.09
∞	1.13	1.69	2.06	2.33	2.53	2.70	2.85	3.08

$$s = \frac{R}{d_2} = \frac{R}{d_n} \tag{2.9}$$

The values of d_2 [7] are given in Table 2.5. The value of d_2 depends on the number k of sets of data used to determine the range and on the number of replicates in the set, n. The value for $k = \infty$ is sometimes called d_n or Hartley's constant.

Let us suppose that the following results have been obtained in chronological order: 2.3, 2.8, 2.2, 2.9, 2.7, 2.4. The lowest value is 2.2 and the highest 2.9 and a single set of data was obtained. The range is therefore 0.7 and s estimated from the range, using eq. (2.9) is therefore 0.7/2.67 = 0.26. The estimation with eq. (2.2) would have yielded $s = 0.29$.

The range is rarely used for $n > 15$. It is useful to note that for $n = 3$ to 10, d_n is very close to $\sqrt{n}$ so that a rapid approximation of s can be obtained by

$$s = R/\sqrt{n} \qquad (3 \leq n < 10) \tag{2.10}$$

In the same way as we can pool variances to obtain a common estimate of the standard deviation from a set of the standard deviations, we can pool ranges. When there are k sets of n data, then:

$$\bar{R} = (1/k) \sum^{k} R_j$$

where R_j is the range of the jth set of data. $\bar{R}$ is called the *average range* and σ, the pooled standard deviation, can be estimated from

$$s = \bar{R}/d_2 = \bar{R}/d_n \tag{2.11}$$

An example of this calculation is given in Table 2.6. Since $\bar{R} = 0.565$ and $n = 4$, it follows that $s = 0.565/2.06 = 0.27$.

R is used less often than s, because it is more vulnerable to extreme values. However, in routine work, for instance in quality control, and for small samples it is used rather often.

TABLE 2.6

Computation of the average range for a characteristic of a product. Four replicates of each sample are measured.

Sample	Concentration				Mean	Range
	(1)	(2)	(3)	(4)		(R_j)
1	9.6	9.8	10.2	9.9	9.875	0.6
2	10.0	10.2	10.0	10.4	10.15	0.4
3	9.3	10.1	9.6	9.9	9.725	0.8
4	9.4	9.9	9.9	9.5	9.675	0.5
5	10.3	9.8	10.1	10.2	10.1	0.5
6	9.9	10.4	10.0	9.9	10.05	0.5
7	10.3	10.4	10.4	10.1	10.3	0.3
8	10.0	9.7	10.0	9.8	9.875	0.3
9	9.9	9.7	9.3	10.0	9.725	0.7
10	10.4	9.5	10.0	9.3	9.8	1.1
11	10.1	10.4	10.2	10.0	10.175	0.4
12	10.2	9.9	10.3	9.9	10.075	0.4
13	9.6	9.8	10.2	10.2	9.95	0.6
14	10.2	10.1	10.0	9.9	10.05	0.3
15	10.3	9.9	10.2	9.5	9.975	0.8
16	9.7	9.9	9.9	10.2	9.925	0.5
17	10.1	9.6	10.1	9.6	9.85	0.5
18	9.6	9.5	10.3	10.4	9.95	0.9
19	9.9	9.8	10.2	10.4	10.075	0.6
20	9.8	10.1	10.2	10.4	10.125	0.6
				Grand mean	= 9.97	$\bar{R}$ = 0.565

2.2 Measurement of quality

2.2.1 Quality and errors

Quality assurance has been defined [8] as a system of activities whose purpose is to provide the producer or user of a product or service with the assurance that it meets defined standards of quality with a stated level of confidence. A process must yield a product with certain characteristics within certain error margins and quality in measurement is obtained if the stated result approaches the true result closely enough, i.e. is not subject to an error larger than that considered acceptable.

A negative definition would therefore be that process quality is to avoid process and measurement errors larger than a given and accepted level and to do that all the time. Clearly, therefore, we must study errors and types of errors.

TABLE 2.7

Types of errors as illustrated by a set of simulated data

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	$\bar{x}$
A	102	98	101	99	100	103	97	100
B	103	97	101.5	98.5	100	104.5	95.5	100
C	112	108	111	109	110	113	107	110
D	102	98	101	99	100	103	125	104
E	103	102	101	100	99	98	97	100

2.2.2 Systematic versus random errors

Suppose the correct value for the result of a process or a measurement is known to be 100. In Table 2.7 several possible sets of replicate results are given. Situations A and B yield the correct mean but the individual results show a dispersion around that mean. One says that the individual results are subject to *random error.* In situation A the dispersion is less than in situation B. Process or measurement A shows more quality. In statistical process control (see Section 2.3.1) one would say that the process capability index for dispersion is better for A than for B, in chemical analysis and in measurement science in general one would say that *precision* (or one of its components, repeatability or reproducibility; see Chapter 13) is better. In situation C all results are clearly too high. There is a *systematic error.* This systematic error is accompanied by some dispersion around the observed mean: there is at the same time a random error. Systematic error is always combined with random error. Much of statistical hypothesis testing is basically needed to make a difference between systematic and random effects: is the observed difference between categories systematic or due to random effects? In statistical process control one describes systematic errors with the *capability index for setting* (see Section 2.3.2) and in metrology one says that there is a *bias* (see Section 2.5) in the measurement result. The term bias is also used in statistical process control. Situations D and E will be discussed in Section 2.6.

2.3 Quality of processes and statistical process control

Statistical process control or *SPC* is concerned with the situation of the probability distribution of a parameter describing a product relative to the *tolerance limits.* Industry tries to develop processes that produce products within the tolerance limits all the time and preferably with a large margin. Such processes are called *capable.* The population of the objects produced should ideally be centered

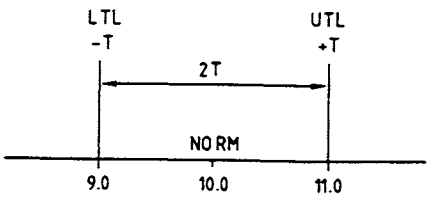


Fig. 2.3. Tolerance limits around a target value (NORM).

around the *norm* or *target* and have a dispersion such that a negligible part of the production falls outside the tolerance limits and must be rejected.

Suppose one produces a bottled product, the concentration of which should be 10.0 (for ease of notation, we will not specify the concentration units), but accepts that it may have values as low as 9.0 and as high as 11.0, i.e. a tolerance T of 1.0 on either side. The lower tolerance limit (LTL) is 9.0 and the upper tolerance limit (UTL) is 11.0 (see Fig. 2.3). Batches with contents outside those limits must be rejected: 10.0 is then the norm and 9.0 and 11.0 the tolerance levels. The difference between the norm and the actual average of the population is a systematic error and should be as small as possible. The dispersion of the population (i.e. the variation due to random error) should be as small as possible. In SPC performance criteria called *process capability indexes* are used to measure both types of errors and to relate them to the *tolerance interval* (2.0 in our example). This interval goes from LTL = NORM – T to UTL = NORM + T, so that it is 2 T wide. In the following sections we will describe the process capability indexes that are used most often. We will follow the terminology as used by Oakland [9] to do this.

2.3.1 Process capability indexes for dispersion

Suppose the probability distribution of a process is normal. The normal distribution is described in more detail in Chapter 3. Suppose also that it is exactly centered around the NORM, then if

$$2T \geq 6\sigma \tag{2.12}$$

only a very small amount of batches ($\leq$0.26%) will need to be rejected (see Fig. 2.4). In SPC, we often use the quantity Cp (process capability index) or CI to characterize the magnitude of σ compared to T

$$Cp = 2T/6\sigma \tag{2.13}$$

If σ is not known, it must be estimated and in SPC we usually apply $\overline{R}$ for this purpose (see eq. (2.11))

$$Cp = 2T\mathrm{d}_n / 6\overline{R} \tag{2.14}$$

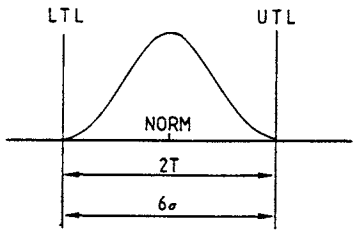


Fig. 2.4. A process with $Cp = 1$.

When $Cp > 1$, at least 3 standard deviations on either side of the mean fall within the tolerance limits (at least when the setting of the process, see Section 2.3.2, is correct). For a normally distributed population this means (see Chapter 3) that at least 99.74% of the objects fall inside the tolerance intervals. It is customary to evaluate Cp as follows:

- $Cp \geq 1.33$: reliable or stable situation. However, some companies require Cp to approach 2. By requiring such strict control on variability shifts in setting may occur without immediately causing values to fall outside the tolerance intervals (see further).
- $1.33 > Cp \geq 1$: control of setting required. Small changes in the setting may lead to a rapid increase of proportion outside the tolerance interval
- $1 > Cp \geq 0.67$: unreliable situation
- $Cp < 0.67$: unacceptable.

In the example of Table 2.6, for $T = 1.0$, $Cp = 2 \times 2.06/6 \times 0.565 = 1.22$. This process would be qualified as requiring control.

Another index of dispersion is the *relative precision index* (RPI). Equation (2.14) can be rewritten for $Cp \geq 1$, i.e. 3 σ limits:

$$2T \geq 6\bar{R}/d_n \text{ or}$$

$$2T/\bar{R} \geq 6/d_n \qquad (2.15)$$

$2T/\bar{R}$ is the RPI. For the example of Table 2.6 with $n = 4$, the RPI is $2 \times 1.0/0.565 = 3.54$.

Since $3.54 \geq 6/2.06$, the condition is met to ensure that 3 σ units on either side of the mean fall within the tolerance interval, at least when there is no systematic error or bias.

2.3.2 Process capability index for setting

This index describes how different the measured mean $\bar{x}$ is from the required mean (= NORM). $\bar{x}$ estimates the true mean μ and $\bar{x}$ – NORM is the estimated bias (see Fig. 2.5)

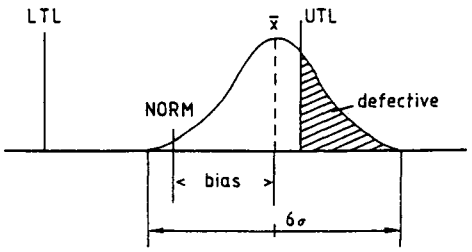


Fig. 2.5. Effect of bias on a process with $Cp = 1$ ($2T = 6\sigma$).

$$C_A = \frac{|\text{NORM} - \bar{x}|}{T} \cdot 100 \tag{2.16}$$

C_A is sometimes called the *index of accuracy*. The term accuracy requires some comment. In this context accuracy measures the systematic error. For a long time, metrologists and analytical chemists have used the term with that same meaning. However, ambiguity was introduced by the ISO definition of the term. According to ISO, accuracy describes the sum of systematic and random errors. For this reason, we no longer use this term when describing systematic errors and prefer the term capability index for setting.

When $C_A \leq 12.5$ the setting of the process is considered reliable, for $12.5 < C_A \leq 25$ control is required, for $25 < C_A \leq 50$ the process is considered unreliable and for $C_A > 50$ it is unacceptable. In the case of Table 2.6:

$$C_A = \frac{|10.0 - 9.97|}{1.0} \times 100 = 3$$

and the setting of the process is reliable.

2.3.3 Process capability indexes for dispersion and setting

The overall *quality index Cpk* is given by

$$Cpk = \frac{\text{distance of } \bar{x} \text{ to the nearest tolerance limit}}{3\sigma} \tag{2.17}$$

Also

$$Cpk = \frac{T - |\bar{x} - \text{NORM}|}{3\sigma} = \frac{T - |\bar{x} - \text{NORM}|}{T} \cdot \frac{T}{3\sigma}$$

$$= Cp\left(1 - \frac{|\bar{x} - \text{NORM}|}{T}\right)$$

or

$$Cpk = Cp\left(1 - \frac{C_A}{100}\right) \tag{2.18}$$

When $C_A = 0$, $Cpk = Cp$. Cpk is also called the corrected process capability index (Cp corrected with C_A). We can show that if $Cpk = 1$ the number of defectives is between 0.13% (when objects falling outside the tolerance levels, all have either too large or else all too small values) and 0.26% in the improbable case that exactly as many bad objects have too low and too high values. For $Cpk \geq 1.33$ the number of defectives is $\leq 0.006\%$. When $Cpk < 1$ the number of defectives is $> 0.13\%$, which is deemed too much and the process is not considered *capable* of achieving the tolerance specifications.

2.3.4 Some other statistical process control tools and concepts

There are seven basic graphical tools [10] in SPC or SQC (statistical quality control), namely the flow chart describing the process, histograms (see Section 2.1.3) to describe the distribution of occurrences, correlation charts (also known as scatter diagrams) (see Chapter 8), run and control charts (see Chapter 7), the cause–effect diagram and the Pareto diagram. The latter two will be discussed here in somewhat more detail.

The cause–effect diagram is also called the fishbone diagram or Ishikawa diagram. Ishikawa is the inventor of this diagram, which he used to introduce process control, i.e. the control of all factors that have an influence on the quality of the product. It often takes the form of Fig. 2.6, and then includes

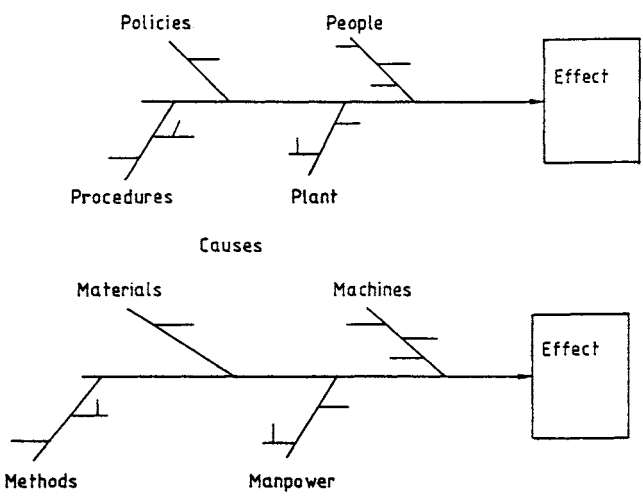


Fig. 2.6. Ishikawa diagrams: the four P and four M versions.

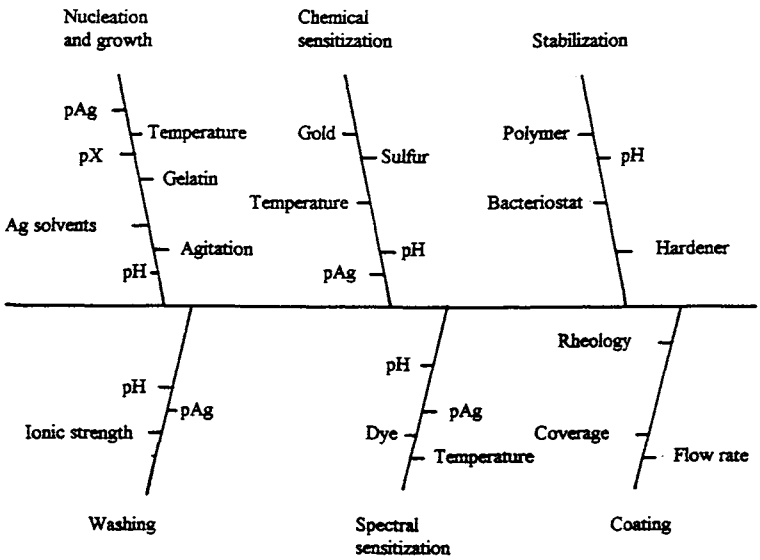


Fig. 2.7. Ishikawa diagram showing process steps and critical parameters for a photographic emulsion manufacturing process (from [11]).

technical, organizational and human factors, but can also be devoted to entirely technical aspects such as in Fig. 2.7 [11]. One identifies major groups of causes that can introduce variation in a process or produce a problem (the effect). Around these major causes one identifies more detailed causes that relate to the major groups.

The Pareto diagram (Fig. 2.8) is also called after its inventor, Vilfredo Pareto, an economist who formulated the 80/20 rule. This Pareto diagram should not be confused with the Pareto optimality diagram or plot described in Chapter 26. Pareto concluded that 20% of the population owns 80% of the wealth and in the same way quality control researchers concluded that 20% of the causes that influence a process often produce 80% of the variation in that process. The chart consists of a histogram listing in order of importance the causes of variation or non-compliance to the requirements and is used as a tool to focus attention on the priority problems. The Pareto chart is sometimes used, too, in connection with latent variable methods (see Chapters 17, 35–36, etc.). Latent variables describe decreasing amounts of variance present in the original data and it is often found that the variance present in a data set characterized by many (hundreds or even more than one thousand) variables can be explained by very few latent variables (much less than 20% in many cases). The amount of variance explained by the so-called first latent variables can then be depicted with a Pareto chart. An example is given in Section 36.2.4.

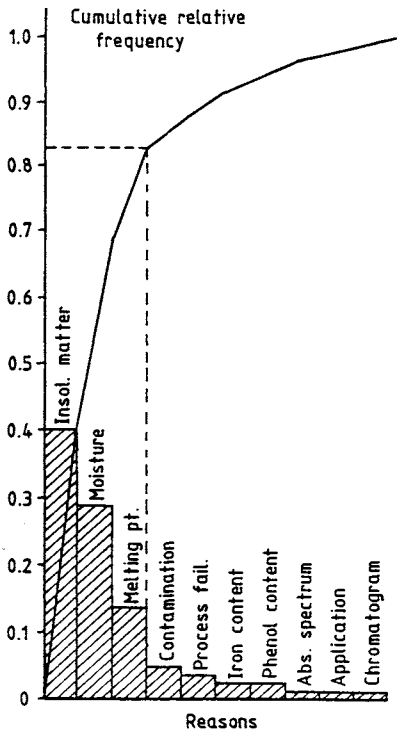


Fig. 2.8. A Pareto diagram. The bar diagram lists reasons for rejection of a batch of dyestuff and the relative frequency of these reasons. The line is the cumulative relative frequency (adapted from [9]).

2.4 Quality of measurements in relation to quality of processes

When we measure the quality of a process, the dispersion of the results is due to two sources, namely, the dispersion due to the process and the dispersion due to the errors in the measurement. We would like to be able to state that the latter source is comparatively small, so that we can conclude with little error that the dispersion observed is due only to the process. Let us therefore try to relate the quality of the measurement to the quality of the process and to the required quality.

Before we do so, we should note that there are two different problems. The one we will discuss here is how good the quality of the measurement should be to control the specifications of a product. The second problem, which will be treated in Chapter 20 is what the characteristics of a measurement process should be to allow control of the process. This characteristic, called measurability, includes also time aspects.

Variances from independent sources are additive (see further Section 2.7) and therefore:

$$\sigma_t^2 = \sigma_p^2 + \sigma_m^2$$

where σ_t^2 is the total variance, σ_p^2 the variance due to the process and σ_m^2 the variance due to the measurement.

If, for instance $\sigma_m = 0.1\ \sigma_p$, then $\sigma_t = 1.005\ \sigma_p$ or, in other words, the contribution of the measurement in the total dispersion is 0.5%. Let us now relate this to tolerance limits in the same way as we did for *Cp* and call this the measurement index [12], *MI*.

$$MI = \frac{2T}{6\sigma_m} \qquad Cp = \frac{2T}{6\sigma_p}$$

MI for $\sigma_m = 0.1\ \sigma_p$ would be equal to 10 for an acceptable process with $Cp = 1$. For such acceptable processes *MI*s of 5 to 3 are usually required. $MI = 3$ for $Cp = 1$ means that $\sigma_m = 0.33\ \sigma_p$ and $\sigma_t = 1.053\ \sigma_p$, so that the contribution of the measurement to total dispersion is about 5%. As an example, let us suppose that we have to develop a measurement method to follow a process. The relative tolerance interval of the process is specified to be 2% and we envisage the use of a titration method with relative standard deviation 0.15%. Does this reach the $MI = 3$ standard? Since $MI = 2/0.9 = 2.22$, the answer is that since $MI < 3$ the method is not sufficiently precise. We should use another method or else carry out each titration in duplicate. Indeed, in that case the mean of the two determinations would be used and the standard deviation on that mean would be $0.15\%/\sqrt{2} = 0.106\%$ and $MI = 3.14$.

2.5 Precision and bias of measurements

The purpose of chemical measurement is in principle to find the true value of a chemical quantity parameter, such as concentration. ISO [4] defines *true value* as: "The value which characterizes a quantity perfectly defined in the conditions which exist at the moment when that quantity is observed (or the subject of a determination). It is an ideal value which could be arrived at only if all causes of measurement error were eliminated and the population was infinite".

As explained in Section 2.2.2, there are two reasons why an analytical result should deviate from the true value, namely the occurrence of either random or systematic errors. When a single analytical result x_i is obtained it differs from the true value μ_0. The difference is the error:

$$e_i = x_i - \mu_0 \tag{2.19}$$

If more measurements are made, i.e. a sample from the population of measurements is obtained, then a mean $\bar{x}$ can be computed for that sample of measurements.

This $\bar{x}$ estimates μ, the mean of the population of measurements and, if the sample is large enough, one can state that $\bar{x} = \mu$. One can then split up eq. (2.19) as

$$e_i = (x_i - \mu) + (\mu - \mu_0) \tag{2.20}$$

The first part, $x_i - \mu$, is the random error of x_i and $\mu - \mu_0$ is the systematic error which will be present in all measurements and therefore also in x_i. Systematic errors lead to inaccuracy and bias, random errors to imprecision. The *accuracy of the mean* is defined by ISO [4] as: "The closeness of agreement between the true value and the mean result which would be obtained by applying the experimental procedure a very large number of times. The smaller the systematic part of the experimental errors which affect the results, the more accurate is the procedure".

A IUPAC document [14] defines at the same time the *bias* as follows: "A measure of the accuracy (or inaccuracy) of the limiting mean is the bias" and "The difference between the population mean and the true value, paying regard to sign". In other words accuracy is the concept, bias the measure.

The latter term can be further specified and in this context *laboratory bias* and *method bias* [15] are of importance. They are discussed further in Chapter 13.

The *precision* is defined as follows [13]: "The closeness of agreement between the results obtained by applying the experimental procedure several times under prescribed conditions. The smaller the random part of the experimental errors which affect the results, the more precise is the procedure".

The measure of precision in analytical chemistry is the standard deviation. Results are often expressed as a relative standard deviation. It should be noted that the experimental standard deviation s estimates σ, the true value of the precision. When the number n of replicate measurements is large enough one can consider that $s = \sigma$. When n is "large enough" it is customary to replace s by σ in equations. What is meant by "large enough" is somewhat subjective. IUPAC [14] stated that one can use σ for $n \geq 10$, but in most textbooks the limit is situated at $n \geq 25$ or $n \geq 30$. According to the exact conditions used, two types of precision are distinguished. They are called *repeatability* and *reproducibility* and are discussed in detail in Chapters 13 and 14.

2.6 Some other types of error

In the previous sections we have focused on systematic and random errors. There are some other sources of error which must be considered, namely:

- *Spurious errors*, leading to *outliers* or *aberrant values*. It may happen that through a wrong operation, going from a wrong setting of an instrument to transcription errors, an atypical value is obtained. This error is clearly neither random nor systematic in nature. In Table 2.7 value 125 in series D is clearly

an outlier. Outliers would falsify the estimation of parameters such as the mean and the standard deviation, and at the same time lead to non-useful descriptions of random and systematic error. From a statistical point of view, outliers must be detected (see Section 5.5), removed and/or one must work with methods resistant to outliers (robust methods, see Chapter 12). From the quality management point of view outliers must be prevented. When an outlier is found, it should be flagged and, when possible, the reason should be ascertained.

- *Drift*, indicative of a process that is not under (statistical) control. This is exemplified in situation E of Table 2.7. If these data were obtained in chronological order then one would conclude that there is a downwards drift. It is an unstated, but always present, hypothesis that processes described in a statistical way, are under *statistical control*. This means that the mean setting and the dispersion of the result are assumed to be constant. The process yielding the data of E would therefore not be under control. When this is the case, it makes no sense to compute statistical parameters. As quality management to a large extent consists of restricting variation, it is evident that drift must be avoided. Drift is not the only type of error occurring when a process is not under control. How to detect them is described in Chapter 7 on quality control.
- *Baseline noise*. While all the other sources of errors are equally relevant for processes and for measurements, this source of error affects only chemical analysis results. Measurements often result from a difference between a signal obtained when the analyte is measured and a signal obtained for a blank. There are different types of *blank* (see Section 13.7.4), but for the moment we can define the blank as consisting of the same material but without the analyte. Both signals show variation due to random error. The variation for the blank signal is also called the *baseline* or *background noise*. The measured signal is due to the analyte plus the baseline noise and, when the signal due to the analyte becomes very small, its contribution cannot be distinguished from that of the baseline. It is then no longer possible to state a result for the analyte. One says that its concentration is below the *detection limit* (and should never say that it is equal to 0). The detection and related limits are discussed in Chapter 13.

2.7 Propagation of errors

When the final result is obtained from more than one independent measurement, or when it is influenced by two or more independent sources of error (for instance measurement and process — see Section 2.4), these errors can accumulate or compensate. This is called the *propagation of errors*.

Random errors accumulate according to the law of propagation of errors given by:

$$\sigma_z^2 = \left(\frac{\partial z}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial z}{\partial x_2}\right)^2 \sigma_{x_2}^2 \tag{2.21}$$

where $z = f(x_1, x_2)$ and x_1 and x_2 must be *independent* variables.

For instance, for the sum of two variables, $z = x_1 + x_2$, eq. (2.21) can be written as:

$$\sigma_{x_1+x_2}^2 = \left(\frac{\partial(x_1 + x_2)}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial(x_1 + x_2)}{\partial x_2}\right)^2 \sigma_{x_2}^2$$

$= 1\sigma_{x_1}^2 + 1\sigma_{x_2}^2$ or

$$\sigma_{x_1+x_2}^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 \tag{2.22}$$

One can verify that:

$$\sigma_{x_1-x_2}^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 \tag{2.23}$$

$$\sigma_{ax_1+bx_2}^2 = a^2\sigma_{x_1}^2 + b^2\sigma_{x_2}^2 \tag{2.24}$$

$$(\sigma_{x_1x_2}/x_1x_2)^2 = (\sigma_{x_1}/x_1)^2 + (\sigma_{x_2}/x_2)^2 \tag{2.25}$$

$$(\sigma_{x_1/x_2}/(x_1/x_2))^2 = (\sigma_{x_1}/x_1)^2 + (\sigma_{x_2}/x_2)^2 \tag{2.26}$$

$$\sigma^2 a \log x = (a\,\sigma_x/x)^2 \tag{2.27}$$

In other words, variances are additive as such when the operation is addition or substraction or as squared relative standard deviations when the operation is multiplication or division. It must be stressed that these equations are correct only when the variables are independent, i.e. not correlated. This is often not the case in practical situations.

Systematic errors are propagated with their signs. If Δz is the systematic error affecting z, then for an additive/substractive relationship:

$$z = a + bx_1 + cx_2 - \mathrm{d}x_3$$

$$\Delta z = b\Delta x_1 + c\Delta x_2 - \mathrm{d}\Delta x_3 \tag{2.28}$$

where Δx_1, etc. are the systematic errors affecting x_1, etc. While random errors do not compensate but accumulate, systematic errors can compensate.

For a multiplicative relationship:

$$z = ax_1x_2/x_3$$

$$\Delta z/z = \Delta x_1/x_1 + \Delta x_2/x_2 - \Delta x_3/x_3 \tag{2.29}$$

In eq. (2.29) we observe that for this type of relationship relative systematic errors are transmitted.

Equations of the type described in this section have assumed a large importance in metrology, because they allow us to describe individual sources of error and to combine them to describe what is called in metrology language the *uncertainty*. This is defined as a range within which the true value should be found. When the uncertainty is expressed as a standard deviation, it is then called a *standard uncertainty*. When there are several sources of error the *combined standard uncertainty* is obtained using the law of propagation of errors (for metrologists: of uncertainties). Eurachem [15], which aims to introduce this terminology in analytical chemistry, states that in analytical chemistry in most cases the *expanded uncertainty* should be used. This defines an interval within which the value of the concentration of an analyte (the *measurand*) is believed to lie with a particular level of confidence (see Chapter 3). It is obtained by multiplying the combined standard uncertainty by a *coverage factor*, k. The choice of the factor k is based on the level of confidence desired. For an approximate level of confidence of 95%, k is 2. The uncertainty concept will not be used further in this book.

2.8 Rounding and rounding errors

Because of the existence of error, not all computed figures are significant and, in principle, they should be rounded. A frequent question is how many figures should be retained. According to the significant figure convention [16], results should be stated so that they contain the figures known with certainty and the first uncertain figure. When carrying out a measurement, such as reading a value on a pH display graduated in tenths of pH, one would write 7.16, because the needle on the display is between 7.1 and 7.2 and the best guess at the next figure is 6.

When the number is the result of a computation, the following rules may be useful [16]:

- After adding or substraction, the results should have the same number of significant numbers after the decimal point as the number being added or subtracted which has the fewest significant figures after the decimal point.
- After multiplication or division, the number of significant figures should equal the smallest number carried by the contributing values.
- When taking a logarithm of a number, we should give as many figures after the decimal point as there are significant figures in the original number.

These rules should be applied in a sensible way. For instance, if in a set of data a certain number is correctly rounded to 99.4 and one has to multiply it with 1.01, yielding 100.394, it should be rounded to 100.4 and not to 100 as one of the above rules would require.

These are the rules that should be applied for the numbers as they are reported. However, during the computations one should not round numbers, because this can lead to *rounding errors*. Ellison et al. [17], who studied the effect of calculator or computer precision and dynamic range concluded that even today finite numerical precision, aggravated by poor choice of algorithm, can cause significant errors. This is one of the reasons why, in this book, we sometimes report more significant figures then we should really do. These numbers are often intermediate results, which are used for further computations, sometimes in another chapter. Rounding them correctly would lead to small discrepancies in the result, when the whole computation is checked with a computer.

References

1. W. Horwitz, Nomenclature for sampling in analytical chemistry. IUPAC, Pure Appl. Chem., 62 (1990) 1193–1208.
2. R.R. Sokal and F.J. Rohlf, Biometry. W.H. Freeman, New York, p. 11, 1981.
3. R. Cleymaet and D. Coomans. Construction of a data set for the investigation of the relationship between the chemical constitution of tooth enamel and oral diseases. Internal Report 1, 1991, TI-VUB, Brussels.
4. ISO 3534-1977 (E/F), Statistics — Vocabulary and Symbols, 1977.
5. G.T. Wernimont, Use of Statistics to Develop and Evaluate Analytical Methods. AOAC, Arlington, p. 33, 1985.
6. G.T. Wernimont, p. 37 in ref. 5.
7. L.S. Nelson, Use of the range to estimate variability. J. Qual. Techn., 7 (1) 1975.
8. J.K. Taylor and H.V. Oppermann, Handbook for the Quality Assurance of Metrological Measurements. National Bureau of Standards, Handbook 145, Gaithersburg, 1986.
9. J.S. Oakland, Statistical Process Control. Wiley, New York (1992).
10. R.A. Nadkarni, The quest for quality in the laboratory. Anal. Chem., 63 (1991) 675A–682A.
11. M.T. Riebe and D.J. Eustace, Process analytical chemistry: an industrial perspective. Anal. Chem., 62 (1990) 65A–71A.
12. A. Van Rossen and L. Segers, Statistical measurement control. Het Ingenieursblad, 1 (1991) 45–50 (in Dutch).
13. W.D. Pocklington, Harmonized protocol for the adoption of standardized analytical methods and for the presentation of their performance characteristics, IUPAC, Pure Appl. Chem., 62 (1990) 149–162.
14. IUPAC, Commission on spectrochemical and other optical procedures for analysis. Pure Appl. Chem., 45 (1976) 99.
15. Eurachem Workshop draft version 5, Quantifying uncertainty in analytical measurement (1994).
16. D. McCormick and A. Roach, Measurement, Statistics and Computation, Analytical Chemistry by Open Learning. John Wiley, Chichester, 1987.
17. S.R. Ellison, M.G. Cox, A.B. Forbes, B.P. Butler, S.A. Hannaby, P.M. Harris and S.M. Hodson, Development of data sets for the validation of analytical instrumentation. J. Assoc. Off. Anal. Chem. 77 (1994) 777–781.

Additional recommended reading

A.J. Duncan, Quality Control and Industrial Statistics, 5th edn. Irwin, Homewood, Illinois, 1986.
D.C. Montgomery, Introduction to Statistical Quality Control, 2nd ed. Wiley, New York, 1991.
T. Pyzdek and R.W. Berger, eds., Quality Engineering Handbook. Dekker, New York, 1992.

Chapter 3

The Normal Distribution

3.1 Population parameters and their estimators

Before discussing the normal distribution as such, we need to enlarge somewhat on the discussion in Chapter 2. Suppose we analyze the concentration of Na^+ in a certain sample. Due to random error, there will be some dispersion of the results of replicate determinations. If we were able to carry out a very large number of such determinations, the results would constitute a population. We would like to know the mean μ and the standard deviation σ of that population, the former to know the true content of Na^+ (assuming there is no systematic error) and the latter to know the precision of the determination. However, it is not possible to carry out so many determinations: let us suppose that $n = 4$ replicate determinations are carried out. The four results constitute a sample in the statistical sense. The mean

$$\bar{x} = (\Sigma x_i)/n \tag{3.1}$$

and the standard deviation

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{(n-1)}} \tag{3.2}$$

of the four replicates are the *sample parameters.* Sample parameters are *estimators* of the population parameters: $\bar{x}$ estimates μ and s estimates σ.

We distinguish the *population parameters* from the sample parameters, also called sample statistics, by writing them as Greek letters (μ, σ). The latter are then written as the corresponding Latin letters (m, s). In most statistical texts, there is one exception, namely the sample mean, which is written as $\bar{x}$. The ISO norms [2] use m instead of μ. We will follow the general practice and use μ. This convention is followed as much as possible in all instances where we want to distinguish between the population parameter and the sample parameters. For instance (see further Chapter 8), a straight-line regression equation is written as:

$$\eta = \beta_0 + \beta_1 x$$

where η is the true response and β_0 and β_1 are respectively the true intercept and slope of the regression line. When we do not know β_0 and β_1 but estimate them from a finite number of points, the estimated parameters are given as

$$\hat{y} = b_0 + b_1 x$$

where b_0 and b_1, the sample parameters, are the estimators of β_0 and β_1, the population parameters; $\hat{y}$ is the estimator of the true response η. Notice that $\hat{y}$ (*y-hat*) is used here as the symbol for the estimator of the true response to distinguish it from y which in regression represents the observed responses (see further Chapter 8).

By considering many random samples of size n from a population and computing a statistic for it, we obtain a distribution of that statistic, called the *sample distribution*. If we carry out many series of four replicate measurements, the means of those sets of four measurements lead to a sample distribution of means for $n = 4$, characterized by its own mean and standard deviation. In this section we are concerned with the mean of such distributions. The standard deviation of this and other sample distributions will be discussed further in Section 3.5.

If the mean of the sampling distribution of a statistical parameter used to estimate a population parameter is equal to that population parameter, then that estimator is called an *unbiased estimator*. The mean $\bar{x}$ is an unbiased estimator of μ. The mean of the sample distribution of the $\bar{x}$ values for sets of $n = 4$ replicate measurements is equal to the population parameter μ. The situation is not so simple for s and σ. As noted in Chapter 2, we divide here by $n - 1$ instead of by n where $n - 1$ is called the number of degrees of freedom (df). The population variance σ^2 is in fact the mean of the squared deviations from μ. At first sight, we should therefore divide by n instead of by $n - 1$ to obtain s^2, the sample variance, and s, the sample standard deviation. However, it can be shown that in this case s^2 would be a *biased estimator* of σ^2. In other words, the mean of the sample distribution of the s^2 values for sets of $n = 4$ replicate measurements would not be equal to the population parameter σ^2. It is to obtain an unbiased estimator of the variance that the term $n - 1$ was introduced in eq. (3.2), although (see further) eq. (3.5) would suggest the use of n. Although $s^2 = (\sum(x_i - \bar{x})^2) / (n - 1)$ is an unbiased estimator of σ^2, s is not an unbiased estimator of σ. It has been shown [1] that s from eq. (3.2) underestimates σ and that the underestimation is a function of n. It is serious only for small sample sizes. The correction factor is 1.253 for $n = 2$, but only 1.064 for $n = 5$ and 1.009 for $n = 30$. Except sometimes when using quality-control charts, it is unusual to correct s to obtain the unbiased estimate. In many application fields, such as in analytical chemistry, it is often stated that $n = 5$ to 8 is needed for a sufficiently good (i.e. precise and unbiased) estimation of s.

3.2 Moments of a distribution: mean, variance, skewness

To summarize the characteristics of a distribution, we can use its *moments*. The *r*th moment of a set of data x_1, ...x_i, ...x_n is equal to

$$m_r^o = \left(\sum_{i=1}^{n} x_i^r \right) /n \qquad \text{or, more briefly: } m_r^o = (\Sigma\, x_i^r)\, /n \tag{3.3}$$

The moment about the mean is defined in the same way with x_i being replaced by $x_i - \bar{x}$. The *rth moment about the mean* or *rth central moment* is therefore:

$$m_r = (\Sigma\, (x_i - \bar{x})^r)\, /n \tag{3.4}$$

It is equal to the average of the deviations of each of the data from the mean to the power *r*.

The dimensionless moment (about the mean) is defined as

$a_r = m_r / s^r$, with s as defined in eq. (3.5)

The first moment of the data, m_1^o, is equal to the *mean*, while the first central moment, m_1, is zero. The mean is one of the descriptors of *central location.*

The second moment about the mean of any distribution (and not only of the normal distribution, as is sometimes thought) is the *variance.* It describes the *dispersion* within the data and its square root is equal to the *standard deviation, s.*

$$m_2 = \Sigma(x_i - \bar{x})^2 /n = s^2 \tag{3.5}$$

The variance is therefore the mean of the squared deviations of the data from the mean. It should be noted here that in this definition the sum of the squared deviations from the mean is divided by n instead of the more usual $n - 1$. The reason for this discrepancy is explained in Section 3.1.

The third moment about the mean is a measure of *skewness*, i.e. the departure of the distribution from symmetry. Distributions such as those in Figs. 3.1a and b, are said to be skewed. The third central moment is used in a dimensionless form by referring it to the standard deviation.

Expressing s as a moment, and since the standard deviation is equal to the square root of variance, we can write

$$a_r = \frac{m_r}{\sqrt{m_2^r}} \tag{3.6}$$

The (*moment*) *coefficient of skewness* is then given by:

$$a_3 = \frac{m_3}{\sqrt{m_2^3}} = \frac{1}{n} \Sigma \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{3.7}$$

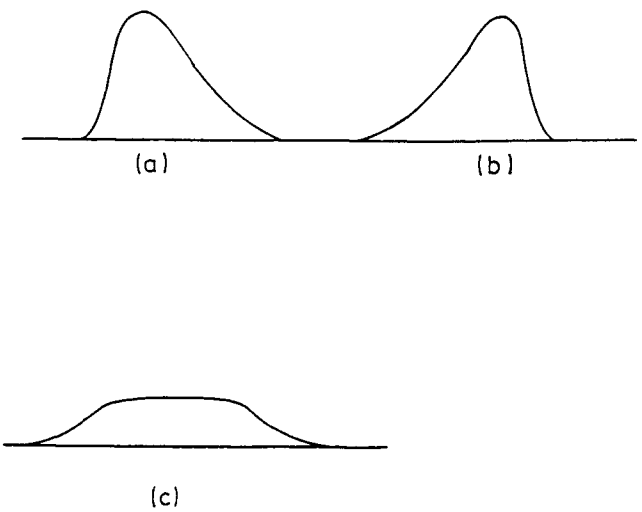


Fig. 3.1. Non-normal distributions: (a) positive skewness, (b) negative skewness, (c) negative kurtosis.

It is also sometimes written as

$b_1 = a_3^2$

When the curve is perfectly symmetric, $a_3 = 0$.

The fourth central moment is used to measure *kurtosis* (also called peakedness; see Fig. 3.1). We often compute

$a_4 = b_2 = m_4 / (m_2^2)$

For a normal distribution $b_2 = 3$. For this reason kurtosis is often defined as $b_2 - 3 = \left(\frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4\right) - 3$ so that kurtosis is then 0 for a normal distribution, positive for a peaked curve (also called leptokurtic) and negative for a flat peak (also called platykurtic).

The equations given in this section are those for a sample distribution. It is also possible to write down the equations for a population distribution.

3.3 The normal distribution: description and notation

The best known probability distribution is the normal distribution. In shorthand notation:

$x \sim N(\mu, \sigma^2)$

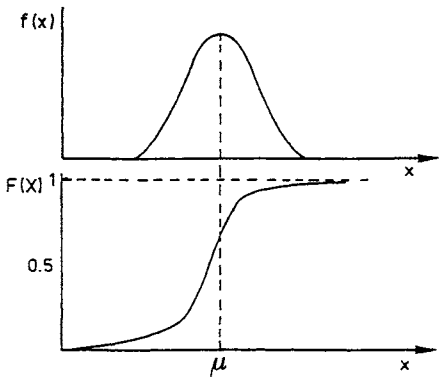


Fig. 3.2. Normal distribution (a) and cumulative normal distribution (b).

This means that the values of x are distributed normally with a mean μ and a variance σ^2. The *normal distribution* or, rather its probability density function, is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} \tag{3.8}$$

and is shown in Fig. 3.2. The factor $\frac{1}{\sigma\sqrt{2\pi}}$ is a normalization factor. It standardizes the area under the curve, so that it is always equal to 1.

The *cumulative normal probability distribution* is given by

$$F(x_0) = \int_{-\infty}^{x_0} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} dx \tag{3.9}$$

and is also shown in Fig. 3.2.

The mean and the standard deviation of such a population are values in a certain scale. For instance, if we were to describe titration results in ml NaOH, the mean would be equal to a certain number of ml NaOH and the standard deviation, too, would have to be described in the same units. To avoid this scale effect, the concept of a *standardized normal distribution* has been developed. The original distribution is transformed by computing

$$z = (x - \mu) / \sigma \tag{3.10}$$

This means that now the original data are described as their deviations from the mean divided by the standard deviation. In other words, the scale used is now a scale in standard deviation units. If a certain number has a $z = 1.5$, this means that its value is higher than that of the mean because it has a plus sign and that its

distance from the mean is equal to 1.5 standard deviation units. This process is often called *standardization, scaling* or *autoscaling*. It is also called the *z-transformation* and z itself is called the *reduced variable* of x or the *standardized deviate* or *standard deviate*. It should be noted that the ISO norms [2] use u instead of z, but because such a large number of statistical texts use z, we have preferred to follow this custom. Since z is normally distributed with as its mean 0 and standard deviation 1, the variance is also 1 and we can write

$$z \sim N(0,1)$$

The probability function for the standardized normal distribution is given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) \tag{3.11}$$

and the corresponding cumulative frequency distribution by

$$\Phi(z_0) = \int_{-\infty}^{z_0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) dz \tag{3.12}$$

3.4 Tables for the standardized normal distribution

All statistical handbooks contain tables of the standardized normal distribution. The principal reason why the z-distribution is used so often is that, because of the standardization, it is possible to use scale independent tables. The main reason for using the tables is that they allow to calculate what proportion of a population has a value smaller or larger than a certain value or is comprised between two boundaries. Since we are interested in areas under parts of the curve and in a probability at a precise point on the z-axis, the tables do not describe the normal distribution as such, but are based on the cumulative distribution.

One of the problems confronting the inexperienced user of statistics is that these tables can be presented in several ways. Let us first note that there are *one-sided* or *one-tailed* and *two-sided* or *two-tailed* tables. The latter tables show what part of the total area falls inside or outside the interval $(-z, +z)$ and usually how much of it falls outside. Such a table is Table 3.1. The table gives the z-value corresponding with certain areas of the sum of the two tails of Fig. 3.3.a. For instance, we can ask what the z-value is such that 5% of all data of a normal distribution will fall outside the range $(-z, +z)$, i.e. 2.5% on each side. In the table we find that for $p = 0.05$, $z = 1.96$. The reason that this is printed in italics is that this value of z will be needed very frequently.

In the same way, we can ask between what values can we find 90% of all values. Then $p = 0.1$ and $z = 1.65$. Suppose that we know that certain titration results are

TABLE 3.1

Values of z and the two-tailed probability that its absolute value will be exceeded in a normal population (see also Fig. 3.3a)

	Second decimal in p									
p	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	∞	2.576	2.326	2.170	2.054	1.960	1.881	1.812	1.750	1.695
0.1	1.645	1.598	1.555	1.514	1.476	1.439	1.405	1.372	1.340	1.311
0.2	1.231	1.254	1.226	1.200	1.175	1.150	1.126	1.103	1.080	1.058
0.3	1.036	1.015	0.994	0.974	0.954	0.935	0.915	0.896	0.878	0.860
0.4	0.842	0.824	0.806	0.789	0.772	0.755	0.739	0.722	0.706	0.690
0.5	0.674	0.659	0.643	0.623	0.613	0.598	0.583	0.568	0.553	0.539
0.6	0.524	0.510	0.496	0.482	0.468	0.454	0.440	0.436	0.412	0.399
0.7	0.385	0.372	0.358	0.345	0.332	0.319	0.305	0.292	0.279	0.266
0.8	0.253	0.240	0.228	0.215	0.202	0.189	0.176	0.164	0.151	0.138
0.9	0.126	0.113	0.100	0.088	0.075	0.063	0.050	0.038	0.025	0.013
p	0.002		0.001	0.0001	0.00001	0.000001	0.0000001		0.00000001	
z	3.090		3.290	3.890	4.417	4.891	5.326		5.730	

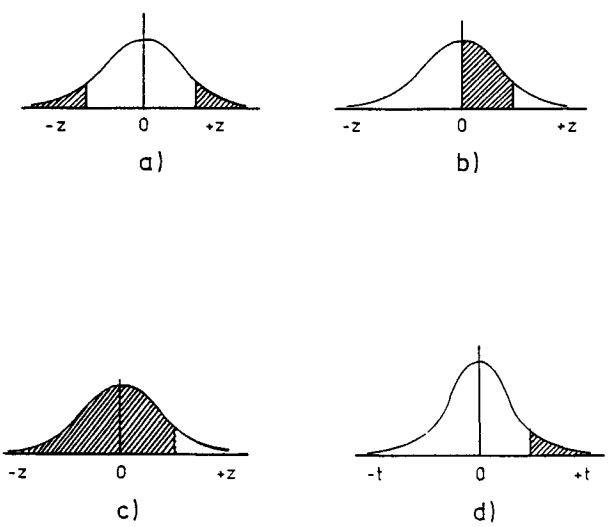


Fig. 3.3. The shaded areas are described by (a) Table 3.1; (b) Table 3.2; (c) Table 3.3; (d) Table 3.4.

normally distributed with $\mu = 5.0$ ml and $\sigma = 0.05$ ml and would like to know in what range 95% of all results will be found. This range is then given by $5.0 \pm 1.96 \cdot 0.05$. In view of what will be discussed in Chapter 4, it is of interest to rephrase the question

TABLE 3.2

Probability p to find a value between 0 and z (see also Fig. 3.3b)

	Second decimal of z										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	
0.0	0.000	0.004	0.008	0.012	0.016	0.020	0.024	0.028	0.032	0.036	
0.1	0.040	0.044	0.048	0.052	0.056	0.060	0.064	0.067	0.071	0.075	
0.2	0.079	0.083	0.087	0.091	0.095	0.099	0.103	0.106	0.110	0.114	
0.3	0.118	0.122	0.125	0.129	0.133	0.137	0.144	0.141	0.148	0.152	
0.4	0.155	0.159	0.163	0.166	0.170	0.174	0.177	0.181	0.184	0.188	
0.5	0.191	0.195	0.198	0.202	0.205	0.209	0.212	0.216	0.219	0.222	
0.6	0.226	0.229	0.232	0.236	0.239	0.242	0.245	0.249	0.252	0.255	
0.7	0.258	0.261	0.264	0.267	0.270	0.273	0.276	0.279	0.282	0.285	
0.8	0.288	0.291	0.294	0.297	0.299	0.302	0.305	0.308	0.311	0.313	
0.9	0.316	0.319	0.321	0.324	0.326	0.329	0.331	0.334	0.336	0.339	
1.0	0.341	0.344	0.346	0.348	0.351	0.353	0.355	0.358	0.360	0.362	
1.1	0.364	0.366	0.369	0.371	0.373	0.375	0.377	0.379	0.381	0.383	
1.2	0.385	0.387	0.389	0.391	0.392	0.394	0.396	0.398	0.400	0.401	
1.3	0.403	0.405	0.407	0.408	0.410	0.411	0.413	0.415	0.416	0.418	
1.4	0.419	0.421	0.422	0.424	0.425	0.426	0.428	0.429	0.431	0.432	
1.5	0.433	0.434	0.436	0.437	0.438	0.439	0.441	0.442	0.443	0.444	
1.6	0.445	0.446	0.447	0.448	0.449	0.450	0.451	0.452	0.453	0.454	
1.7	0.455	0.456	0.457	0.458	0.459	0.460	0.461	0.462	0.462	0.463	
1.8	0.464	0.465	0.466	0.466	0.467	0.468	0.469	0.469	0.470	0.471	
1.9	0.471	0.472	0.473	0.473	0.474	0.474	0.475	0.476	0.476	0.477	
$z =$	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	
$F(z) =$	0.477	0.482	0.486	0.489	0.492	0.494	0.495	0.496	0.497	0.498	
$z =$	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
$F(z)$	0.4987	0.4990	0.4993	0.4995	0.4997	0.4998	0.4998	0.4998	0.4999	0.49995	0.49997

as follows: determine decision limits, beyond which results will be rejected, such that 5% of all values fall outside, half on each side. The answer of course remains the same.

Two other examples of z tables are given in Table 3.2 and 3.3. Because of the symmetry of the normal distribution, these two tables give p-values only for positive z-values. Table 3.2 gives the areas between two boundaries. One boundary is $z = 0$ and the table gives the area between this value of z and the chosen value (see Fig. 3.3b). *Example*: a large number of determinations was carried out on the same sample and the results are known to be normally distributed with $\mu = 215$ and $\sigma = 35$. What percentage of determinations will fall between the boundaries 200 and 250? First we compute the corresponding z-values.

TABLE 3.3

Probability to find a value lower than z (see also Fig. 3.3c)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.625	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.799	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.866	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.892	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.950	0.951	0.952	0.953	0.954
1.7	0.955	0.956	0.957	0.958	0.960	0.961	0.962	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977

$z_1 = (200 - 215)/35 = -0.43$

$z_2 = (250 - 215)/35 = 1$

The area between $z = 0$ and $z = 0.43$ is 0.166 or nearly 17% and between $z = 0$ and $z = 1$ it is 34%. We can conclude that 51% of all data are comprised between 200 and 250.

Table 3.3 is a one-sided table, also called cumulative table. It gives the area below a certain value of z (see also Fig. 3.3c). Suppose that for the same data as given above, we want to know how large the probability is of finding a result above 250. Since z for that value is equal to 1, consultation of the table shows that $p = 0.84$. This is the probability of finding a value lower than $z = 1$. It follows that the probability for values above $z = 1$, (i.e. in this case, values higher than 250) is $1 - 0.84 = 0.16$ or 16%.

Tables 3.1, 3.2 and 3.3 contain the same information and therefore we should be able to use any of them for each of the different examples discussed. For example, let us consider the titration example, with which we illustrated the use of Table 3.1. Table 3.2 covers only half of the normal distribution, i.e. 50% of the values that

occur, so that when all those values are included $p = 0.5$. In this half distribution, it gives the area between the apex of the distribution ($z = 0$) and the decision boundary that delimits the higher tail. Since the area for both tails together is 5%, that for the higher one will include 2.5%. The boundary is thus situated so that 50% – 2.5% = 47.5% ($p = 0.475$) is included between $z = 0$ and the boundary. For $p = 0.475$, one finds $z = 1.96$ as with Table 3.1.

In Table 3.3 we include the whole distribution up to the higher tail. This means that we should determine the z-value that bounds the higher tail. The area up to the higher tail includes 97.5%. The z for which Table 3.3 gives $p = 0.975$ is again 1.96.

3.5 Standard errors

If we take random samples of size n from a population with mean μ and standard deviation σ, then the sample distribution of the means, $\bar{x}$, will be close to normal with mean μ and standard deviation

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} \tag{3.13}$$

$\sigma_{\bar{x}}$ is the standard deviation of the means for samples with size n. It is also called the *standard error on the mean* or SEM. It follows that we can also write:

$$s_{\bar{x}} = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n(n-1)}}$$

where $s_{\bar{x}}$ estimates $\sigma_{\bar{x}}$. The approximation of normality will be better when n increases, but the approximation is quite good, even for small n. It should be emphasized that the population from which the samples are taken to obtain the means need not be normal. In Fig. 3.4, two clearly non-normal distributions are given. Taking samples of size n from these distributions will lead to the normal distributions of the means of the n results shown in the figure. The distribution of means of n individual non-normally distributed data will approach the normal distribution better when the sample size n increases. When rigorously stated, this is known as the *central limit theorem.*

The distribution of the sample means becomes progressively sharper when the sample size n is increased: the means of samples conform more to the mean (i.e. estimate better the mean) for larger n. This is shown in Fig. 3.5 where the distribution of samples of $n = 1$ (i.e. individual measurements), $n = 4$ and $n = 9$ from the same population $N(\mu, \sigma^2)$ are compared. The mean of the three distributions is μ. The standard deviation for $n = 4$ and $n = 9$ is respectively $\sigma/\sqrt{4} = \sigma/2$ and $\sigma/\sqrt{9} = \sigma/3$.

It should be noted that $s_{\bar{x}}$ or $\sigma_{\bar{x}}$ should not be used as measures of dispersion to evaluate the precision of a measurement or the capability of a process. We must then use s or σ, since we are interested in the dispersion of individual results; $s_{\bar{x}}$ gives an idea, however, about the confidence we can have in the mean result (see Section 3.6).

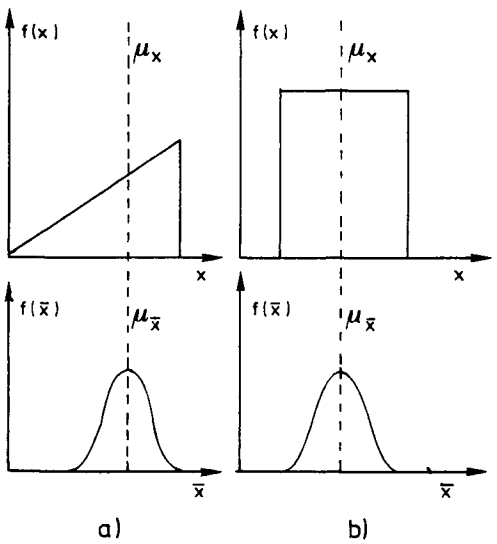


Fig. 3.4. Means of samples taken from the non-normal triangular (a) and rectangular (b) distributions are normally distributed.

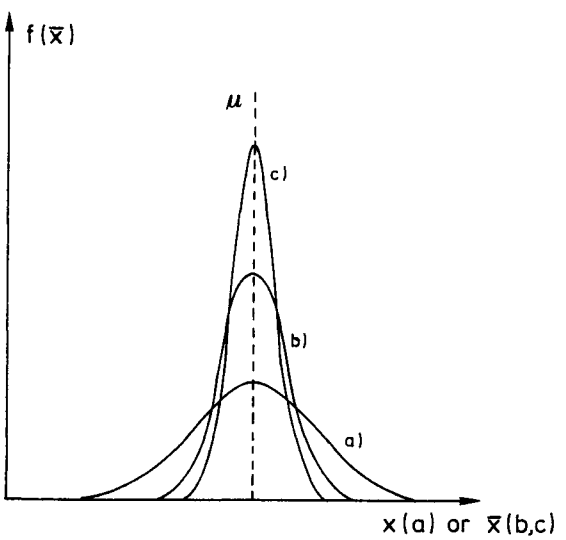


Fig. 3.5. The sharpness of a distribution of means depends on the sample size. Distribution (a) is the population distribution; (b) samples of $n = 4$ from (a); (c) samples of $n = 9$ from (a).

The standard deviation of a sample distribution is often called a *standard error.* In this section we have studied the sample distribution of means and the standard deviation of that distribution is the standard error of the mean. This can be applied

to other sample distributions. For instance, we could have determined the sample distribution of the standard deviation of samples with size n. For a normal distribution of the original population, this would then have yielded a standard error of the standard deviations, $\sigma_s = \sigma/\sqrt{2n}$.

3.6 Confidence intervals for the mean

In Section 3.4, it was computed that for a normal distribution, 95% of the data (or 95% of the area under the curve) fall within the limits $z = -1.96$ to $z = +1.96$. This can be rephrased to state that 95% of the data fall within the limits $\mu \pm 1.96\sigma$. This is true for all normal distributions and, since sample means are normally distributed, it is true also for the distribution of means. We can state, therefore, that 95% of all sample means of size n must fall within the limits

$$\mu \pm 1.96\sigma/\sqrt{n}$$

Suppose we take a sample of size n from a population, we carry out the n measurements and compute $\bar{x}$. This $\bar{x}$ is an estimator of μ, the population mean. Suppose also that the standard deviation, σ, is known (how to proceed when σ is not known is explained in Section 3.7). There is then a probability of 95% that $\bar{x}$ will fall in the range $\mu \pm 1.96\ \sigma/\sqrt{n}$ (see Fig. 3.6). The statement

$$\mu - 1.96\ \sigma / \sqrt{n} < \bar{x} < \mu + 1.96\ \sigma / \sqrt{n} \tag{3.14}$$

is therefore correct in 95% of cases. This type of statement will be written in future as $\bar{x} = \mu \pm 1.96\ \sigma/\sqrt{n}$. It should be noted that this is considered to mean that $\bar{x}$ lies in the interval $\mu - 1.96\ \sigma / \sqrt{n}$ to $\mu + 1.96\ \sigma / \sqrt{n}$ and not that it is equal to one or both of these boundaries. It follows from (3.14) that

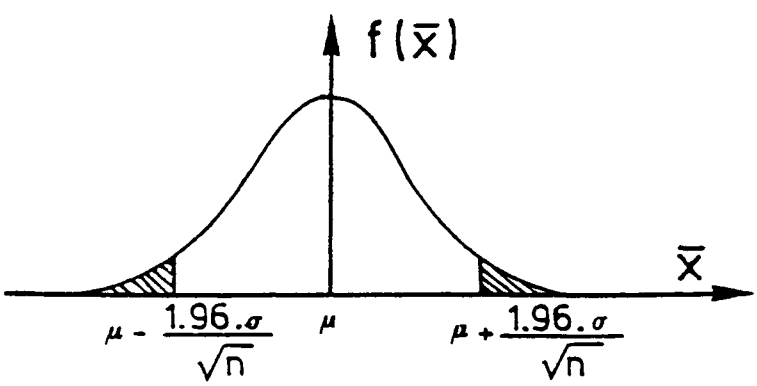


Fig. 3.6. There is 5% probability (the shaded area) that $\bar{x}$, the mean of a sample of n results, has a value more than $1.96 \cdot \sigma/\sqrt{n}$ distant from μ.

$$\mu = \bar{x} \pm 1.96\,\sigma/\sqrt{n} \tag{3.15}$$

is also correct in 95% of cases. This means that we can estimate μ, which is unknown, by determining $\bar{x}$ for n measurements and at the same time describe the uncertainty of that estimate by writing eq. (3.15). In 95% of all cases the resulting statement will be correct. In general:

$$\mu = \bar{x} \pm z\sigma/\sqrt{n} \tag{3.16}$$

with $100 - \alpha$% probability or *confidence*, where α is derived from a z-table. For instance,

$\mu = \bar{x} \pm 1.645\sigma/\sqrt{n}$ with 90% confidence.

The limits in eq. (3.16) are called the *confidence limits* (for instance, with z = 1.96, the 95% confidence limits). The range between the limits is called the *confidence interval*. Confidence limits or intervals can be stated in %, or as fractions. A confidence of 90% is equivalent to one of 0.90.

Suppose now that a certain material has been analyzed and that a result has been obtained of 10.10 ± 0.10, where the ± 0.10 describes the 95% confidence interval. In other words, 10.10 is an estimate of the unknown μ and there is 95% probability that the interval 10.00 to 10.20 contains μ. It is possible that the analyst is not happy with this result because he wants the 95% confidence interval to be smaller, say ± 0.05. How can this be achieved? The 0.10 was computed as

$$1.96\,\frac{\sigma}{\sqrt{n}} = 0.10$$

The standard deviation σ is typical of the measurement process. It is the population standard deviation and therefore a constant for that population. The only thing which can be changed is n. Let us call the sample size to obtain the smaller confidence limits, N. Then

$$1.96\,\frac{\sigma}{\sqrt{N}} = 0.05$$

It follows that $N = 4\,n$.

By increasing the sample size, we can narrow the confidence limits. Because of the dependence on $\sqrt{n}$, the n required to obtain certain confidence limits may of course be impractical in some experimental situations. Nevertheless, this simple example demonstrates that by choosing a correct sample size, the confidence interval can, at least in theory, be restricted to what is considered an acceptable range. This is a very important notion. Indeed, hypothesis tests (see Chapter 4) such as the t-test and many others can be linked to considerations of confidence limits.

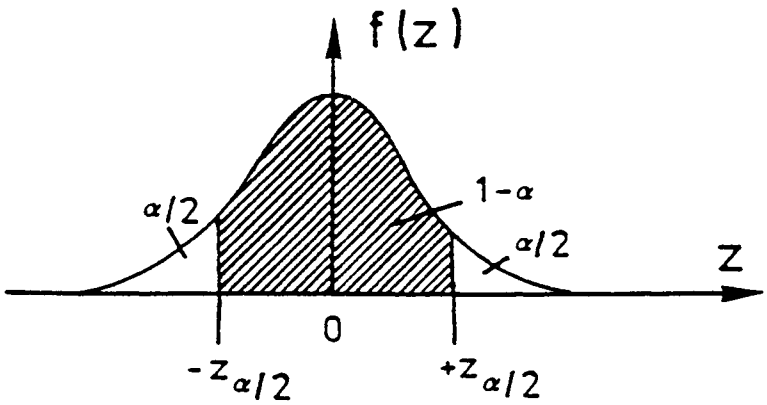


Fig. 3.7. A standardized normal distribution curve with the parameter α.

It follows that sample size will also be important in hypothesis tests and, more precisely, it will be shown that the sample size determines what kind of difference (for instance between a mean and a given value) a test can detect (see Section 4.8).

The notation used until now can be generalized by writing that the $(1-\alpha)100\%$ confidence interval around the mean is given by

$$\bar{x} \pm z_{\alpha/2} \, (\sigma / \sqrt{n}) \tag{3.17}$$

The meaning of parameter α for a standardized normal distribution is illustrated in Fig. 3.7. The fact that α/2 is used means that the interval is two-sided. If $\alpha = 5\%$, then the limits are made such that they exclude 2.5% on each side. There is a probability of 2.5% that μ will be situated outside the limits and lower than $\bar{x}$ and, equally, there is a probability of 2.5% of finding a value outside the confidence limits but higher than $\bar{x}$. Using our notation of eq. (3.17), we would write

$$\bar{x} \pm z_{0.025} \, (\sigma / \sqrt{n})$$

3.7 Small samples and the t-distribution

Equation (3.17) contains σ, the population standard deviation. This is a problem because this equation tries to estimate the unknown population parameter μ and its confidence limits from the sample parameter $\bar{x}$ using a population (and therefore also usually unknown) parameter σ. When $n \geq 30$ (some practitioners put the limit at 25), then s as defined by eq. (3.2) is considered a sufficiently good estimator of σ and one may write for the $(1 - \alpha) \cdot 100\%$ confidence interval

$$\mu = \bar{x} \pm z_{\alpha/2} \, (s / \sqrt{n}) \; (n \geq 30) \tag{3.18}$$

For $n < 30$, s is an uncertain estimate of σ. A correction is required and this is obtained by replacing z by t, so that:

$$\mu = \bar{x} \pm t_{\alpha/2}\,(s/\sqrt{n})\ (n < 30) \tag{3.19}$$

is the $(1-\alpha)\cdot 100\%$ confidence interval for sample sizes $n < 30$. The t-values are derived from tables of the *t-distribution.*

The notation of eq. (3.19) is found in many statistics books. Again it should be noted that, although an ISO norm [3] exists, there is no standardization in practice. ISO, for instance, writes (3.19) as:

$$\bar{x} - (t_{0.975}/\sqrt{n})\,s < m < \bar{x} + (t_{0.975}/\sqrt{n})\,s$$

Often, and we will follow this practice when we consider it useful, we write down the number of degrees of freedom, for which t is determined

$$\mu = \bar{x} \pm t_{\alpha/2,(n-1)}\,(s/\sqrt{n})$$

As for the z-tables, there are many different t-tables available. One possible layout is shown in Table 3.4. For the confidence interval for the mean the number of degrees of freedom (df) is $n - 1$. For instance, to obtain the 95% confidence interval for a sample size of $n = 10$, one consults the table at df = 9 and $t_{0.025} = 2.262$, so that $\mu = \bar{x} \pm 2.262(s/\sqrt{n})$. One notes that for df = ∞, $t_{0.025} = z_{0.025} = 1.96$. Also, at $n = 30$, $t_{0.025} = 2.04$, which is considered close enough to 1.96. The t-distribution is broader at the base and more peaked around the centre than the z-distribution (see Fig. 3.8). The higher the number of degrees of freedom, k, is, the closer it comes to the z distribution. The t-distribution is also known as Student's distribution. Thus, for small sample sizes the confidence interval is broader than when a large ($n > 30$) sample size is used or than when one knows σ, for instance, from prior

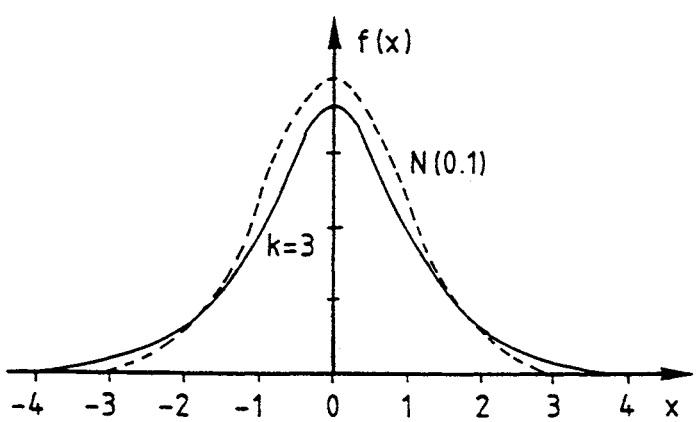


Fig. 3.8. The t-distribution for 3 degrees of freedom compared with the z-distribution.

TABLE 3.4

One-sided t-table (see also Fig. 3.3d)

	Area in upper tail of t-distribution						
df	0.10	0.05	0.025	0.01	0.005	0.0025	0.001
1	3.078	6.314	12.706	31.821	63.657	127.32	318.310
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327
3	1.638	2.353	3.182	4.541	5.841	7.453	10.215
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785
8	1.397	1.860	2.306	2.896	3.355	3.832	4.501
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485
24	1.318	1.711	2.064	2.492	2.797	3.090	3.467
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435
27	1.314	1.703	2.052	2.473	2.771	3.056	3.421
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232
120	1.289	1.658	1.980	2.358	2.617	2.860	3.160
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090

experimentation. When using smaller sample sizes, we pay a double price: the confidence in the estimate of the population mean is less precise (the confidence interval is larger), because we use the broader t-distribution and because we divide by a smaller n in eq. (3.19).

It is also useful to note that confidence intervals for the mean can be obtained by using the range. The procedure is described, for instance, in the annex to the ISO-norm [3].

3.8 Normality tests: a graphical procedure

As we will see in later chapters, many statistical tests are based on the assumption that the data follow a normal distribution. It is far from evident that this should be true. Distributions can be non-normal and procedures or tests are needed to detect this departure from normality. Also, when we consider that a distribution is normal, we can make predictions of how many individual results out of a given number should fall within certain boundaries, but, again, we then need to be sure that the data are indeed normally distributed. Sometimes one will determine whether a set of data is normally distributed because this indicates that an effect occurs that cannot be explained by random measurement errors. This is the case for instance in Chapters 22 and 23, where the existence of a real effect will be derived from the non-normality of a set of computed effects. In this section a graphical procedure is described that permits us to indicate whether a distribution is normal or not. A second graphical procedure, the box plot, will be described in Chapter 12. Graphical procedures permit us to visually observe whether the distribution is normal. If we want to make a formal decision, a hypothesis test is needed. Such tests are described in Chapter 5.

The graphical procedure applied here is called the *rankit procedure*. It is recommended by ISO [2] and we shall consider the numerical examples given in that international norm to explain how the method works. The example concerns the measurement of breaking points of threads. Twelve threads are tested and the following results are obtained:

2.286; 2.327; 2.388; 3.172; 3.158; 2.751; 2.222; 2.367; 2.247; 2.512; 2.104; 2.707.

We can reason that a result such as 2.104 must be representative of the lower tail of the distribution, 3.172 the higher tail, and results such as 2.367 and 2.388 the central part of the distribution. To have a better look, it seems logical to rank the data, yielding the following series:

2.104; 2.222; 2.247; 2.286; 2.327; 2.367; 2.388; 2.512; 2.707; 2.751; 3.158; 3.172.

To determine for which part of the distribution each number is representative, let us first look at a simpler example and suppose that only three numbers were given, 2.104, 2.367 and 3.172. We would then split up the range in four subranges, namely <2.104, 2.104–2.367, 2.367–3.172 and >3.172, and would consider that 2.104 is therefore located such that 25% of all data that could be obtained from the

distribution would fall below it, that 2.367 is located such that 50% fall below it, etc. In other words, the cumulative frequency of 2.104 would be equal to 25%, i.e. $100/(n+1)\%$, for 2.367 it would be $(100{\cdot}2)/(n+1)\%$, etc.

Let us now turn back to the thread data and discuss this using a more statistical vocabulary. We can now state that the cumulative frequency of data is equal to or lower than 2.104, in short the cumulative frequency of 2.104, is equal to 1 (since there is one observation ≤ 2.104) and that its cumulative relative frequency is given by

$$\text{cumulative \% frequency} = (100 \times \text{cumulative frequency}) / (n + 1) \text{ or}$$

$$(100{\cdot}1)/(12 + 1) = 7.7\%.$$

The cumulative frequency of 2.222 is 2 and the cumulative relative frequency is 15.4% and for 3.172 the respective values are 12 and 92.3%.

The following step is to assume that the data indeed come from a normal distribution. The value with a cumulative relative frequency of 7.7% is equivalent to the value that delimits a lower tail of a normal distribution with an area of 7.7%. Expressed in z-values by using one of the Tables 3.1, 3.2 or 3.3, this is equal to -1.43. By proceeding in this way for all the data, one obtains ranked z-values, also called ranked normal deviates or *rankits*. This yields Table 3.5.

It can now be shown that, when the data are indeed normally distributed, a graph of x against z yields a straight line. The result for the example is shown in Fig. 3.9. This figure also illustrates the weakness of this graphical method. It is sometimes (as is the case here) difficult to decide whether the points fall on a straight line or not. Nevertheless, it is a useful way of looking at the data, and in many cases, as illustrated further, it leads to clear conclusions.

TABLE 3.5

Computation of normal deviates from a set of ranked data. The measurements are strengths of threads in Newton [2].

Measurement (x)	Cumulative frequency	Cumulative % frequency	z
2.104	1	7.7	−1.43
2.222	2	15.4	−1.02
2.247	3	23.1	−0.74
2.286	4	30.8	−0.50
2.327	5	38.5	−0.28
2.367	6	46.1	−0.10
2.388	7	53.8	+0.10
2.512	8	61.5	+0.28
2.707	9	69.2	+0.50
2.751	10	76.9	+0.74
3.158	11	84.6	+1.02
3.172	12	92.3	+1.43

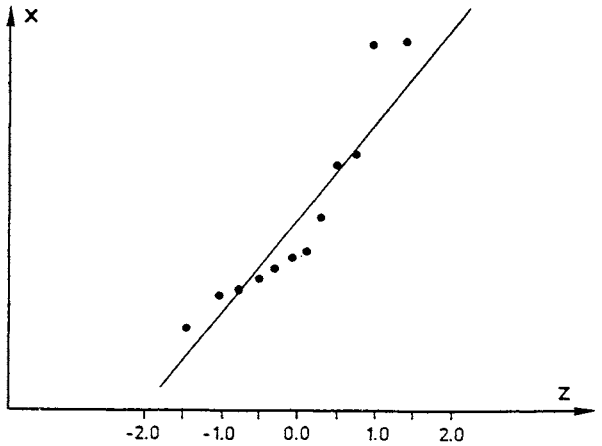


Fig. 3.9. Rankit method applied to the data of Table 3.5.

To make the procedure outlined above easier to carry out in practice, we can also use normal probability paper. The z-axis is then replaced by a cumulative probability axis. Fundamentally, this type of paper uses the straight-line relationship between z and x for a normal distribution, but it skips a step in the calculations by giving the axis the values of the percent cumulative frequency corresponding with the z-values. Applying this to the data we are examining here, leads to Fig. 3.10 and, of course, the interpretation is the same as for Fig. 3.9. When a straight line is obtained, we conclude that the distribution is normal.

There may be several reasons why an experimentally obtained set of data is found to be not normally distributed. This is illustrated with two examples described by Feinberg and Ducauze [4]. The first concerns a set of Pb measurement by AAS on the same portion of beef liver. The results are given in Table 3.6 and the rankit-line is shown in Fig. 3.11a. The line is clearly not straight. Closer inspection reveals that this may be due to the two highest results. After elimination of these two points, we obtain Fig. 3.11b. Now a straight line can be drawn through the points. The effect was due to two outlying points. The underlying distribution is normal, but outliers distort it. The presence of outliers can be seen

TABLE 3.6

Results of Pb determinations (in mg/kg) in the same portion of beef liver (from Feinberg and Ducauze [4])

0.965	0.975	1.040	1.095	1.105
1.135	1.135	1.165	1.167	1.180
1.200	1.210	1.210	1.232	1.232
1.242	1.300	1.362	1.945	2.185

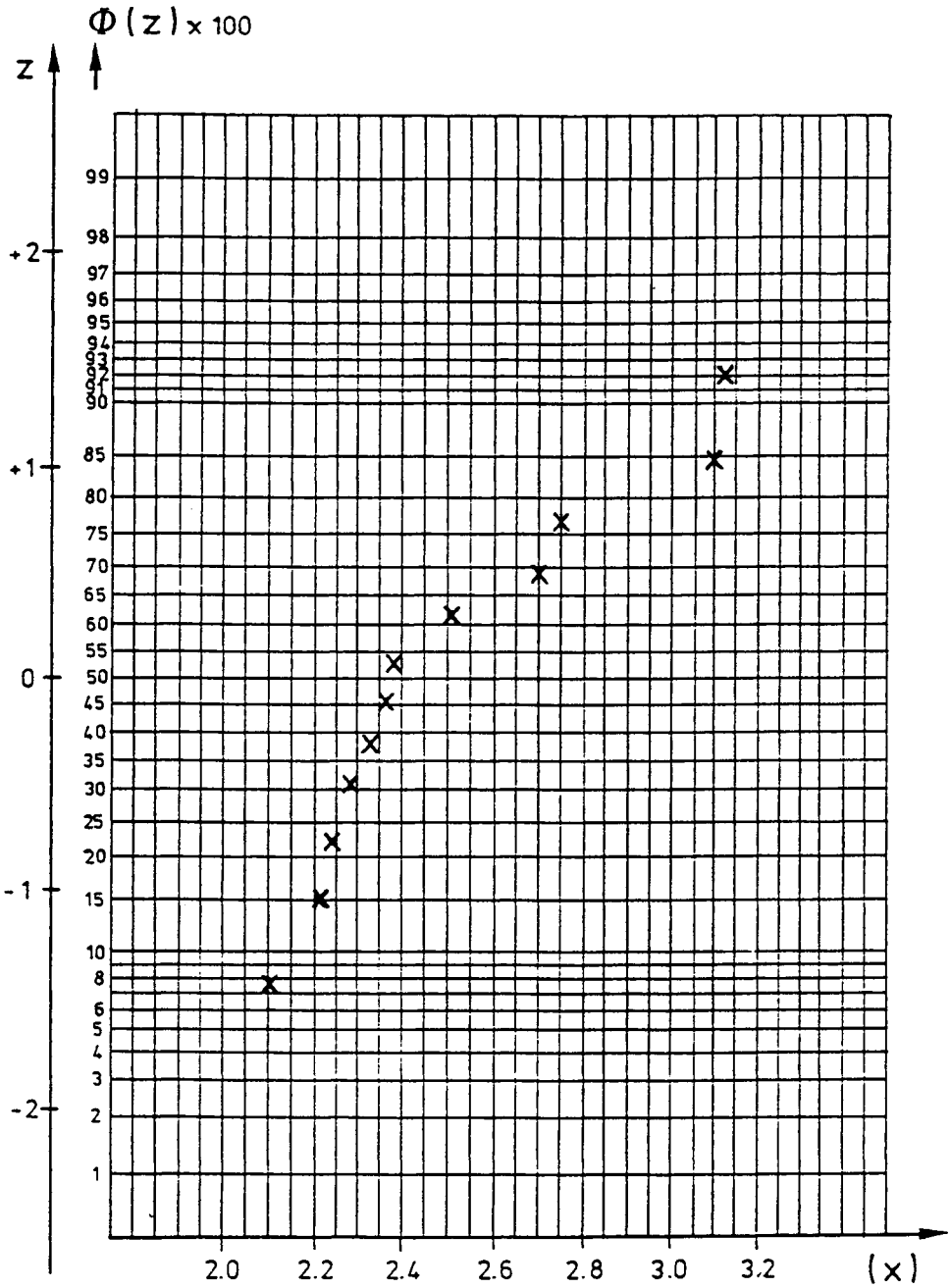


Fig. 3.10. Rankit method: use of probability paper for the data of Fig. 3.9 and Table 3.5.

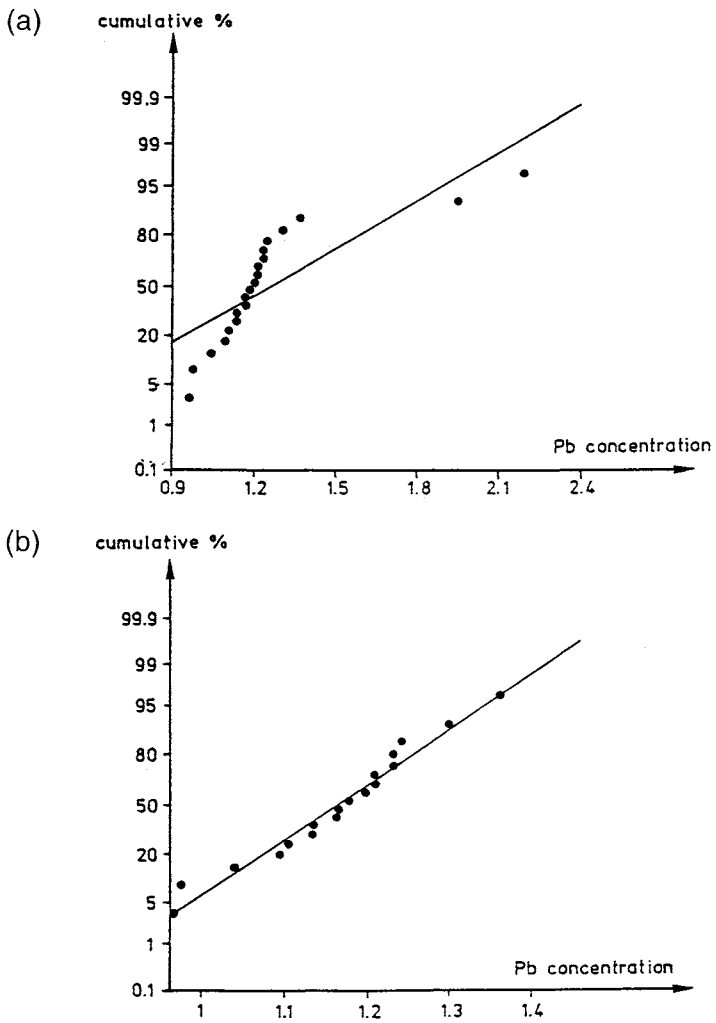


Fig. 3.11. (a) Rankit method for the AAS data of Table 3.6; (b) Rankit method for the AAS data after elimination of the two highest data.

as a source of non-normality. In fact, we will treat these data again in Chapter 5 with outlier tests and a formal test for normality (the Kolmogorov–Smirnov test) and the evaluation of all these approaches together confirms that the two results may be considered to be outliers.

The second example concerns bacterial counts on ground meat. The data are given in Table 3.7 and the rankit-line is shown in Fig. 3.12a. It is known that this type of data follows a *lognormal distribution*. This means that the ln (x_i) or log (x_i) are normally distributed. In Fig. 3.12b the rankit-line for the natural logarithms of

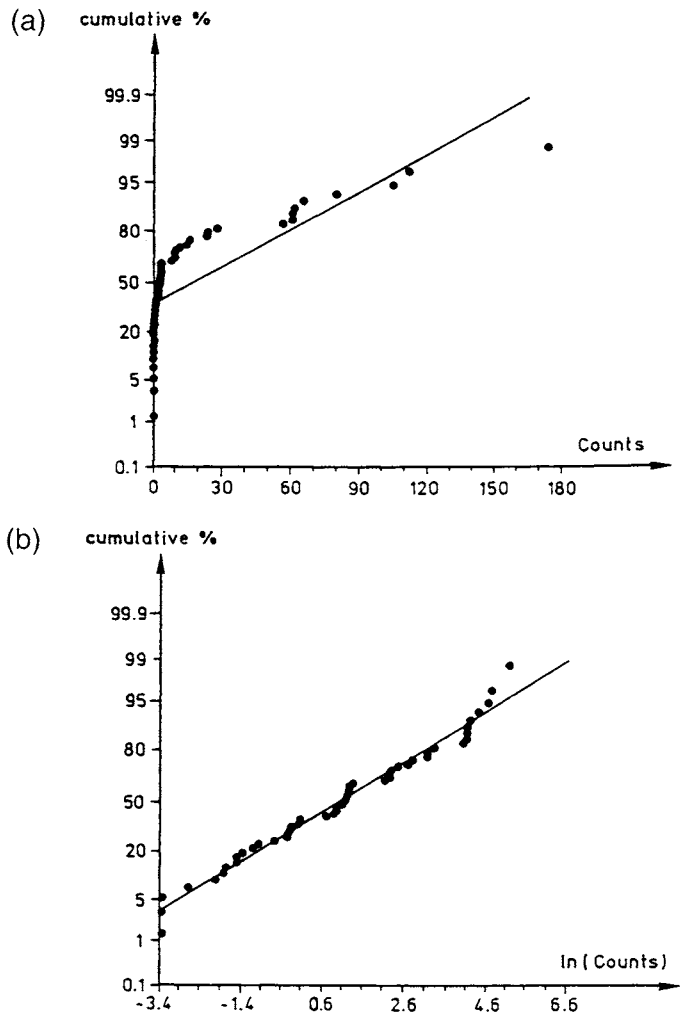


Fig. 3.12. (a) Rankit method for the bacteriological data of Table 3.7; (b) Rankit method for the log of the same data.

the counts is shown and the graph indeed indicates that they can be considered to be normally distributed.

When larger numbers of data are available, it is convenient to first group them into classes. This can be demonstrated again with the data on fluoride in the enamel of teeth of young children of Table 2.1. The cumulative frequencies (on probability paper) or the equivalent z-values (on the usual linear graph paper) are plotted against the class marks. The result on probability paper is shown in Fig. 3.13. We can conclude that the fluoride data are normally distributed.

TABLE 3.7

Bacteriological counts of 50 samples of ground meat (from Feinberg and Ducauze, [4]).

0.035	0.035	0.036	0.069	0.136
0.164	0.171	0.222	0.226	0.258
0.327	0.380	0.560	0.780	0.800
0.840	0.860	1.010	1.050	2.020
2.440	2.600	2.600	3.000	3.230
3.300	3.340	3.500	3.600	3.600
3.760	8.500	9.400	9.500	10.100
12.000	15.000	16.230	23.700	24.100
28.200	57.000	61.000	61.000	62.000
66.000	80.000	105.000	112.000	174.000

All data were divided by 10^5.

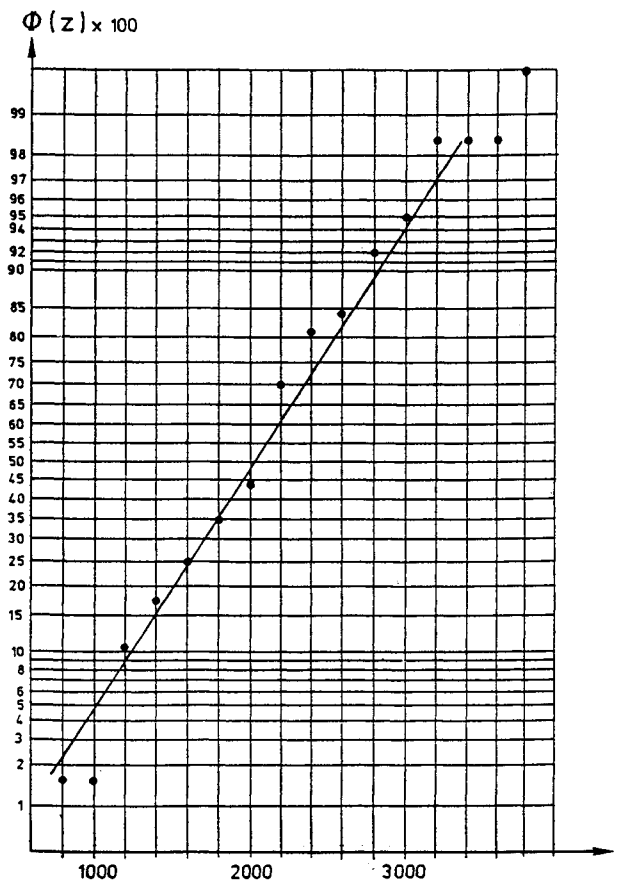


Fig. 3.13. Rankit method for the fluoride data of Table 2.1.

3.9 How to convert a non-normal distribution into a normal one

The bacteriological example already provides a clue to how to make non-normal distributions normal, namely by *transformation*. The transformation we carried out in the previous section is called the *log-transformation*. Log-normal distributions are frequently found in nature, particularly when the variable studied has a natural zero (such as weight, length, etc.). In this case simple normality around the mean could include negative values. The log-transformation is however not the only transformation one can think of. Other often used transformations are the square root transformation ($y = \sqrt{x}$), the inverse transformation ($y = 1/x$), the square transformation $y = x^2$. A special type of transformation, mainly useful when one studies proportions, is the arcsine transformation [5].

A procedure to find the transformation that best approaches normality has been described by Box and Cox [6]. Its approach is very similar to that used for finding a transformation to straighten a line (see Chapter 8). They propose the following general equation:

$$\begin{aligned} y_i &= \frac{(x_i + k_1)^{\lambda} - k_2}{\lambda} \qquad \text{for } \lambda \neq 0 \\ y_i &= \log (x_i + k_1) \qquad \text{for } \lambda = 0 \end{aligned} \tag{3.20}$$

This procedure requires us to find optimal values for the three parameters λ, k_1 and k_2 and therefore an optimization procedure such as the Simplex (Chapter 26) would be needed. For this reason, we usually simplify this to

$$\begin{aligned} y_i &= x_i^{\lambda} \qquad \text{for } \lambda \neq 0 \\ y_i &= \log x_i \qquad \text{for } \lambda = 0 \end{aligned} \tag{3.21}$$

We then select a criterion that describes similarity to (or distance from) normality. This can be the Kolmogorov–Smirnov d-value (see Chapter 5), but other criteria are also possible, such as skewness. The latter is chosen here. We will use a_3. As explained in Section 3.2, for a perfectly symmetric distribution a_3 should be close to 0. The procedure consists in computing $y_i = x_i^{\lambda}$ in function of λ (and if $y_i = \log x_i$ for $\lambda = 0$). For each λ, the skewness of the distribution of the y_i is then obtained. This yields Table 3.8. and Fig. 3.14. The optimal λ (i.e. yielding the lowest a_3) = 0, so that we should indeed choose the log transformation.

TABLE 3.8

Moment coefficient of skewness, a_3, in function of λ in eq. (3.21) for the data of Table 3.7

λ	a_3	λ	a_3
−2	0.505	0.1	0.035
−1.9	0.502	0.2	0.083
−1.8	0.498	0.3	0.125
−1.7	0.494	0.4	0.163
−1.6	0.489	0.5	0.197
−1.5	0.483	0.6	0.229
−1.4	0.476	0.7	0.259
−1.3	0.468	0.8	0.288
−1.2	0.457	0.9	0.318
−1.1	0.445	1	0.347
−1	0.429	1.1	0.377
−0.9	0.409	1.2	0.407
−0.8	0.386	1.3	0.437
−0.7	0.357	1.4	0.468
−0.6	0.323	1.5	0.498
−0.5	0.283	1.6	0.527
−0.4	0.237	1.7	0.555
−0.3	0.185	1.8	0.583
−0.2	0.130	2	0.635
−0.1	0.073		
0	0.017		

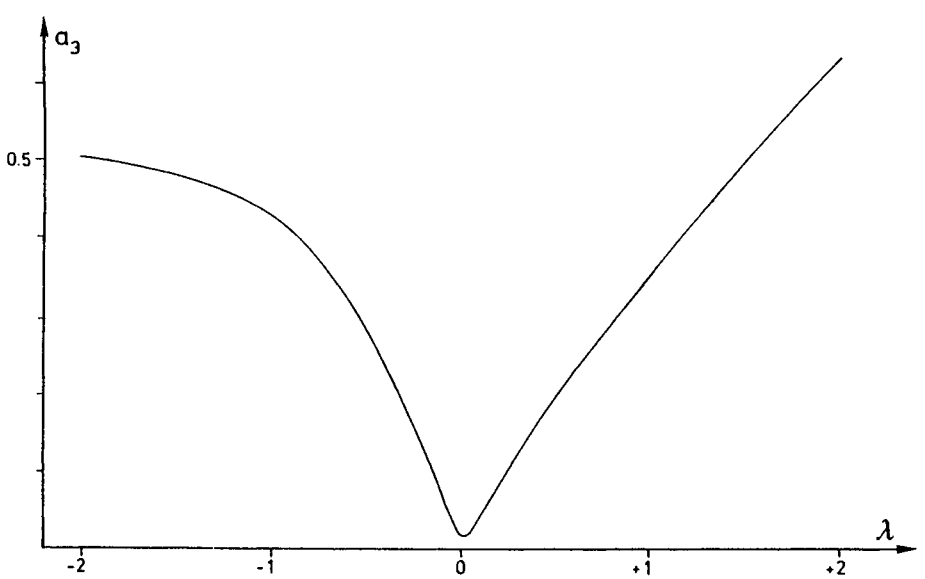


Fig. 3.14. Selection of transformation equation to normality. Skewness, as measured by a_3, against λ in eq. (3.20).

References

1. J. Gurland and R.C. Tripathi, Am. Stat., 25 (1971) 30–32.
2. ISO norm 2854-1976 (E) Statistical interpretation of data. Techniques of estimation and tests relating to means and variances, 1976.
3. ISO norm 2602-1980 (E) Statistical interpretation of test results – Estimation of the mean – Confidence interval, 1980
4. M. Feinberg and C. Ducauze, Exprimer le résultat d'une analyse: une approche statistique et informatique. Analusis, 12 (1984) 26.
5. R.R. Sokal and F.J. Rohlf, Biometry. Freeman, New York, 2nd ed., 1981 p. 427.
6. G.E.P. Box and D.R. Cox, J. Roy, Stat. Soc. B, 26 (1964) 211.

Chapter 4

An Introduction to Hypothesis Testing

4.1 Comparison of the mean with a given value

Let us consider the following situation which is described in more detail in Chapter 13 on Method Validation. To investigate possible bias we have prepared a powder containing all the ingredients of a formulated drug in known amounts. Suppose that the amount of drug added is 100.0 mg.

Example 1: Four determinations are carried out ($n = 4$). The mean, $\bar{x} = 98.2$ and the standard deviation is known to be 0.80 from prior experience ($\sigma = 0.80$).

Example 2: Six replicate determinations are carried out with the following results:

98.9 – 100.3 – 99.7 – 99.0 – 100.6 – 98.6 ($n = 6$, $\bar{x} = 99.5$, $s = 0.81$)

Note that in this case the standard deviation is not known, but estimated from the 6 replicate results.

We now need to decide whether the mean obtained, $\bar{x} = 98.2$ (Example 1) or 99.5 (Example 2), is really different from the amount, $\mu_0 = 100$, we should find. Notice that $\bar{x}$ is only an estimate of the true result, μ, the population mean, that would be found if we were to carry out an infinite number of replicate determinations. Therefore, we conjecture whether it is true that μ, as estimated by $\bar{x}$, is equal to μ_0. This is an example of a hypothesis. To ascertain whether one can accept the hypothesis to be true, a *hypothesis test* is carried out.

Hypothesis testing is a very important part of statistics. Here, we will apply it to test whether the mean of observed results should be considered equal to a given value. Hypothesis testing can also be used to investigate whether the means or standard deviations of two or more series of results are equal, whether the slope of a regression line is really 0 (or 1, according to the context), etc. Many of the following chapters will be devoted to developing appropriate hypothesis tests. In this chapter our aim is to introduce the subject of hypothesis testing. We will do this by considering the following hypothesis:

Hypothesis: the mean, μ, of a population of measurement results estimated from a relatively small set of observed results, $\bar{x}$, is equal to a given value μ_0.

The test needed to test this hypothesis is the easiest to understand and we will use it to explain how a hypothesis test is carried out in general and to consider some questions common to all hypothesis tests.

We should make here an important note. *Statistical significance* means that a difference between two numbers (here 100 and 98.2 or 100 and 99.5) is considered real. It does not necessarily mean that the difference is relevant to the problem under study. For instance, if the test were to conclude that the difference between 100 and 99.5 is significant, this does not necessarily mean that the method studied is declared incorrect. In fact, the method developer probably will be quite pleased with the outcome and use the method, because a difference of 0.5% is in this application of no consequence, even if it is statistically significant. Decision-making should therefore be a two-step process. One should first ask whether a difference is practically relevant and then whether it is also statistically significant. The concept of relevant difference is introduced from Section 4.7 onwards.

4.2 Null and alternative hypotheses

The hypothesis as formulated above is that there is no difference between μ and μ_0. This is called a *null hypothesis* and the customary short hand notation for it is:

H_0: $\mu = \mu_0$

or, since μ_0 in both examples is equal to 100.0,

H_0: $\mu = 100.0$

For the case that the null hypothesis is not true, we need to formulate an alternative. This is referred to as the *alternative hypothesis*, H_1. Here we will formulate it simply as:

H_1: $\mu \neq 100.0$

It must be noted that this choice is not evident. In Section 4.9, we will see that instead of H_1: "is different from", there are situations where it is preferable to state H_1: "is greater than" or, of course, H_1: "is smaller than".

When carrying out a hypothesis test, it is good practice to state clearly at the outset what both hypotheses are. In our example,

H_0: $\mu = 100.0$

H_1: $\mu \neq 100.0$

4.3 Using confidence intervals

Let us first consider Example 1. The 95% confidence interval around $\bar{x}$ is given by:

$$98.2 \pm 1.96\,(\sigma/\sqrt{n}) = 98.2 \pm 1.96 \cdot 0.40 = 98.2 \pm 0.78$$

When the target value (here 100.0) is inside the confidence interval, then we consider it as compatible with $\bar{x}$. We would then conclude that $\bar{x} = 98.2$ is not an improbable value for $\mu = 100.0$ and that therefore

$$\mu = \mu_0 = 100.0$$

and we would accept H_0. Accepting the null hypothesis does not imply that we have proven that the hypothesis is true. The only thing we can conclude is that the data are compatible with H_0 and that there is not enough evidence to reject H_0.

When the confidence interval around $\bar{x}$ does not contain μ_0, then we would reject the null hypothesis because the value of $\bar{x}$ is improbable for a $\mu = 100.0$. Indeed, there is only a 5% probability that μ has a value situated outside the confidence interval.

In the case of Example 1, since μ_0 is outside the confidence interval around $\bar{x}$, we would reject the null hypothesis and our conclusion would be to accept H_1: $\mu \neq \mu_0$. In Example 2, σ is not known. We therefore have to use a t-value to construct the confidence interval (see Section 3.7):

$$99.5 \pm t_{0.025,5}\, s/\sqrt{n}$$

$$\text{or } 99.5 \pm 2.57 \frac{0.81}{\sqrt{6}} = 99.5 \pm 0.85$$

$\mu_0 = 100.0$ falls inside this interval and therefore we consider that $\bar{x} = 99.5$ is consistent with $\mu = 100.0$ and accept

$$H_0: \mu = \mu_0 = 100.0$$

Let us now summarize how we have carried out the hypothesis test. We have carried out the following steps:

1. We have stated the null and alternative hypothesis. For both examples:

   $H_0: \mu = \mu_0 = 100.0$

   $H_1: \mu \neq 100.0$

2. We have decided that $\alpha = 5\%$.
3. We have defined a confidence interval around $\bar{x}$ at the $100 - \alpha = 95\%$ level.

   $\bar{x} \pm 1.96(\sigma/\sqrt{n}$ ($n > 25$ or σ known) (4.1a)

or

$$\bar{x} \pm t(s/\sqrt{n} \text{ (otherwise)} \tag{4.1b}$$

4. We investigated whether μ_0 falls within the confidence interval.
5. If the answer to our question was "yes" we accepted H_0, if it was "no" then we rejected H_0 (and accepted H_1).
6. Presentation of results (see Section 4.5).

It is very important to understand that the same decision scheme can always be followed: it is valid for all hypothesis tests. In other words, when we have a confidence interval we can carry out a hypothesis test. Let us consider an example. In Chapter 8, we will learn how to estimate a regression line. The estimate of the intercept of such a line is given by b_0. This estimates β_0, the true intercept. In Chapter 13, we will see circumstances where we would like to know whether we can accept that $\beta_0 = 0$. Let us see how we would carry out the hypothesis test for a situation where $b_0 = 0.10$. We would comply with the following reasoning.

1. H_0: $\beta_0 = 0.0$

   H_1: $\beta_0 \neq 0.0$

2. $\alpha = 5\%$
3. We have not learned yet how to determine a confidence interval around b_0, but suppose it is found to be $b_0 \pm 0.15$. The confidence interval is then $b_0 - 0.15$ to $b_0 + 0.15$ or [–0.05, 0.25].
4. Does 0.0 fall within [–0.05, 0.25]?
5. The answer is "yes". Therefore we accept H_0: $\beta_0 = 0.0$
6. Presentation of results (see Section 4.5).

4.4 Comparing a test value with a critical value

There is a second way in which hypothesis tests can be carried out. Fundamentally, it is exactly the same as the method described in the previous section, but it looks somewhat different. As we saw in Chapter 3.6, we can state that 95% of all sample means $\bar{x}$ of size n fall within the limits $\mu \pm 1.96\ \sigma/\sqrt{n}$. If we suppose that H_0: $\mu = \mu_0$ is true, then this statement is correct for all $\bar{x}$ falling within the interval $\mu_0 \pm 1.96\ \sigma/\sqrt{n}$. In Fig. 4.1 the distribution of the $\bar{x}$ around μ_0 for the first example is given, once in the original units (mg) (Fig. 4.1a) and once in z units (Fig. 4.1b). We know that 95% of all means compatible with H_0: $\mu = \mu_0$ are situated within $z = -1.96$ and $z = +1.96$ of the standardized normal distribution of means. The 95% acceptance interval for H_0: $\mu = \mu_0$ in z-units is therefore given by $-1.96 \leq z \leq +1.96$ or $|z| < 1.96$. We can also express the distance of the observed $\bar{x}$ from μ_0 in z-units.

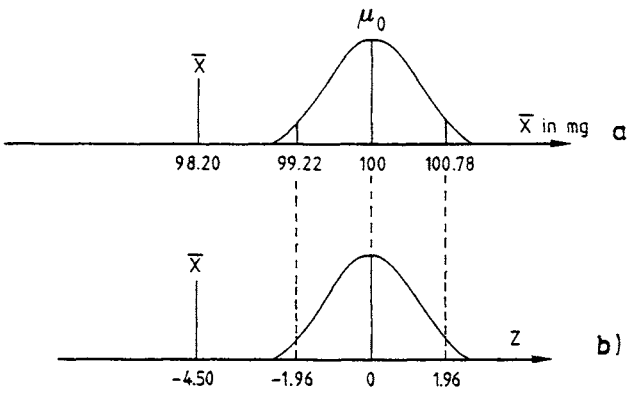


Fig. 4.1. The distribution of $\bar{x}$-values for Example 1 (see text) that would be obtained if $\mu_0 = \mu$: (a) in original units, (b) in standard deviate units.

If this experimental z is larger in absolute value than 1.96, the $\bar{x}$ being tested falls outside the acceptance interval for H_0 and H_1 will be accepted. Otherwise, if $|\bar{x} - \mu_0|/(\sigma/\sqrt{n}) \leq 1.96$, we will accept H_0.

This way of presenting a hypothesis test is different because it does not explicitly apply confidence limits. It is, however, very important to realize that both ways of presenting a hypothesis test lead to exactly the same conclusion.

Let us summarize the second way of presenting a hypothesis test for both examples used in this chapter and, first, for Example 1

1. State the hypotheses

   H_0: $\mu = \mu_0 = 100.0$

   H_1: $\mu \neq 100.0$

2. $\alpha = 5\%$
3. What is the critical z-value? It is $z_{crit} = 1.96$
4. What is the z-value for the $\bar{x}$ being tested? It is computed with eq. (4.2)

$$|z| = \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \tag{4.2}$$

   and this is in this case equal to

$$|z| = \frac{|98.2 - 100.0|}{0.8/\sqrt{4}} = 4.50$$

5. If $|z| < |z_{crit}|$, then accept H_0. In this case $|z| = 4.5 > 1.96$, so that H_0 is rejected (and H_1 accepted). We conclude that $\mu \neq 100.0$
6. Presentation of results (see Section 4.5).

For the second example, steps 1 and 2 are exactly the same, so that we give only the following steps.

3. What is the critical t-value? Since in this example, σ is not known and $n < 25$ to 30, one uses s and t. For 5 degrees of freedom,

   $|t_{crit}| = 2.57$

4. What is the t-value corresponding to 99.5? The equation used is:

$$|t| = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \tag{4.3}$$

   In this case:

$$|t| = \frac{|99.5 - 100.0|}{0.813/\sqrt{6}} = 1.51$$

5. Since $|t| = 1.51 < 2.57$, H_0 is accepted and one concludes that

   $\mu = 100.0$

6. Presentation of results (see Section 4.5).

We have seen now two ways of testing a hypothesis:

1. Determining the confidence interval around $\bar{x}$ and observing whether μ_0 falls in it.
2. Determining a critical z or t-value and observing whether this is exceeded or not.

We have a preference for method 1, because it does not only yield a decision on accepting H_0 or H_1, but also gives a compact and informative summary of the measurement result: it is more data oriented. On the other hand, method 2 allows us to give p-values (see next section).

Significance testing should not be applied as a "yes" or "no" procedure, except perhaps in a regulatory context where rules have to be followed. Scientifically, there is no reason to make entirely opposite conclusions when $p = 0.048$ and $p = 0.052$. As noted by Box, Hunter and Hunter [1] "significance testing in general has been a greatly overworked procedure".

4.5 Presentation of results of a hypothesis test

The experimental or calculated z-value of Example 1 ($z = -4.5$) coincides with an $\alpha \cong 0.000005$ and the experimental t-value of Example 2 ($t = -1.51$, df = 5) with an $\alpha = 0.19$. To make a distinction between the *a priori* α-value (usually $\alpha = 0.05$) and the one actually obtained, it is customary to write that $p \cong 0.000005$ (Example 1) and $p = 0.19$ (Example 2) instead of α.

If $p > \alpha$, as is the case for Example 2, then the probability of making an error by stating that there is an effect (i.e. a difference between μ and μ_0) is too large and it

is preferable to state that there is no significant effect (shorthand notation NS). This also means that we will find that μ_0 is inside the confidence interval or that $|z| < z_{crit}$. We can now fill in point 6 of Sections 4.3 and 4.4 for Example 2 to read:

6. $p = 0.19$ (NS)

For the first example, $p < \alpha$, which means that μ_0 is outside the 100 $(1 - \alpha)$% confidence interval. We write the p value and, to give an idea of the confidence we have in the result, add $p < 0.05$, $p < 0.01$, $p < 0.001$, etc. as happens to be the case. Writing $p < 0.05$ also implies in such a case that $p > 0.01$. For Example 1, we would fill in point 6. to read:

6. $p \sim 0.000005$ ($p < 0.00001$).

4.6 Level of significance and type I error

Let us return to Section 4.3 where it was decided to use confidence intervals as decision criteria. For the example introduced in that section it was decided to reject all values outside the limits 97.42–98.98 as not belonging to the probability distribution around $\bar{x}$ and to conclude that all values of μ_0 outside that interval are not compatible with $\mu = \mu_0$. The confidence intervals were chosen so as to include 95% of the probability distribution.

We should now focus on the other 5% and to do this we must turn the argument around. Let us suppose that $\mu = \mu_0 = 100$. For any value of $\bar{x}$ within the range 99.22–100.78, we would conclude that $\mu = \mu_0$. Indeed, in all these cases, the confidence interval around $\bar{x}$ would include $\mu_0 = 100.0$. However, there is 5% probability that a value of $\bar{x}$ outside the range 99.22–100.78 would be obtained when $\mu = \mu_0 = 100.0$. Nevertheless, we have decided that we would consider such values as inconsistent with $\mu = \mu_0$ and would consider $\bar{x}$'s with such values as indicating that $\mu \neq \mu_0$. In these 5% of cases, we would therefore make an error. The 5% is called the *level of significance* and is equal to the *probability of (incorrectly) rejecting the null hypothesis when it is true*. The error we make in this way is called a *type I error* or also the α *error*.

4.7 Power and type II Errors

There is also a type II error. To understand this we should consider Fig. 4.2. Let us again consider the situation where the null hypothesis is true: $\mu = \mu_0 = 100.0$. The $\bar{x}$-values that would be obtained would be situated with 95% probability in the range 99.22–100.78. The value of 98.20 is outside this range and the confidence interval around 98.20 (97.42–98.98) would not include 100.0. Therefore, we would reject H_0 on finding a value of 98.20 and would also do so for any other value below

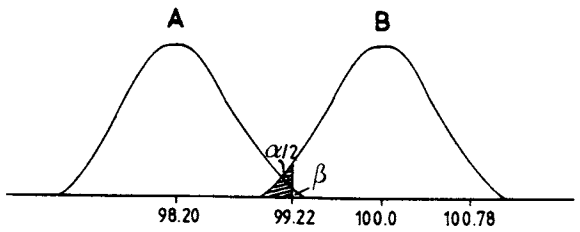


Fig. 4.2. Type I (α) and II (β) errors for Example 1. Distribution A is the distribution of $\bar{x}$-values that would be obtained if the measurement were biased ($\mu = 98.20$). Distribution B is the distribution of $\bar{x}$-values that would be obtained if the measurement were unbiased ($\mu = \mu_0 = 100.0$).

99.22. The limits are chosen such that there is a 5% probability that we would make an α or type I error ($\alpha = 0.05$ or 5%). This is so because in 5% of all cases a set of 4 determinations with an unbiased method ($\mu = \mu_0$) for which $\sigma = 0.80$ will yield a value outside the limits.

Let us now suppose that the method is indeed biased with a bias -1.80 and that the population mean μ of the determinations is therefore 98.20. We might wonder whether that bias could go undetected. If we were to find a value of $\bar{x}$ higher than 99.22, we would conclude that there is no bias because the confidence interval around 99.22 would include 100.0. We would accept the H_0 hypothesis, i.e. conclude that $\mu = \mu_0$, while in fact this is not true. We have now made a *type II error* or β-*error, which consists in (incorrectly) accepting that* H_0 *is true, while in fact it is not.*

There is a relationship between the two types of error. Let us compute how large the β-error would be for our example. Knowing that there is a bias of -1.80 and that therefore $\mu = 98.20$, what is the probability of finding a value higher than 99.22? The z-value for 99.22 on the distribution centred around 98.20 is given by:

$$z = \frac{99.22 - 98.20}{0.8/\sqrt{4}} = 2.55$$

Using Table 3.2 we see that the fraction of values with $z \geq 2.55$ is 0.006. The β-error is 0.006 or 0.6%. For this example, we can summarize that when we accept an α-error of 5%, we incur a risk of $\beta = 0.6\%$ of falsely accepting H_0, when there is a true bias of -1.8.

Let us now suppose that $\alpha = 5\%$ is considered too large, i.e. having a 1 out of 20 probability to decide that H_0 must be rejected when it should not. We would like to reduce the risk and therefore set $\alpha = 1\%$. The lower decision limit around 100.00 within which $\bar{x}$-values would lead to the conclusion $\mu = \mu_0 = 100$ would now be

$$100.00 - 2.57\left(\frac{0.8}{\sqrt{4}}\right) = 98.97$$

The probability that an $\bar{x}$ higher than this decision limit would be obtained out of the population centred around $\mu = 98.20$ has now grown larger. Since

$$z = \frac{98.97 - 98.20}{0.8/\sqrt{4}} = 1.92$$

the probability of finding a higher value than 98.97 is 2.6%. There is a probability of 2.6% that one would not detect the bias.

Let us suppose that there is a somewhat smaller bias, $\mu = 98.40$. As before, we can compute the z-value for 99.22 for a distribution centred around 98.40. This would then be 2.04, resulting in a β-error of 2.1% for $\alpha = 5\%$; for $\alpha = 1\%$, $z = 1.43$ and $\beta = 7.6\%$. Clearly, decreasing α increases the β-error. For this reason, we would not reduce α to very low values. Very often, $\alpha = 0.05$ and, in fact, this has become so standard that a good reason is needed to replace it by $\alpha = 0.01$ in step 2 of the hypothesis testing process as explained in Sections 3 and 4.

Incidentally, we can now define what is called the *power* of a test. This is the probability of *correctly* rejecting H_0 when it is false or, in other words, how likely the test is to detect a statistically significant difference. Since β is the probability of accepting H_0 under those circumstances, the power of a test is given by $1 - \beta$. For $\alpha = 5\%$ and $\mu = 98.4$, $\beta = 0.021$ and the power of the test $= 1 - 0.021 = 0.979$ or 97.9%.

The comparison of the β-error for $\mu = 98.2$ and $\mu = 98.4$ shows that, all circumstances being equal, β grows as the difference between 100 and μ becomes smaller. This is common sense. We are more likely not to detect a bias when that bias is small. It is also common sense that β will be larger when σ is larger and n smaller. The two distributions of Fig. 4.2 then overlap to a larger extent. In summary, for a given α, β grows as $|\mu_0 - \mu|$ decreases, σ increases or n decreases.

The effect of n will be investigated further in Section 4.8. Let us return to the difference $\mu_0 - \mu$. The effect of this difference is often described with a *power curve*. This is a plot of $1 - \beta$ (the power) as a function of the $|\mu - \mu_0|$. When $|\mu - \mu_0|$ is small the probability that H_0 will be rejected is also small or β is large and the power, $1 - \beta$, again small. The larger $|\mu - \mu_0|$ becomes the larger the power becomes. When $|\mu - \mu_0|$ is sufficiently large, the power becomes virtually 1. This is shown in Fig. 4.3 together with the so-called *operating characteristic curve* (*OC curve*). This is the curve relating β and $|\mu - \mu_0|$ for a given α, σ and n. The two curves, of course, give the same information.

The power of a test is sometimes called its *sensitivity*. This is for instance the case when one carries out clinical tests. In this context, α and β considerations are very important; α represents then the probability of obtaining what is called a *false positive* and β that of obtaining a *false negative* conclusion, the conclusion being that a patient suffers from some disease. This is discussed at greater length in Section 16.1.3.

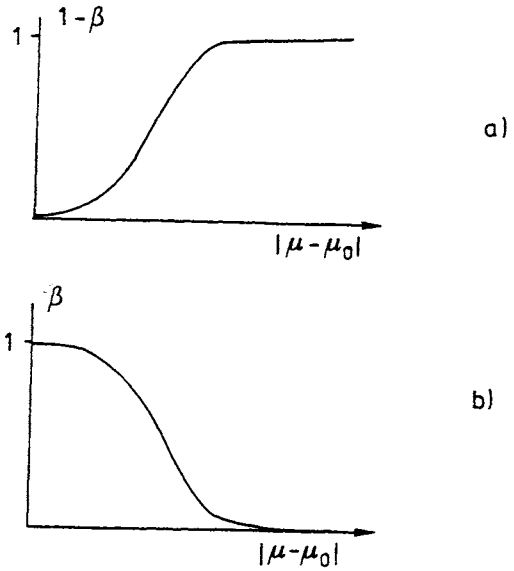


Fig. 4.3. (a) Power curve. (b) Operating characteristic curve.

4.8 Sample size

So far, we have set an α-level and a sample size n, and, for a given $|\mu_0 - \mu|$ and σ or s, we were then able to compute β. Another question is to determine what sample size n is large enough to achieve the purpose of our test with sufficient confidence. To express it in terms of our examples, β is the probability that we will not detect a certain bias although it exists. Stated in this way β clearly is important and we should ask the question: how large should the sample size n be to detect a given difference $|\mu_0 - \mu|$, which is considered relevant, with a given σ or s, so that there is only $\alpha\%$ probability of deciding that there is a difference when there is none, and $\beta\%$ probability of not detecting the difference when it does exist. Stated again in terms of our example:

- how many replicates n should be analyzed to detect a bias of at least $|\mu_0 - \mu|$ in a procedure with a known precision, σ, or a precision, s, estimated from the experiment, so that there is a probability of not more than $\alpha\%$ to decide there is a bias, when there is in fact none and, at the same time a probability of not more than $\beta\%$ that a bias larger than $|\mu_0 - \mu|$ will go undetected?

If we call δ the minimum difference that we want to detect, we can verify that (see Fig. 4.4):

$$n \geq \left[(z_{\alpha/2} + z_{\beta})\, \sigma/\delta\right]^2 \tag{4.4}$$

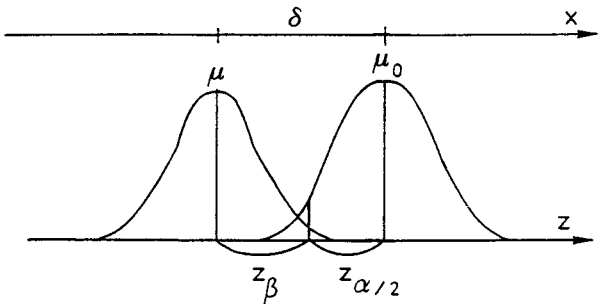


Fig. 4.4. Derivation of equation for minimum sample size.

for known σ where $z_{\alpha/2}$ and z_β are the value of z for the stated α and β, respectively. Indeed from the example given above, it can be understood that the decision limit is situated at $z_{\alpha/2}$ from μ_0 and z_β from μ. If a certain difference $\delta = |\mu - \mu_0|$ must be detected at a given α and β-level, for instance $\alpha = \beta = 5\%$, then, expressed in z units $\delta \geq z_{\alpha/2} + z_\beta = z_{0.025} + z_{0.05}$. In the original units and still assuming σ is known

$$\delta \geq (z_{\alpha/2} + z_\beta)\, \sigma / \sqrt{n}$$

which yields eq. (4.4) by rearrangement.

We can also make use of graphs published by ISO [2]. In Fig. 4.5a the required sample size n can be derived for a known σ and in Fig. 4.5b for an s, estimated by the experiment. Let us consider first only Fig. 4.5a. The abscissa is λ, where

$$\lambda = \frac{|\mu - \mu_0|}{\sigma} \tag{4.5}$$

The parameter λ is in fact the effect or bias we want to detect expressed in standard deviation units. It may be that a smaller effect than λ exists, but this is considered of no practical interest by the experimenter. Let us return to Example 1. The known amount $\mu_0 = 100$ mg. A bias of 1.5 mg is judged to be relevant and should be detected with a $\beta = 0.05$. Since it is known that $\sigma = 0.8$ mg, $\lambda = 1.5/0.8 = 1.87$ and, from Fig. 4.5a, $n \geq [(1.96 + 1.65)/1.87]^2 = 3.73$ is derived, then $n = 4$ determinations need to be carried out.

Suppose now that σ had been 1.6 mg. Then $\lambda = 0.93$ and n should then have been 15. Equally, if the bias to be detected had been 1.0 instead of 1.5, then, for $\sigma = 0.8$, $\lambda = 1.25$ and n should have been larger than 8.33, i.e. 9.

If we compare Fig. 4.5a with Fig. 4.5b, we observe that the two sets of lines are about the same for $n > 25$. This is the limit above which in Chapter 3 it was accepted that the experimental s may be equated with σ and where probability calculations can be performed with z as a parameter. Below $n = 25$, the divergence increases. Below that limit, calculations are performed with t-values, which become

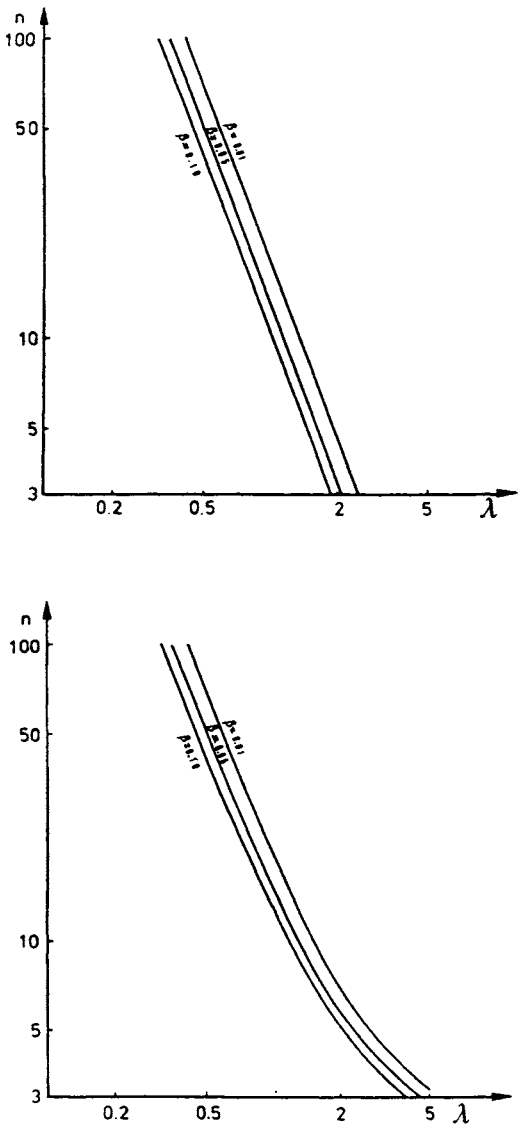


Fig. 4.5. Sample size required to detect a certain bias $\lambda = \delta/\sigma$ (situation a, σ known) or $\lambda = \delta/s$ (situation b, s obtained from the experiment) at the $\alpha = 5\%$ level of significance. Adapted from Ref. [1].

progressively larger compared with z-values as n decreases. For instance, for s (instead of σ) = 0.8 and $|\mu_0 - \mu| = 1.5$, i.e. $\lambda = 1.87$, we derive that $n = 6$ (instead of 4, i.e. 1.5 times more). Incidentally, analytical chemists very often work with $n = 6$ replicates in studies concerning bias and usually do not know σ, which means they determine s. At the $\alpha = 0.05$, $\beta = 0.05$ level, they are therefore able to detect a bias of 1.87 s.

A philosophical point should be made here. As already stated above, the practice of hypothesis testing is biased towards the use of α over β. The reason is that statisticians applying their methods in biology and the social sciences were conservative and were mainly concerned to avoid jumping wrongly to the conclusion that there is an effect; they therefore stressed a low probability α of wrongly deciding there is an effect, when there is none. This is not necessarily always the best approach. In the example given in this section, it is just as important not to decide there is no effect (no bias) when in fact the effect exists. Not including considerations about β encourages sloppy work. Indeed, the best way of achieving a low α is a high σ (bad precision) and low n (few replicates). β then becomes high, i.e. the probability of missing an effect, although it exists, increases. Considerations about α and β depend strongly on the consequences of making a wrong decision. In the toxicological study of a drug, we should be as certain as possible not to conclude wrongly that there is no toxicological effect, since this would lead to undesirable side-effects in the drug. In pharmacological studies on the same drug, we would try to avoid wrongly concluding that there is an effect, since this would lead to the use of drugs that have no real therapeutic effect. The toxicologist needs a small β-error, the pharmacologist a small α-error.

If we do not know σ and determine s from the experiment in which we are investigating whether the bias or effect exists, there is of course a problem. If we do not know before the experiment how large s is, we cannot compute n. In this case, we can work as follows. If at the given α level an effect is detected, then we accept H_1 and reject H_0. If no effect is detected, eq. (4.1b) is used to determine how large β is for the n-value used, the s found and the observed δ. If this is smaller than a level set *a priori*, we accept H_0 and reject H_1. Otherwise, we note that we cannot reject H_0 and accept H_1, but reserve judgement because n was too low and β therefore too high.

It should be noted, that although this is the correct procedure, it is often not applied. However, this depends on the field of study. For instance, in clinical trials β considerations are often included to determine minimal sample size; in analytical chemistry and in chemistry in general they are usually not.

4.9 One- and two-sided tests

In the preceding sections the hypotheses were

H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$

or to write H_1 in another way

H_1: $\mu > \mu_0$ or $\mu < \mu_0$

In the context of the example given in these sections, it is just as bad to find that the analysis method yields too high ($\mu > \mu_0$) or too low ($\mu < \mu_0$) results. This is what is called a *two-sided, two-tail or two-tailed hypothesis*. The former term is preferred by ISO.

There are situations where we are concerned only about "greater than" or "smaller than", and not both of them at the same time. For instance, suppose that ore is bought to produce metal A. The seller guarantees that there is 10 g/kg of A in the ore. The buyer is interested only in ascertaining that there is enough A in the product. If there is more, this will be all the better. The hypothesis test will be formulated by the buyer as follows:

H_0: $\mu \geq 10$ g/kg

H_1: $\mu < 10$ g/kg

The hypothesis $\mu > 10$ g/kg will not be tested as such by the buyer. The buyer (consumer) tests that the risk of having less metal A than expected does not exceed a given probability. The producer might decide to test that he does not deliver more metal A than needed, in other words that he does not run a higher risk than that acceptable to him of delivering too much A. This leads to the concept of *consumer/producer risk* and the application of *acceptance sampling* techniques (see Chapter 20). Another example is the following. A laboratory is testing whether a substance remains stable on storage. The initial concentration is known to be 100.0 mg/l. After a certain time, the sample is analyzed six times. The mean $\bar{x}$ is 94.0 mg/l. It estimates a mean μ and one is concerned whether μ is lower than 100.0, taking into account that it is known that $\sigma = 8.0$ (i.e. a relative standard deviation of 8%). The hypotheses are

H_0: $\mu \geq 100.0$

H_1: $\mu < 100.0$

These are examples of a *one-sided, one-tail or one-tailed test.*

Let us consider the normal distribution of $\bar{x}$ around μ of Fig. 4.6. The hypotheses are:

H_0: $\mu \geq \mu_0$

H_1: $\mu < \mu_0$

At the $\alpha = 0.05$ level for a two-sided test, the interval in which H_0 will be accepted is between $z = -1.96$ and $z = +1.96$. For a one-sided test with H_1: $\mu < \mu_0$, we would accept an $\alpha = 0.05$ probability to incorrectly reject H_0, i.e. conclude that μ is lower than μ_0 when it is in fact at least equal. Therefore the decision limit must be set so that 5% of all cases fall below it, i.e. at $z = -1.65$.

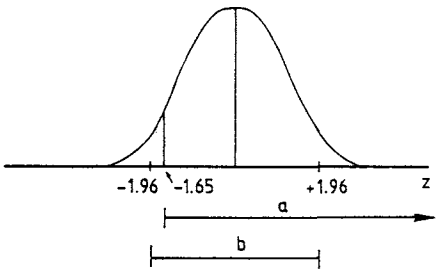


Fig. 4.6. One-sided decision limit (at −1.65) compared to two-sided limits (between −1.96 and +1.96). (a) Interval in which H_0 would be accepted for a one-sided test; (b) interval in which H_0 would be accepted for a two-sided test.

All $\bar{x}$ below $\mu_0 - 1.65(\sigma/\sqrt{n})$ will lead to rejection of H_0: $\mu \geq \mu_0$ and to acceptance of H_1: $\mu < \mu_0$. Observe that an effect is more easily detected. For instance for a calculated $z = -1.80$, the two-sided test would have led to acceptance, while the one-sided test, thanks to the additional information about the type of H_1, would lead to rejection.

In terms of confidence limits, we can reword the preceding paragraph as follows. All $\bar{x}$ such that $\bar{x} + 1.65(\sigma/\sqrt{n})$ is smaller than μ_0 will lead to rejection of H_0. We have used a one-sided confidence interval. We observe that the upper limit of the one-sided interval is closer to $\bar{x}$ than the same limit for the two-sided interval. The results between $z = -1.65$ and -1.96 now lead to rejection of H_0, while this would not have been the case for two-sided tests. We should also observe that the limit of a 95% one-sided interval is equal to that of the 90% two-sided confidence interval. The one-sided test detects more easily a difference and therefore it is more powerful than the two-sided test.

When there is an *a priori* reason to carry out the one-sided test instead of the two-sided test, this should be preferred. In this context a warning should be given. Let us go back to Example 1 of Section 4.1. We might (incorrectly) reason as follows. Since $\bar{x} = 98.2$, it is smaller than μ_0; therefore, the hypothesis $\mu > \mu_0$ should not be included. The original two-sided hypothesis test thus becomes a one-sided test and the test becomes more powerful. This reasoning is not acceptable. If there is no *a priori* reason to state a one-sided hypothesis test, then the hypothesis should be two-sided.

Let us now apply this to the stability on storage example and go through the steps of Section 4.3.

Step 1:

H_0: $\mu \geq \mu_0$

H_1: $\mu < \mu_0$ with $\mu_0 = 100.0$

Step 2:

$\alpha = 0.05$

Step 3: The confidence interval around $\bar{x}$ is a one-sided confidence limit given by

$$94.0 + 1.65 \frac{8.0}{\sqrt{6}} = 99.4$$

Step 4: $\mu_0 = 100.0$ is higher than the upper confidence limit.

Step 5: Reject H_0, accept H_1. The amount found is significantly lower than 100.0 mg/l. The substance is not stable on storage.

Step 6: $z = (100.0 - 94.0)/(8.0/\sqrt{6}) = 1.84$. p(one-sided) = 0.033.

The report would therefore include the statement $p < 0.05$ and preferably also: ($p = 0.033$).

It should be noted that we used z and the one-sided version of eq. (4.1a) because σ is known in our example. As in Section 4.3, t should be used instead of z in cases where this is appropriate.

We can also include β-considerations for a one-sided test. In this case the equation for the sample size n becomes

$$n = \left[(z_\alpha + z_\beta)\, \sigma / \delta\right]^2 \tag{4.6}$$

for known σ. Since $z_{\alpha/2}$ (eq. (4.4)) is larger than z_α (eq. (4.6)), n for a one-sided test may be smaller than for a two-sided one to have the same probability of not detecting the relevant difference δ.

4.10 An alternative approach: interval hypotheses

The hypotheses described so far are also called *point hypotheses* in contrast to the *interval hypotheses* we will shortly introduce in this section. We will do this with an example about the stability of drugs in biological fluids, which was described by Timm et al. [3]. The problem is to decide whether a drug in blood remains stable on storage. A certain amount of degradation (up to 10%) is considered acceptable. In statistical terminology, we want to exclude a degradation higher than 10% with 95% probability. Suppose that we add a known amount of 100.0 mg to the blood, then, following our procedures from Sections 4.1 to 4.9, we could only carry out a hypothesis test with as hypotheses:

H_0: $\mu \geq 100.0$ $\qquad$ H_1: $\mu < 100.0$

This is a one-sided test and therefore, to obtain the one-sided 95% confidence limit, we will compute the two-sided 90% confidence interval (see 4.9). Since the test is

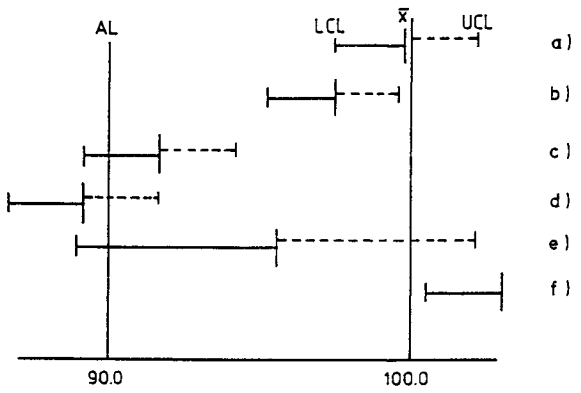


Fig. 4.7. Different situations in a stability on storage test. $\bar{x}$ is the mean result obtained, UCL, the upper one-sided 95% confidence limit used in the point hypothesis test and LCL the lower one-sided 95% confidence limit used in the interval hypothesis test. The interval between UCL and LCL is the 90% two-sided confidence interval.

one-sided, we will automatically accept H_0, when the mean result obtained by carrying out n replicate tests is higher than 100.0 (situation f) of Fig. 4.7. In situations a) and e) we would also conclude that there is no significant difference because 100.0 is situated in the confidence interval, in the one-sided case the interval $(-\infty)$ – UCL (upper confidence limit). In situations b) and c) we would find that the difference is significant, but we would probably add that since our best estimate, the mean, is higher than 90.0. the difference is acceptable. In d) we would conclude that there is a significant difference and that it is unacceptably high. This approach is not satisfying because we observe from Fig. 4.7 that in situations c) and e) there is a higher than 5% probability (one-sided) that the loss due to degradation exceeds 10.0, since the lower confidence limit is situated below 90.0. In other words, the β-error exceeds 5%.

In interval hypothesis testing, one accepts as not different those cases where the confidence interval around the estimated value is completely inside the acceptance interval, here 90.0 – $+\infty$. This confidence interval is now the interval LCL (lower confidence limit) – $(+\infty)$. In statistical terms the interval hypothesis for the one-sided example can be written as

H_0: $\mu \leq AL$

H_1: $\mu > AL$

where AL is the acceptance limit 90.0.

Let us again consider Fig. 4.7. Situations a) and f) are completely inside the acceptance interval. This is also the case for b). In this case, there is a significant degradation because 100.0 is not in the two-sided 90% confidence interval, but it

is not relevant since, with 95% probability, it does not exceed the 10% degradation limit.

In cases c), d) and e), we would reject the conclusion that there is no relevant degradation. This was already so for the point interval hypothesis for d) but not for situations c) and e). The rejection in c) is due to the fact that it is a marginal case. It is more probable that the degradation is acceptable than that it is not, but the probability that it is not is too high. In situation e) the uncertainty on the results is too high. This is connected with the β-error for the point hypothesis test: if we had computed the β-error for the point hypothesis test, we would have concluded that the β-error for finding a difference of 10.0 would have been too high. We observe that the interval hypothesis test is more conservative in accepting H_0, but also that it must be more useful in certain cases since it takes into account considerations about which difference is relevant and avoids high β-errors. For a two-sided test, we would write:

H_0: $LAL \geq \mu$ or $UAL \leq \mu$

H_1: $LAL < \mu < UAL$

where LAL and UAL are the lower acceptance limit and upper acceptance limit, respectively.

H_0 is rejected and therefore H_1 accepted at the 95% level of confidence when the 90% two-sided confidence interval around $\bar{x}$ is completely included in the acceptance interval. This 95–90 rule may appear strange, but it can be rationalized as follows. The interval hypothesis test can be described as consisting of two one-sided point hypothesis tests:

H_{01}: $\mu \leq LAL$ $\qquad$ H_{02}: $\mu \geq UAL$

H_{11}: $\mu > LAL$ $\qquad$ H_{12}: $\mu < UAL$

each being carried out at the 95% one-sided confidence level. H_0 will be rejected when both H_{01} and H_{02} fail and we should remember that the limit of a 95% one-sided confidence interval is equal to one of the limits of the 10% two-sided interval. When $\bar{x}$ < centre of the acceptance interval, one knows *a priori* that, when $LAL < \mu$ (H_{11} accepted), H_{12} will be accepted automatically since UAL must then be larger than μ. In practice therefore, in this simple situation a single one-sided hypothesis test is required. This is, however, not always the case.

This approach is not the more usual one. As explained, it has been proposed by Timm et al. [3] for stability studies of drugs in blood and also by Hartmann et al. [4] for method validation purposes, but it is not the standard approach in those two fields. It has been accepted, however, as the standard approach for bioequivalence studies [5]. Such studies are carried out to show that the bioavailability of a new drug formulation is comparable to that of an existing one. To avoid β-errors, i.e.

accepting bioavailability when it does not exist, one has come to apply the interval hypothesis testing approach. Although this approach is to be preferred in many cases, we will not apply it systematically in what follows because it is so unusual in most application fields.

References

1. G.E.P. Box, W.G. Hunter and J.S. Hunter, Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building. Wiley, New York, 1978.
2. ISO Standard 3494-1976, Statistical interpretation of data — Power of tests relating to means and variances, 1976.
3. U. Timm, M. Wall and D. Dell, A new approach for dealing with the stability of drugs in biological fluids. J. Pharm. Sci., 74 (1985) 972–977.
4. C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, Y. Vander Heyden, P. Vankeerberghen and D.L. Massart, Reappraisal of hypothesis testing for method validation: detection of systematic error by comparing the means of two methods or of two laboratories. Anal. Chem., 67 (1995) 4491–4499.
5. V.W. Steinijans and D. Hauschke, Update on the statistical analysis of bioequivalence studies. Int. J. Clin. Pharmacol. Ther. Toxicol., 28 (1990) 105–110.

Chapter 5

Some Important Hypothesis Tests

5.1 Comparison of two means

In the previous chapter we learned how to carry out a hypothesis test. We applied it to compare a sample mean, $\bar{x}$, with a known value, μ_0. Now the comparison of two independent sample means, $\bar{x}_1$ and $\bar{x}_2$, will be described.

Depending on the experimental design two different approaches have to be considered. The first is for the comparison of the means of *two independent samples*. In this case we want to compare the means of two populations and to do this a sample from each population is taken independently. For example to compare the nitrogen content in two different wheat flours replicate determinations in each flour are performed. Or, in the comparison of two digestion procedures prior to the determination of nitrogen in wheat flour, the same flour is analyzed independently by means of the two procedures.

The latter comparison could also be performed by means of *paired samples*. In that case different flours are used. An aliquot of each flour is analyzed by means of both procedures. The two samples thus obtained are paired, each pair being composed of the same flour. Consequently there is a one-to-one correspondence between the members of the samples which implies that there are equal numbers of observations in both samples. Pairing in this example is interesting since different flours are included in the comparison of the two digestion procedures.

5.1.1 Comparison of the means of two independent samples

Two different approaches which depend on the sample size can be considered.

5.1.1.1 Large samples

$\bar{x}_1$ and $\bar{x}_2$ are estimates of μ_1 and μ_2, based on respectively n_1 and n_2 observations. We have to test the hypothesis that there is no difference between μ_1 and μ_2. Therefore the null hypothesis is formulated as:

$H_0: \mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$)

and the alternative hypothesis, for a two-sided test:

$H_1: \mu_1 \neq \mu_2$ (or $\mu_1 - \mu_2 \neq 0$)

For a one-sided test the alternative hypothesis is:

$H_1: \mu_1 > \mu_2$ (or $\mu_1 - \mu_2 > 0$)

or

$H_1: \mu_1 < \mu_2$ (or $\mu_1 - \mu_2 < 0$)

depending on the problem that is considered.

For large samples, (n_1 and $n_2 \geq 30$), taken from any distribution of x (i.e. even a distribution that is not normal), the mean $\bar{x}$, is normally distributed with a variance σ^2/n (see Section 3.5). Consequently $\bar{x}_1$ and $\bar{x}_2$ are normally distributed variables with variance σ_1^2/n_1 and σ_2^2/n_2, respectively. If the null hypothesis is true, $\bar{x}_1 - \bar{x}_2$ is also normally distributed with a mean zero and variance $(\sigma_1^2/n_1 + \sigma_2^2/n_2)$, since the variance of the sum (or the difference) of two independent random variables is the sum of their variances.

Therefore the statistic used in the comparison of $\bar{x}_1$ and $\bar{x}_2$ is:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \tag{5.1}$$

If σ_1 and σ_2 are not known, s_1^2 and s_2^2 calculated from eq. (3.2) are considered as good estimators of the population variances σ_1^2 and σ_2^2. In that case one calculates:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \tag{5.2}$$

For a two-sided test H_0 is accepted if $|z|$, the absolute value of z obtained from eq. (5.1) or eq. (5.2), is smaller than the critical z-value at the chosen significance level. For a one-sided test with H_1 specified as $\mu_1 > \mu_2$, H_0 is accepted if $z < z_{crit}$. If H_1 is specified as $\mu_1 < \mu_2$, H_0 is accepted if $z > -z_{crit}$. At $\alpha = 5\%$, $z_{crit} = 1.96$ for a two-sided test and $z_{crit} = 1.645$ for a one-sided test.

The test can, of course, also be carried out by calculating the $(1 - \alpha)$ 100% confidence interval for $\mu_1 - \mu_2$. The 95% confidence interval for a two-sided test is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{s_1^2/n_1 + s_2^2/n_2} \tag{5.3}$$

The null hypothesis is accepted if this interval contains the value 0.

For a one-sided test the 95% confidence limit would be calculated as:

$$(\bar{x}_1 - \bar{x}_2) - 1.645 \sqrt{s_1^2/n_1 + s_2^2/n_2} \tag{5.4}$$

or

$$(\bar{x}_1 - \bar{x}_2) + 1.645 \sqrt{s_1^2/n_1 + s_2^2/n_2} \tag{5.5}$$

depending on whether the alternative hypothesis is $H_1:\mu_1 > \mu_2$ or $H_1:\mu_1 < \mu_2$, respectively. In the former situation the null hypothesis is accepted if the value 0 exceeds the lower confidence limit, in the latter situation if the value 0 is smaller than the upper confidence limit.

As an example, suppose that in the comparison of two digestion procedures prior to the determination of nitrogen in wheat flour, the following results are observed:

Procedure 1: $\bar{x}_1 = 2.05$ g/100 g $\quad s_1^2 = 0.050 \quad (n_1 = 30)$

Procedure 2: $\bar{x}_2 = 2.21$ g/100 g $\quad s_2^2 = 0.040 \quad (n_2 = 32)$

Procedure 1 was suspected beforehand of resulting in some loss of nitrogen during the digestion. Consequently the hypothesis to be tested is $H_0:\mu_1 = \mu_2$ against the alternative $H_1:\mu_1 < \mu_2$. The calculated z-value (eq. (5.2)) is obtained as:

$$z = \frac{2.05 - 2.21}{\sqrt{0.050/30 + 0.040/32}} = -2.96$$

Since the test is one-sided, at the 5% level of significance, this value has to be compared with –1.645. The null hypothesis is rejected and it can be concluded that procedure 1 indeed yields lower results ($p < 0.05$).

The calculation of the one-sided upper 95% confidence limit (eq. (5.5)) yields:

$$-0.16 + 1.645\sqrt{0.05/30 + 0.04/32} = -0.071$$

This is smaller than 0 and of course would lead to the same conclusion.

5.1.1.2 Small samples

As already mentioned in Chapter 3, if n is small and σ is not known, s^2 is not a precise estimator of the population variance. The tests described in the previous section, which are based on the normal distribution, cannot be applied if n_1 and/or $n_2 < 30$. The t-distribution (see Section 3.7) should be used instead. The t-test is now based on the following assumptions:

1. The samples with mean $\bar{x}_1$ and $\bar{x}_2$ are drawn from normally distributed populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. Tests described in Chapter 3 can be used to check this assumption. If normality cannot be shown or assumed (from previous knowledge) a non-parametric test (see Chapter 12) should be performed.

2. The variances σ_1^2 and σ_2^2 estimated by s_1^2 and s_2^2 are equal. In Section 5.4 it will be explained how this assumption can be tested. If the latter condition is fulfilled, a pooled variance s^2 (see Chapter 2), which is an estimate of the common variance of the two populations can be obtained:

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 + n_2 - 2)} \tag{5.6}$$

It is necessary before the estimation of the mean and the standard deviation of both samples to check that the data do not contain outlying observations which may have a large influence on these parameters. As shown in Section 3.8 the presence of outliers is a source of non-normality. Tests for the detection of outliers are described in Section 5.5.

The t-test is then performed by calculating the statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}} \tag{5.7}$$

This calculated t-value is compared with the critical t-value at the chosen significance level, α, and $n_1 + n_2 - 2$ degrees of freedom. For a two-sided test H_0 is accepted if $|t| < t_{crit}$. For a one-sided test with H_1 specified as $\mu_1 > \mu_2$, H_0 is accepted if $t < t_{crit}$. If H_1 is specified as $\mu_1 < \mu_2$, H_0 is accepted if $t > -t_{crit}$.

In a similar way as described earlier, the test could also be carried out by calculating the $(1 - \alpha)$ 100% confidence interval for $\mu_1 - \mu_2$.

The same example as in the previous section but with smaller sample sizes n_1 and n_2 will be considered:

Procedure 1: $\bar{x}_1 = 2.05$ g/100 g $\quad s_1^2 = 0.050 \quad (n_1 = 8)$

Procedure 2: $\bar{x}_2 = 2.21$ g/100 g $\quad s_2^2 = 0.040 \quad (n_2 = 7)$

It will be assumed that both populations from which the samples are drawn are normally distributed. From the F-test (see Section 5.4) it can be concluded that the hypothesis $\sigma_1^2 = \sigma_2^2$ is acceptable. Consequently, at the 5% significance level, the null hypothesis H_0: $\mu_1 = \mu_2$ can be tested against the alternative H_1: $\mu_1 < \mu_2$ by means of a t-test. First the pooled variance s^2 (eq. (5.6)) is calculated:

$$s^2 = \frac{7 \times 0.050 + 6 \times 0.040}{13} = 0.045$$

The calculated t (eq. 5.7) is obtained as:

$$t = \frac{-0.16}{\sqrt{0.045(1/7 + 1/8)}} = -1.46$$

The critical t-value for a one-sided test and 13 degrees of freedom is 1.771 (see Table 3.4). Since $t > -t_{crit}$ the null hypothesis can be accepted. Consequently from these results no difference between the two digestion procedures can be detected.

The calculation of the one-sided upper 95% confidence limit, which is obtained from

$$(\bar{x}_1 - \bar{x}_2) + t_{0.05}\sqrt{s^2(1/n_1 + 1/n_2)}$$

and yields

$$-0.16 + 1.771\sqrt{0.045(1/7 + 1/8)} = 0.034$$

would lead to the same conclusion since this limit exceeds 0.

If the condition of a homogeneous variance ($\sigma_1^2 = \sigma_2^2$) is not fulfilled, the test cannot be applied as described earlier since a pooled variance cannot be obtained. The Cochran test can then be used. It is based on the comparison of

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \tag{5.8}$$

with a critical value given by:

$$t' = \frac{t_1(s_1^2/n_1) + t_2(s_2^2/n_2)}{(s_1^2/n_1) + (s_2^2/n_2)} \tag{5.9}$$

where, for $\alpha = 0.05$, t_1 represents the critical t-value for $n_1 - 1$ degrees of freedom and t_2 represents the critical t-value for $n_2 - 1$ degrees of freedom. With $n_1 = n_2$, $t' = t_1 = t_2$. The test statistic in eq. (5.8) is obtained in the same way as in eq. (5.2) but it is tested differently.

Consider for the previous example the following results:

Procedure 1: $\bar{x}_1 = 2.05$ g/100 g $\quad s_1^2 = 0.050 \quad (n_1 = 9)$

Procedure 2: $\bar{x}_2 = 2.21$ g/100 g $\quad s_2^2 = 0.010 \quad (n_2 = 8)$

The means have not changed but the variances and sample sizes are different. From the F-test (see Section 5.4) it is concluded that $\sigma_1^2 \neq \sigma_2^2$. Therefore the Cochran test has to be used for the comparison of both means. The calculated t-value:

$$t = \frac{0.16}{\sqrt{\frac{0.050}{9} + \frac{0.010}{8}}} = -1.94$$

has to be compared with $-t'$. Since from

$$t' = \frac{1.860(0.050/9) + 1.895(0.010/8)}{(0.050/9) + (0.010/8)} = 1.87$$

it follows that $t < -t'$ it is concluded that procedure 1 yields lower values ($p < 0.05$).

5.1.2 Comparison of the means of two paired samples

As already mentioned earlier two samples are paired if there is a one-to-one correspondence between the members of the samples. An example will illustrate

this. The nitrogen amount is determined in 8 different flour samples. For the digestion of each of these samples two different procedures are used. The results are the following:

Flour	1	2	3	4	5	6	7	8
Procedure 1	2.0	1.4	2.3	1.2	2.1	1.5	2.4	2.0
Procedure 2	1.8	1.5	2.5	1.0	2.0	1.3	2.3	2.1

These are paired samples since each value obtained for the first procedure has to be compared with a specific value obtained for the second procedure. This situation obviously is different from that described in Section 5.1.1 where the samples are independent since there is no specific connection between the observations from the samples. This connection exists in the above example since what is important is the comparison of 2.0 and 1.8 which both are results for the first flour, 1.4 and 1.5 which are the results for the second flour and so on. That information is taken into account by considering the differences between the paired observations:

$d_i = x_{1i} - x_{2i}$

The mean of these differences is:

$$\bar{d} = \frac{\sum d_i}{n} \tag{5.10}$$

where n represents the number of pairs. $\bar{d}$ is an estimate of the true but unknown mean difference δ. If there is no difference between the means obtained by both procedures $\delta = 0$. Therefore the null hypothesis can be formulated as:

H_0: $\delta = 0$

and the alternative hypothesis for a two-sided test:

H_1: $\delta \neq 0$

For a one-sided test the alternative hypothesis is:

H_1: $\delta > 0$

or

H_1: $\delta < 0$

In this way the problem has been reduced to the comparison of a mean with a given value (here 0) and tests similar to those described in Chapter 4 can be performed. Depending on the sample size they are based on a normal distribution or on a t-distribution.

5.1.2.1 Large samples

For large samples ($n \geq 30$) the statistic

$$z = \frac{\bar{d} - 0}{s_d / \sqrt{n}} \tag{5.11}$$

is calculated; s_d is the standard deviation of the differences and consequently $s_d/\sqrt{n}$ is the standard deviation of the mean difference. At the significance level α this calculated value has to be compared with the critical z-value which is 1.96 for a two-sided test and 1.645 for a one-sided test. For a two-sided test H_0 is accepted if $|z| < z_{crit}$. For a one sided test with H_1 specified as $\delta > 0$, H_0 is accepted if $z < z_{crit}$. If H_1 is specified as $\delta < 0$, H_0 is accepted if $z > -z_{crit}$.

5.1.2.2 Small samples

For small samples ($n < 30$) a t-test is performed

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}} \tag{5.12}$$

where t has $(n - 1)$ degrees of freedom. The test is only valid if the differences d_i are normally distributed and have the same variance.

If for our example of paired samples mentioned earlier, we want to know whether the two digestion procedures yield the same results (two-sided test; $\alpha = 0.05$) the calculations proceed as described in Table 5.1. Since

$$\bar{d} = \frac{\sum d_i}{n} = 0.05$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} = 0.16$$

TABLE 5.1

Comparison of the means of two paired samples

Flour	Procedure 1	Procedure 2	d_i
1	2.0	1.8	0.2
2	1.4	1.5	–0.1
3	2.3	2.5	–0.2
4	1.2	1.0	0.2
5	2.1	2.0	0.1
6	1.5	1.3	0.2
7	2.4	2.3	0.1
8	2.0	2.1	–0.1

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{0.05}{0.16/\sqrt{8}} = 0.88$$

The critical t-value for a two-sided test and 7 degrees of freedom is 2.365 (see Table 3.4). Since $|t| < t_{crit}$ there is no significant difference between both digestion procedures for the analysis of nitrogen in flour.

5.2 Multiple comparisons

If more than two means have to be compared we could reason that 2 by 2 comparisons using a t-test will reveal which means are significantly different from each other. However in these comparisons the same mean is used several times and consequently the t-tests are not independent of each other. As a result, when all population means are equal, the probability that at least one comparison will be found to be significant increases. Even if all population means are equal, the more comparisons are made the more probable it is that one or more pairs of means are found to be statistically different. One way to overcome this problem is by adjusting the α value of the individual comparisons, α' such that the overall or joint probability corresponds to the desired value. α' is then obtained from:

$$\alpha' = 1 - (1 - \alpha)^{1/k} \tag{5.13}$$

with α' the significance level for the individual comparisons; α the overall significance level; k the number of comparisons.

This adjustment is sometimes referred to as *Bonferroni's adjustment* and the overall significance level, α, is called the *experimentwise error rate*. If α is small α' can also be approximated by α/k.

For example in the comparison of 5 means, 10 t-tests have to be performed ($k = 10$). If, when the null hypothesis is true, we want an overall probability of at least 95% ($\alpha = 0.05$) that all the observed means are equal, we have to take $\alpha' = 0.005$. Therefore the individual comparisons have to be performed at a significance level of 0.005. Critical t-values at a significance level of 0.005 have then to be used in the comparisons to ensure an overall significance level $\alpha = 0.05$.

It follows that, the more comparisons are made, the larger the differences between the pairs of means must be in order to decide, from a multiple comparison, that they are significant. In the following example 5 different digestion procedures for the determination of N in flour have been applied. The results obtained are:

Procedure	1	2	3	4	5
$\bar{x}_i$	2.21	2.00	1.95	2.15	2.20
s_i^2	0.04	0.05	0.05	0.03	0.04
n_i	8	8	8	8	8

TABLE 5.2

Calculated t-values for the comparison of 5 different digestion procedures to determine N in flour.

Comparison	s^2_{pooled}	$\lvert\bar{x}_1 - \bar{x}_2\rvert$	$\lvert t_{cal}\rvert$
1–2	0.045	0.21	1.98
1–3	0.045	0.26	2.45
1–4	0.035	0.06	0.64
1–5	0.040	0.01	0.10
2–3	0.050	0.05	0.45
2–4	0.040	0.15	1.50
2–5	0.045	0.20	1.89
3–4	0.040	0.20	2.00
3–5	0.045	0.25	2.36
4–5	0.035	0.05	0.53

Do some of the procedures yield significantly different results?

The calculated t-values (t_{cal}) for the 10 different possible comparisons, as calculated for two independent samples from eq. (5.7), are summarized in Table 5.2.

In order to ensure an overall significance level $\alpha = 0.05$, the individual comparisons have to be performed at a significance level $\alpha' = 0.005$. Since in our example the means obtained for the different digestion procedures are based on the same number of observations ($n_i = 8$) all calculated t-values must be compared with the same (two-sided) critical t-value, $t_{0.005,14} = 3.326$. Consequently no differences between the digestion procedures can be detected since all calculated t-values are lower than this critical t-value. Note that if the individual comparisons were incorrectly performed at a significance level $\alpha = 0.05$, two significant tests would result since $t_{0.05,14} = 2.145$.

The Bonferroni adjustment, as given in eq. (5.13) is also necessary for t-tests that do not involve computations from the same data. Such independent t-tests are for example computed in the following situation. To validate a method (see also Section 13.5.4) the whole range of concentrations for which the method is intended to be used must be considered. Therefore recovery experiments at different concentration levels, covering the range to be determined can be performed. The validation involves t-tests, at each concentration level, to compare the mean concentration found with the known concentration added (see Chapter 4). These t-tests are independent since each of them uses different data. To make a joint confidence statement that e.g. with 95% probability all found and added concentrations are equal, each individual t-test has to be performed at a significance level α' as given by eq. (5.13).

Other multiple comparison procedures are described in Section 6.3 where the analysis of variance (ANOVA) is introduced. An analysis of variance reveals

whether several means can be considered to be equal. However if they are found not to be equal ANOVA does not indicate which mean (or means) are different from the others. If this information is wanted multiple comparison procedures have also to be used.

5.3 Beta error and sample size

The β error has been defined in Section 4.7 as the probability of incorrectly accepting the null hypothesis when, in fact, the alternative hypothesis is true. Here it corresponds to the probability that for given sample sizes (n_1 and n_2) and variances (σ^2 or s^2) and for a specified significance level α, a certain difference between the two means will not be revealed by the test used although it exists. Alternatively, it is possible to determine the sample size ($n_1 = n_2 = n$) necessary to detect a certain difference between the two means so that there is 100 α% probability of detecting a difference when in fact there is none and 100 β% probability of not detecting a difference when it does exist.

Both these problems can be solved by using graphs published by ISO [1], which for the first kind of problem allow us to determine β and for the second kind of problem allow us to determine the sample size, n. Such a graph has already been introduced in Section 4.7 with respect to a test of the difference between a mean and a given value. It is impossible here to explain all the different graphs necessary to determine the β error and the sample size for the different tests previously described (z and t-tests, one and two-sided tests, $\alpha = 0.5$ and $\alpha = 0.1$). One example will illustrate how to use them here. The data are those from Section 5.1.1.2. Suppose that we wish to know the probability that a real difference between the means of the two digestion procedures equal to 0.15, will not be detected. According to ISO [1], the following value has to be calculated:

$$\lambda = \frac{|\mu_1 - \mu_2|}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

where s^2 is the pooled variance which for our example equals 0.045. Consequently $\lambda = 0.15/\sqrt{0.045(1/8 + 1/7)} = 1.4$. Figure 5.1 shows the value of β as a function of λ for a one-sided test and $\alpha = 0.05$. For 13 degrees of freedom one finds (by interpolation) β to be about 0.6. Consequently, using the t-test with a significance level $\alpha = 0.05$, the probability of not detecting a real difference between the two means equal to 0.15 (s^2 being equal to 0.045) is about 60%. This value can be reduced by increasing the sample size and this can be derived from Fig. 5.2 which shows the value of n ($= n_1 = n_2$) as a function of λ. The latter is now calculated as:

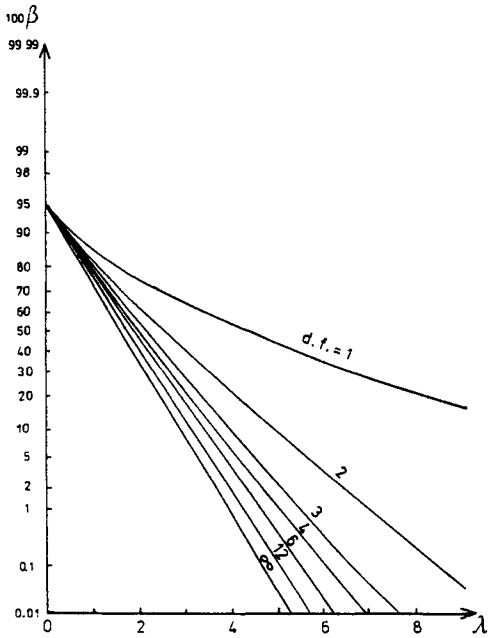


Fig. 5.1. Operating characteristic curve for the one-sided t-test ($\alpha = 0.05$). For the meaning of λ see text. Adapted from Ref. [1]

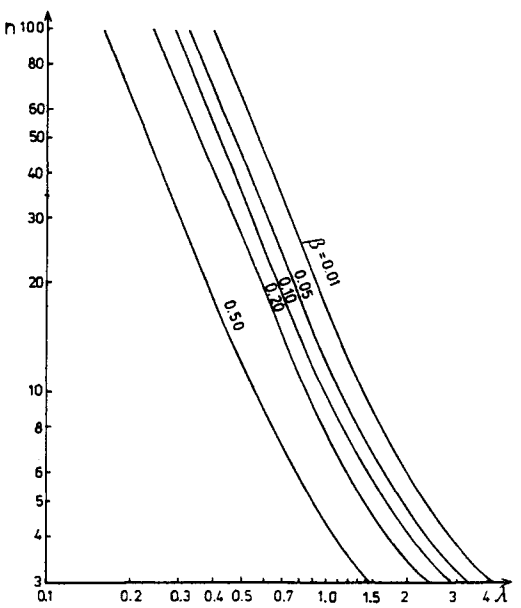


Fig. 5.2. Sample size required to detect a certain bias with the one-sided t-test ($\alpha = 0.05$). For the meaning of λ see text. Adapted from Ref. [1].

$$\lambda = \frac{|\mu_1 - \mu_2|}{s\sqrt{2}} = 0.5$$

If we decide that we do not want a probability of more than 0.10 of accepting the hypothesis that $\mu_1 = \mu_2$ when actually $\mu_1 - \mu_2 = 0.15$ it follows from Fig. 5.2 that n should be at least 36.

5.4 Comparison of variances

5.4.1 Comparison of two variances

The comparison of two variances σ_1^2 and σ_2^2, estimated by s_1^2 and s_2^2 is performed by means of an F-test:

$$F = s_1^2/s_2^2 \tag{5.14}$$

in which s_1^2 is the larger of the two variances. By dividing the largest variance by the smallest variance an F-value equal to or larger than unity is obtained. This calculated F is compared with the critical F-value at the chosen significance level. This critical value is derived from tables of the F distribution such as the one shown in Table 5.3. The critical value, which depends on the two sample sizes, is found at the intersection of the column df_1 (= $n_1 - 1$ = the degrees of freedom corresponding to s_1^2) and the row df_2 (= $n_2 - 1$ = the degrees of freedom corresponding to s_2^2). The F-test performed can again be two-sided (H_0: $\sigma_1^2 = \sigma_2^2$;H_1: $\sigma_1^2 \neq \sigma_2^2$) or one-sided ($H_0$: $\sigma_1^2 = \sigma_2^2$;$H_1 = \sigma_1^2 > \sigma_2^2$).

For the first example treated in Section 5.1.1.2 where two procedures for the determination of nitrogen were compared it was concluded that the two variances σ_1^2, estimated by $s_1^2 = 0.05$ ($n_1 = 8$), and σ_2^2, estimated by $s_2^2 = 0.04$ ($n_2 = 7$), are equal ($\alpha = 0.05$). This conclusion was reached as follows:

$F = 0.05/0.04 = 1.25$

Since the alternative hypothesis is H_1: $\sigma_1^2 \neq \sigma_2^2$ a two-sided test has to be performed. Therefore the critical F-value is obtained from Table 5.3.a (the critical F-value for a one-sided test at $\alpha = 0.025$ corresponds to the critical F value for a two-sided test at $\alpha = 0.05$). Since $df_1 = 7$ and $df_2 = 6$, $F_{0.05;7,6} = 5.70$. The calculated F-value (1.25) being smaller than the critical value (5.70), the null hypothesis that both variances are equal is accepted.

For the F-test ISO [1] also gives graphs that allow us, for the particular case where the two samples are of the same size, to determine the β error or for a given β to determine the common size n of the samples.

TABLE 5.3a

Critical F-values for a one-tailed test ($\alpha = 0.025$)

	df_1																		
df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647	779	864	899	922	937	948	956	963	968	976	985	993	997	1001	1005	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

TABLE 5.3b

Critical F-values for a one-tailed test ($\alpha = 0.05$)

	df_1																		
df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	199	215	224	230	234	237	239	240	242	244	246	248	249	250	251	252	253	254
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

5.4.2 Comparison of a variance with a given value

To compare a variance, s^2, with a known value σ_0^2 the following test statistic is generally calculated:

$$T = (n-1)s^2/\sigma_0^2 \tag{5.15}$$

Since T is distributed as χ^2 with n-1 degrees of freedom [2] the test consists in comparing the calculated T with the tabulated χ^2 given in Table 5.4. For a two-sided test at the 5% significance level (H_0: $\sigma^2 = \sigma_0^2$; H_1: $\sigma^2 \neq \sigma_0^2$) the null hypothesis is rejected if $T \geq \chi^2_{0.025;n-1}$ or if $T \leq \chi^2_{0.975;n-1}$. For the one-sided test, H_0: $\sigma^2 = \sigma_0^2$; H_1: $\sigma^2 > \sigma_0^2$, H_0 is rejected if $T \geq \chi^2_{0.05}$ while for the one-sided test, H_0: $\sigma^2 = \sigma_0^2$; H_1: $\sigma^2 < \sigma_0^2$, the null hypothesis is rejected if $T \leq \chi^2_{0.95}$.

TABLE 5.4

Critical values of Chi-square (the α-values represent the area to the right of the critical χ^2 in one tail of the distribution)

df\$^{\alpha}$	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.001
1	0.000 2	0.001 0	0.003 9	0.015 8	2.71	3.84	5.02	6.63	10.83
2	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	13.82
3	0.12	0.22	0.35	0.58	6.25	7.81	9.35	11.34	16.27
4	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	18.47
5	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	20.52
6	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	22.46
7	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.47	24.32
8	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	26.13
9	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	27.88
10	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	29.59
11	3.05	3.82	4.57	5.58	17.27	19.67	21.92	24.72	31.26
12	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	32.91
13	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	34.53
14	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	36.12
15	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	37.70
16	5.81	6.91	7.96	9.31	23.54	26.30	28.84	32.00	39.25
17	6.41	7.56	8.67	10.08	24.77	27.59	30.19	33.41	40.79
18	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.80	42.31
19	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	43.82
20	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	45.32
21	8.90	10.28	11.59	13.24	29.61	32.67	35.48	38.93	46.80
22	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	48.27
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	49.73
24	10.86	12.40	13.85	15.66	33.20	36.41	39.37	42.98	51.18
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	52.62
26	12.20	13.84	15.38	17.29	35.56	38.88	41.92	45.64	54.05
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	55.48
28	13.57	15.31	16.93	18.94	37.92	41.34	44.46	48.28	56.89
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	58.30
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	59.70

For example, we want to test if the variance, $s^2 = 1.24$, obtained from 11 observations has a true known value, $\sigma_0^2 = 1$ (H_0: $\sigma^2 = 1$; H_1: $\sigma^2 \neq 1$). T calculated from eq. (5.15) equals 12.4. Since $\chi^2_{0.025;10} = 20.48$ and $\chi^2_{0.975;10} = 3.25$ the null hypothesis is accepted that σ^2, estimated by s^2, equals 1.

It can be verified from Tables 5.3 and 5.4 that $\chi^2_{\alpha;n-1} = (n-1)\, F_{\alpha;n-1,\infty}$ and $\chi^2_{1-\alpha;n-1} = (n-1)/F_{\alpha;\infty,n-1}$. Therefore an F-test could also be applied by computing:

$$F = s^2/\sigma_0^2 \quad \text{if} \quad s^2 > \sigma_0^2$$

or

$$F = \sigma_0^2/s^2 \quad \text{if} \quad \sigma_0^2 > s^2$$

The former F-value is compared with $F_{\alpha;n-1,\infty}$ and the latter with $F_{\alpha;\infty,n-1}$ in the usual way.

The test can also be performed by considering the 95% confidence interval for σ^2 which for our example is obtained as:

$$\frac{n-1\ s^2}{\chi^2_{0.025;n-1}} < \sigma^2 < \frac{n-1\ s^2}{\chi^2_{0.975;n-1}} \tag{5.16}$$

or also as:

$$\frac{s^2}{F_{0.025;n-1,\infty}} < \sigma^2 < s^2\, F_{0.025;\infty,n-1}$$

For our example this yields:

$$\frac{12.4}{20.48} < \sigma^2 < \frac{12.4}{3.25}$$

or also:

$$\frac{1.24}{2.05} < \sigma^2 < 1.24 \times 3.08$$

Therefore:

$$0.61 < \sigma^2 < 3.82 \qquad \text{and} \qquad 0.78 < \sigma < 1.95$$

The known true variance, $\sigma_0^2 = 1$, being contained in this interval, the null hypothesis that $\sigma^2 = 1$ is accepted. Notice that, since the χ^2 distribution is not symmetrical the confidence interval cannot be written in the form $s^2 \pm e$.

5.5 Outliers

Several of the hypothesis tests described up to now assume that the data are normally distributed. From what we learned in Chapter 3 we know that the normal distribution is completely characterized by the mean and the standard deviation. However the presence of an outlier, a value which is not representative for the rest of the data, can have a great influence on these parameters. Different outlier tests have been described but the problem is that they do not always yield the same result. Rechenberg [3] compared eight different procedures to test 4 suspect values in a series of 21 observations. Depending on the test used, zero or up to four outliers were detected. This shows that outlier rejection by statistical tests should not be carried out as a routine matter. They should rather be performed to identify problem samples. It is important to carefully examine whether an assignable cause for the outlier(s) can be found (e.g. a clerical error, a computational error, an error in the analysis). If this is the case, the outlier can be corrected or removed from the data. The occurrence of multiple outliers can be an indication that the analysis method is not under control and that corrective actions have to be taken.

As indicated by the Analytical Methods Committee of the Royal Chemical Society [4], rejection of outliers on a statistical basis from data aimed at defining the variability of an analytical method, can result in an important underestimation of the variance. This is also illustrated by Goetsch and coworkers [5] who evaluated some outlier problems in collaborative studies.

If outliers are removed from a data set it should always be reported that outlying observations were present.

5.5.1 Dixon's test

Dixon's test is one of the most popular tests for the detection of an outlier because it is easy to calculate. It is based on a comparison of the difference between the suspect value and its direct or a close neighbour with the overall range or a modified range. Consider a set of n data $x_i (i = 1,2,...n)$ arranged in order of increasing magnitude. Depending on the sample size the following test statistics are calculated:
for $n = 3$ to 7:

$$Q_{10} = (x_2 - x_1) / (x_n - x_1)$$

or

$$Q_{10} = (x_n - x_{n-1}) / (x_n - x_1)$$

depending on whether x_1 or x_n is the suspect value

for $n = 8$ to 12:

$$Q_{11} = (x_2 - x_1) / (x_{n-1} - x_1)$$

or

$$Q_{11} = (x_n - x_{n-1}) / (x_n - x_2)$$

for $n \geq 13$

$$Q_{22} = (x_3 - x_1) / (x_{n-2} - x_1)$$

or

$$Q_{22} = (x_n - x_{n-2}) / (x_n - x_3)$$

In the literature sometimes another statistic ($Q_{21} = (x_3 - x_1)/(x_{n-1} - x_1)$) or ($Q_{21} = (x_n - x_{n-2})/(x_n - x_2)$) and other critical values are found for $n = 11$ to 13.

The calculated Q-value is compared in the usual way with the critical value at the chosen significance level. An outlier is detected if the calculated Q exceeds the critical Q. Critical Q-values are given in Table 5.5. The one-sided values in the table apply to test an observation at a predesigned end of the data set while for an observation that seems suspect after an inspection of the data the two-sided values have to be used.

As an example consider the following data arranged in increasing order: 22.1, 22.4, 22.9, 23.0, 23.5, 23.7, 23.9, 26.5. If, after the inspection of these data, we suspect the value 26.5 of being too high the following statistic is calculated ($n = 8$):

$$Q_{11} = (x_n - x_{n-1}) / (x_n - x_2)$$
$$= (26.5 - 23.9) / (26.5 - 22.4)$$
$$= 0.634$$

From Table 5.5 the critical Q-value for $n = 8$ and $\alpha = 0.05$ is 0.608. The calculated value is larger and therefore 26.5 is considered to be an outlier at the 0.05 level of significance.

Problems can arise when the test is repeatedly used for the detection of multiple outliers since these can mask each other. It can be checked that if in the above example the value 23.9 is changed to 26.0, yielding a data set with two suspect values, no outlier is detected since $Q_{11} = (26.5 - 26.0) / (26.5 - 22.4) = 0.122$. Multiple outlier tests such as described further are then more appropriate.

The Dixon test is the outlier test originally recommended by ISO [7] for inter-laboratory tests. A table of 2-sided critical values is used here since in collaborative studies outliers at both ends of the data set are equally likely. In its

TABLE 5.5

Critical Q-values for testing outliers (extracted from a more extensive table by Beyer [6])

n		One-sided α		Two-sided α	
		0.05	0.01	0.05	0.01
3	Q_{10}	0.941	0.988	0.970	0.994
4		0.765	0.889	0.829	0.926
5		0.642	0.780	0.710	0.821
6		0.560	0.698	0.628	0.740
7		0.507	0.637	0.569	0.680
8	Q_{11}	0.554	0.683	0.608	0.717
9		0.512	0.635	0.564	0.672
10		0.477	0.597	0.530	0.635
11		0.450	0.566	0.502	0.605
12		0.428	0.541	0.479	0.579
13	Q_{22}	0.570	0.670	0.611	0.697
14		0.546	0.641	0.586	0.670
15		0.525	0.616	0.565	0.647
16		0.507	0.595	0.546	0.627
17		0.490	0.577	0.529	0.610
18		0.475	0.561	0.514	0.594
19		0.462	0.547	0.501	0.580
20		0.450	0.535	0.489	0.567
21		0.440	0.524	0.478	0.555
22		0.430	0.514	0.468	0.544
23		0.421	0.505	0.459	0.535
24		0.413	0.497	0.451	0.526
25		0.406	0.489	0.443	0.517
26		0.399	0.486	0.436	0.510
27		0.393	0.475	0.429	0.502
28		0.387	0.469	0.423	0.495
29		0.381	0.463	0.417	0.489
30		0.376	0.457	0.412	0.483

latest draft document however ISO [8] prefers the single and double Grubbs' test explained in the next section. ISO also gives a procedure for the treatment of outliers which is described in Chapter 14. The test is repeatedly performed until no more extreme values are detected. Outlying observations, which are significant at the 1% level are called outliers and are always removed. If the outlying observations are significant at the 5% level they are called *stragglers* and are only discarded if they can be explained.

5.5.2 Grubbs' test

The maximum normalized deviation test described by Grubbs and Beck [9] is based on the calculation of:

$$G = (x_i - \bar{x}) / s \tag{5.17}$$

with x_i the suspected outlier (either the highest or the lowest result), $\bar{x}$ the sample mean and s the sample standard deviation. The absolute value of G is compared with the critical values for one largest or one smallest value given in Table 5.6.

The Grubbs' statistic for the detection of two outliers (either the two highest or the two lowest results) is obtained as:

$$G = SS_{n-1,n} / SS_0 \quad \text{or} \quad G = SS_{1,2} / SS_0 \tag{5.18}$$

with SS_0 the sum of squared deviations from the mean for the original sample ($= \sum(x_i - \bar{x})^2$) and $SS_{n-1,n}$ and $SS_{1,2}$ the sum of squared deviations obtained after removal of the two highest or the two lowest values, respectively. Critical values for the double-Grubbs' test (two largest or two smallest values) are also given in Table 5.6. Notice that here outliers are detected if the test statistic of eq. (5.18) is smaller than the critical value.

The single-Grubbs' test can also be performed by calculating the percentage reduction in the standard deviation when the suspect point is rejected:

$$R = 100\,(1 - s_1 / s) \tag{5.19}$$

with s the original sample standard deviation and s_1 the standard deviation obtained after removal of the suspect value. This test is equivalent with the one of eq. (5.17) because their critical values are related [10]. The latter (eq. (5.19)) is recommended for the detection of a single outlier in collaborative studies by the AOAC [11].

The Grubbs' pair statistic for the detection of 2 outliers, which is also part of the AOAC procedure is calculated in the same way but in this test s_1 is the standard deviation obtained after removal of a pair of suspect values (either situated at the same or different ends of the data sets). Critical values for two-sided single value and pair value tests performed in this way can be found in references [10] and [11].

Application of the single outlier test (eq. (5.17)) to our example yields the following G-value

$$G = (26.5 - 23.5)/1.36 = 2.206$$

which is larger than the two-sided critical value for $n = 8$ and $\alpha = 0.05$ (2.126).

For the data set with two suspect values (22.1; 22.4; 22.9; 23.0; 23.5; 23.7; 26.0; 26.5) application of the double-Grubbs' test yields:

TABLE 5.6

Two-sided critical values for the Grubbs' test. (For the single Grubbs' test outliers give rise to values which are larger than the critical values while for the double Grubbs' test they give rise to values which are smaller than the critical values).

n	One largest or One smallest α		Two largest or Two smallest α	
	0.05	0.01	0.05	0.01
3	1.155	1.155	–	–
4	1.481	1.496	0.0002	0.0000
5	1.715	1.764	0.0090	0.0018
6	1.887	1.973	0.0349	0.0116
7	2.020	2.139	0.0708	0.0308
8	2.126	2.274	0.1101	0.0563
9	2.215	2.387	0.1492	0.0851
10	2.290	2.482	0.1864	0.1150
11	2.355	2.564	0.2213	0.1448
12	2.412	2.636	0.2537	0.1738
13	2.462	2.699	0.2836	0.2016
14	2.507	2.755	0.3112	0.2280
15	2.549	2.806	0.3367	0.2530
16	2.585	2.852	0.3603	0.2767
17	2.620	2.894	0.3822	0.2990
18	2.651	2.932	0.4025	0.3200
19	2.681	2.968	0.4214	0.3398
20	2.709	3.001	0.4391	0.3585
21	2.733	3.031	0.4556	0.3761
22	2.758	3.060	0.4711	0.3927
23	2.781	3.087	0.4857	0.4085
24	2.802	3.112	0.4994	0.4234
25	2.822	3.135	0.5123	0.4376
26	2.841	3.157	0.5245	0.4510
27	2.859	3.178	0.5360	0.4638
28	2.876	3.199	0.5470	0.4759
29	2.893	3.218	0.5574	0.4875
30	2.908	3.236	0.5672	0.4985
31	2.924	3.253	0.5766	0.5091
32	2.938	3.270	0.5856	0.5192
33	2.952	3.286	0.5941	0.5288
34	2.965	3.301	0.6023	0.5381
35	2.979	3.316	0.6101	0.5469
36	2.991	3.330	0.6175	0.5554
37	3.003	3.343	0.6247	0.5636
38	3.014	3.356	0.6316	0.5714
39	3.025	3.369	0.6382	0.5789
40	3.036	3.381	0.6445	0.5862

$G = SS_{n-1,n}/SS_0 = 1.89 / 18.52 = 0.1021$

where

$SS_{n-1,n} = (22.1 - 22.9)^2 + (22.4 - 22.9)^2 + ... + (23.7 - 22.9)^2$

$SS_0 = (22.1 - 23.8)^2 + (22.4 - 23.8)^2 + ... + (26.5 - 23.8)^2$

Since the G-value is smaller than the two-sided critical value for the double-Grubbs' test (0.1101 for $n = 8$ and $\alpha = 0.05$) the two highest values are considered to be outliers. We come to the same conclusion if we consider the data set introduced in Section 3.8 (Table 3.6). The non-normality of the data was ascribed to the two highest values. The double-Grubbs' test reveals that these values indeed are outliers since:

$G = SS_{n-1,n} / SS_0 = 0.179 / 1.670 = 0.1072$

which is smaller than the two-sided critical value for $n = 20$ and $\alpha = 0.05$ (0.4391).

As already mentioned ISO [8] now recommends the use of the single and double Grubbs' test as just described.

5.6 Distribution tests

Distribution tests or *goodness-of-fit tests* allow us to test whether our data follow a particular probability distribution. They are based on the comparison of an observed distribution with an expected or theoretical distribution. In this section the Chi-square and the Kolmogorov–Smirnov test are introduced to test normality. This is an important application since most statistical tests are based on the assumption that the data follow a normal distribution. Both tests are appropriate for the following situations:

- when the theoretical distribution is completely specified. For the normal distribution this means that σ and μ are known. The observed distribution is then compared with a particular normal distribution with known σ and μ.
- when the theoretical distribution is derived from the data themselves. In that case σ and μ are estimated by the sample standard deviation, s, and mean, $\bar{x}$, respectively, and the question is whether the distribution can be considered to be normal.

Since most often it is required to test whether the data are normally distributed and not whether they follow a particular normal distribution only the last situation will be considered.

TABLE 5.7

Chi-squared test for normality applied to the fluoride data of Table 2.1

(1) Class interval	(2) z for upper limit	(3) Cumulative relative expected frequencies	(4) Relative expected frequencies	(5) Expected frequencies (E_i)	(6) Observed frequencies (O_i)	(7) $(O_i - E_i)^2/E_i$
700– 900	–2.02	0.023	0.023	1.45 ⎤	1 ⎤	
900–1100	–1.68	0.047	0.024	1.51 ⎦ 5.67	0 ⎦ 7	0.3120
1100–1300	–1.34	0.090	0.043	2.71	6	
1300–1500	–1.00	0.159	0.069	4.35 ⎤	4 ⎤	
1500–1700	–0.66	0.255	0.096	6.05 ⎦ 10.40	5 ⎦ 9	0.1885
1700–1900	–0.32	0.375	0.120	7.56	6	0.3219
1900–2100	0.01	0.504	0.129	8.13	6	0.5580
2100–2300	0.35	0.637	0.133	8.38	16	6.9289
2300–2500	0.69	0.755	0.118	7.43	7	0.0249
2500–2700	1.03	0.848	0.093	5.86 ⎤	2 ⎤	
2700–2900	1.37	0.915	0.067	4.22 ⎦ 10.08	5 ⎦ 7	0.9411
2900–3100	1.70	0.955	0.040	2.52 ⎤	2 ⎤	
3100–3300	2.04	0.977	0.022	1.39	2	
3300–3500	2.38	0.992	0.015	0.95 5.31	0 5	0.0181
3500–3700	2.72	0.996	0.004	0.25	0	
3700–3900	3.06	0.999	0.003	0.20 ⎦	1 ⎦	
						$\Sigma = X^2 = 9.2934$

5.6.1 Chi-square test

The test will be illustrated by means of an example. The Chi-square test applied to the fluoride data of Table 2.1 is given in Table 5.7. The observed frequencies for these data (column (6)), grouped into classes, are obtained from Table 2.2. To test whether these observations are normally distributed we proceed as follows:

1. The distribution mean and standard deviation are estimated from the 63 measurements yielding $\bar{x} = 2092.3$ and $s = 591.7$.
2. The upper class limits are transformed into standardized deviates (column(2)) by applying eq. (3.10).
3. From Table 3.3 the cumulative probabilities to find a value smaller than z (column (3)) are obtained. They represent the cumulative relative expected frequencies. Notice that for negative z-values the probability is: $p\,(< -z) = 1 - p\,(< z)$.
4. The relative expected frequency for each class (column(4)) is derived from column (3).
5. The expected frequencies (column (5)) are obtained by multiplying the relative expected frequencies by $n = 63$.
6. The test requires that the expected frequencies are not too small. The accepted convention is that they should at least be equal to 5. Therefore the data are regrouped as shown in Table 5.7 in order to have an expected frequency of at least 5 in each class. Of course the corresponding observed frequencies also have to be regrouped.
7. The following test statistic is calculated (column (7)):

   $X^2 = \sum(O_i - E_i)^2/E_i$

   with O_i and E_i the observed and expected frequency, respectively, for each class.
8. If the null hypothesis that the data are normally distributed holds, X^2 is approximately distributed as χ^2. Therefore X^2 is compared with tabulated χ^2-values at $k - 3$ degrees of freedom, k being the number of classes used in the calculation. In the comparison of an observed frequency distribution with a particular normal distribution (i.e. μ and σ known) there are $k - 1$ degrees of freedom. For our example the tabulated value of χ^2 with 5 degrees of freedom at the 5% significance level, obtained from Table 5.4, equals 11.07. Since X^2 (= 9.293) is smaller, the null hypothesis that the fluoride data are drawn from a normal distribution is accepted. This confirms the indication of normality already obtained from the graphical rankit method given in Fig. 3.13.

TABLE 5.8

The Kolmogorov–Smirnov test applied to the measurement of breaking points of threads introduced in Section 3.8

(1)	(2) z	(3) F_{E_i}	(4) F_{O_i}	(5) $d_i = \lvert F_{O_i} - F_{E_i} \rvert$	(6) $d_i^- = \lvert F_{O_{i-1}} - F_{E_i} \rvert$
2.104	−1.17	0.121	0.083	0.038	0.121
2.222	−0.84	0.201	0.167	0.034	0.118
2.247	−0.77	0.221	0.250	0.029	0.054
2.286	−0.66	0.255	0.333	0.078	0.005
2.327	−0.54	0.295	0.417	0.122	0.038
2.367	−0.43	0.334	0.500	0.166	0.083
2.388	−0.37	0.356	0.583	0.227	0.144
2.512	−0.02	0.492	0.667	0.175	0.091
2.707	0.53	0.702	0.750	0.048	0.035
2.751	0.65	0.742	0.833	0.091	0.008
3.158	1.79	0.963	0.917	0.046	0.130
3.172	1.83	0.966	1.000	0.034	0.049

5.6.2 Kolmogorov–Smirnov test

The χ^2-test requires the data to be presented as frequencies by grouping them into classes and is therefore not applicable for small samples. Generally the test is not used with $n < 50$. Since, the Kolmogorov–Smirnov test treats all observations separately it is suitable for small samples. The test, which is applicable only to continuous distributions, consists in determining the largest difference between two cumulative relative frequency distributions: the observed distribution, here denoted F_O, and the expected distribution, here denoted F_E. It will be illustrated by means of the example concerning the measurement of breaking points of threads, introduced in Section 3.8. The Kolmogorov–Smirnov test applied to these data, with mean $\bar{x} = 2.5201$ and $s = 0.3554$, is summarized in Table 5.8. The second column gives the standardized deviation for each observation from the mean. The cumulative relative expected frequencies, F_E, in column (3) are then obtained from these standardized deviates as in the previous section by consulting Table 3.3. Since there are twelve observations, the relative observed frequency for each observation is $1/12 = 0.0833$, from which the cumulative relative observed frequencies, F_O, of column (4) are obtained.

Both the distribution F_E and F_O are represented in Fig. 5.3. The test consists in determining the largest difference between the two curves. The expected distribution, F_E, being a continuous distribution, the differences between the two distributions are computed as shown in columns (5) and (6). The differences d_i are obtained as $\lvert F_{O_i} - F_{E_i} \rvert$ and the differences d_i^- as $\lvert F_{O_{i-1}} - F_{E_i} \rvert$. These are illustrated for the second and the seventh observation in Fig. 5.3. The reason why d_i^- has to be taken into account becomes obvious if we consider the difference between both distributions around the

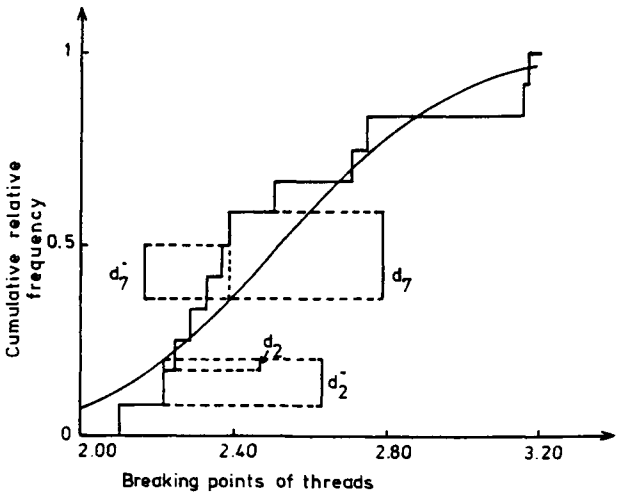


Fig. 5.3. Expected (smooth curve) and observed (stepped curve) cumulative frequency distribution of breaking points of threads (data from Table 5.8).

second observation. The largest difference around that point is obtained just below 2.222 and is calculated as d_2^- .

An inspection of all d_i and d_i^- values in Table 5.8 reveals that the maximum difference between the two distributions is $d_7 = 0.227$. This value being smaller than the critical value at $n = 12$ and $\alpha = 0.05$ of Table 5.9 (= 0.242) the data can be considered to be normally distributed.

Critical values for the Kolmogorov–Smirnov test in the case of comparison with a completely specified expected distribution (μ and σ known) can be found in Sokal and Rohlf [2].

The test as previously described can only be applied to continuous distributions and in the absence of tied values. Therefore it is for example not applicable to the Pb data of Table 3.6 in which several ties occur. Sokal and Rohlf [2] describe an approximate test in which, by grouping the data, frequencies instead of the individual observations are used. The test applied to the Pb data is summarized in Table 5.10. The data were grouped into 7 classes and the cumulative relative expected frequencies, F_{E_i}, are obtained as described in Section 5.6.1 and Table 5.7. The cumulative relative observed frequencies, F_{O_i}, are calculated by dividing the cumulative observed frequencies in the 5th column of Table 5.10 by $n = 20$.

The largest difference between both distributions F_O and F_E is 0.209. From Table 5.9 it follows that the 5% critical value for $n = 20$ is 0.192. Therefore the conclusion from Fig. 3.11a that the data are not normally distributed is confirmed. It can be verified that after elimination of the two highest results, which are suspected of being outliers, the data fit a normal distribution as is also indicated by Fig. 3.11b.

TABLE 5.9

Critical values for the Kolmogorov-Smirnov test (expected distribution derived from the data)

n	0.10	0.05	0.01
4	0.346	0.376	0.413
5	0.319	0.343	0.397
6	0.297	0.323	0.371
7	0.280	0.304	0.351
8	0.265	0.288	0.333
9	0.252	0.274	0.317
10	0.241	0.262	0.304
11	0.231	0.251	0.291
12	0.222	0.242	0.281
13	0.215	0.234	0.271
14	0.208	0.226	0.262
15	0.201	0.219	0.254
16	0.195	0.213	0.247
17	0.190	0.207	0.240
18	0.185	0.202	0.234
19	0.181	0.197	0.228
20	0.176	0.192	0.223
25	0.159	0.173	0.201
30	0.146	0.159	0.185
40	0.128	0.139	0.162
100	0.082	0.089	0.104
400	0.041	0.045	0.052
900	0.028	0.030	0.035
>30	$0.84/\sqrt{n}$	$0.90/\sqrt{n}$	$1.05/\sqrt{n}$

TABLE 5.10

Kolmogorov–Smirnov test applied to the Pb data of Table 3.6 which contain tied values

Class interval	z for upper limit	F_{E_i}	Observed frequencies	Cumulative observed frequencies	F_{O_i}	$d_i = \lvert F_{O_i} - F_{E_i} \rvert$
0.960–1.160	–0.32	0.375	7	7	0.35	0.025
1.160–1.360	0.36	0.641	10	17	0.85	0.209
1.360–1.560	1.03	0.848	1	18	0.90	0.052
1.560–1.760	1.71	0.956	0	18	0.90	0.056
1.760–1.960	2.38	0.992	1	19	0.95	0.042
1.960–2.160	3.06	0.999	0	19	0.95	0.049
2.160–2.360	3.73	0.9999	1	20	1.00	0.0001

References

1. ISO norm 3494-1976 (E) Statistical interpretation of data — Power of tests relating to means and variances.
2. R.R. Sokal and F.J. Rohlf, Biometry, The Principles and Practice of Statistics in Biological Research. W.H. Freeman, New York, 1981.
3. W. Rechenberg, Zur Ermittlung von Ausreizern. Fresenius Zeitschrift für Analytische Chemie, 331 (1988) 513–519.
4. Analytical Methods Committee, Robust Statistics — How not to reject outliers. Part 1. Basic concepts. Analyst, 114 (1989) 1693–1697.
5. P.-H. Goetsch, Ch. Junge and W. Kroenert, Evaluation of collaborative studies with special consideration of the outlier problem. J. Assoc. Off. Anal. Chem., 69 (1986) 401–403.
6. H. Beyer, Handbook of Tables for Probability and Statistics, 2nd edition. CRC Press, Boca Raton, FL, 1968.
7. ISO norm 5725-1986 (E) Precision of test methods - Determination of repeatability and reproducibility for a standard test method by inter-laboratory tests.
8. ISO 5725-2:1994(E), Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method.
9. F.E. Grubbs and G. Beck, Extension of sample size and percentage points for significance tests of outlying observations. Technometrics, 14 (1972) 847–854.
10. P.C. Kelly, Outlier detection in collaborative studies. J. Assoc. Off. Anal. Chem., 72 (1990) 58–64.
11. A.O.A.C., Guidelines for collaborative study procedure to validate characteristics of a method of analysis. J. Assoc. Off. Anal. Chem. 71 (1988) 161–172.

Additional recommended reading

J.N. Miller, Tutorial review. Outliers in experimental data and their treatment. Analyst, 118 (1993) 455–461.

Chapter 6

Analysis of Variance

6.1 One-way analysis of variance

6.1.1 Terminology — examples

In Chapter 5 hypothesis tests for the comparison of two means were discussed. It is sometimes necessary to compare more means as shown in Table 6.1a. The data of Table 6.1a were taken from a study carried out to determine whether dissolution methods have an effect on the result obtained for the determination of Fe in a multivitamin/trace element formulation [1]. Each column gives the results obtained on 6 separate samples pretreated according to a certain procedure. The first applied dry ashing, a second microwave digestion, another consisted in using a strong acid, filtration and determination of Fe in the filtrate, etc. The question is whether there is an effect of the pretreatment on the result obtained. The results of method SZC are clearly different from the rest, as shown in Fig. 6.1. However, how can we arrive at this conclusion in a statistical way? The data given are real, but not all applications are so easy to decide. One way of doing this would be to compare each column mean with each other using a *t*-test. How to do this correctly was described in Chapter 5.2 (the Bonferroni correction). We should note in passing that the computations were carried out on data which were later rounded to obtain Tables 6.1a and b. For this reason small differences are possible if the computations are carried out starting with the data in these tables.

Instead of immediately asking the question: which means are different, we can first ask a more general question: does the factor differing between the columns have an effect on the means of those columns? In other words, do all the dissolution methods yield the same result, or do one or more affect the results in a different way from the others? If the latter were the case this would have an influence on the total variance of all the data of Table 6.1a. In the case that all the methods really give the same result, that variance would be determined exclusively by the precision of the methods. Each separate result x_{ij} could then be written as follows

$$x_{ij} = \mu + e_{ij} \tag{6.1}$$

where x_{ij} is the ith result in the jth column, μ the true mean and e_{ij} the deviation of

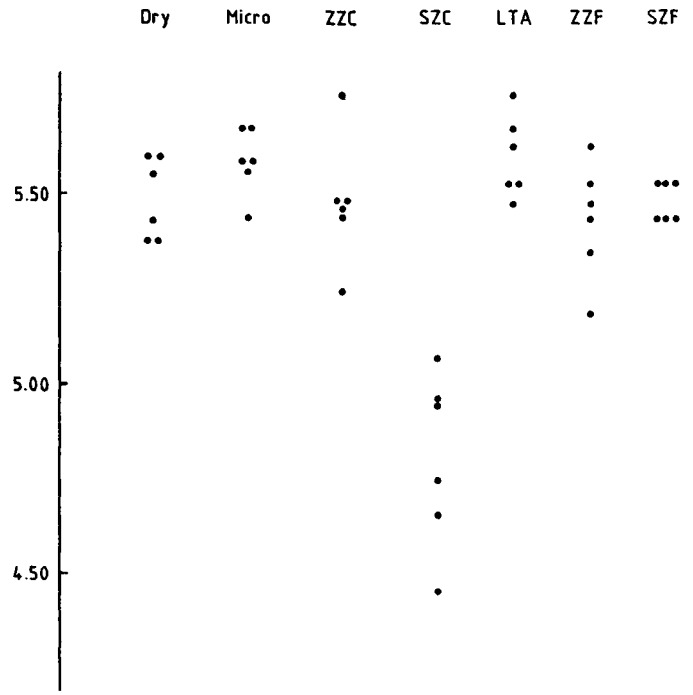


Fig. 6.1. Visual representation of the data of Tables 6.1a and b.

x_{ij} from μ. The e_{ij} are assumed to be normally distributed with mean = 0 and variance $\sigma_e^2 = \sigma^2$.

When there is an effect of the methods, we can write

$$x_{ij} = \mu + a_j + e_{ij} \qquad \sum a_j = 0 \tag{6.2}$$

where a_j is the *effect* of the jth pretreatment method relative to the overall mean. The variance of x_{ij} around μ would now no longer be determined solely by e_{ij} and thus not be equal to σ^2. The term a_j introduces additional variance in the data, so that it would be larger than σ_e^2. We may thus expect that analyzing the variance will be a way of finding out whether certain factors have an effect.

Only one single possible effect is studied in Table 6.1a: the pretreatment procedure. This table is then called a *one-way layout* and the hypothesis test applied to test whether the treatment has an effect is called a *one-way analysis of variance* or *one-way ANOVA*.

Before looking at the statistical computations, let us consider some other examples of situations that could lead to data such as those given in Table 6.1a. The same type of data could, for instance, also be the result of an interlaboratory study (see Chapter 14), carried out to investigate a specific analytical procedure. Supposing that this is carried out on the same formulation described above, and supposing

TABLE 6.1a

Concentrations of Fe (in mg/100 g) in a vitamin/mineral formulation determined by AAS using different dissolution methods [1]

	Dry	Micro	ZZC	SZC	LTA	ZZF	SZF
j	1	2	3	4	5	6	7
	5.59	5.67	5.75	4.74	5.52	5.52	5.43
	5.59	5.67	5.47	4.45	5.47	5.62	5.52
	5.37	5.55	5.43	4.65	5.66	5.47	5.43
	5.54	5.57	5.45	4.94	5.52	5.18	5.43
	5.37	5.43	5.24	4.95	5.62	5.43	5.52
	5.42	5.57	5.47	5.06	5.76	5.33	5.52
$\bar{x}_j$	5.48	5.57	5.47	4.80	5.59	5.43	5.48
s_j	0.11	0.093	0.16	0.23	0.11	0.15	0.05

TABLE 6.1b

Concentrations of Fe (in mg/100 g) in a vitamin/mineral formulation in different samples from the same lot. The data are synthetic and are the same as in Table 6.1a.

	Sample						
j	1	2	3	4	5	6	7
	5.59	5.67	5.75	4.74	5.52	5.52	5.43
	5.59	5.67	5.47	4.45	5.47	5.62	5.52
	5.37	5.55	5.43	4.65	5.66	5.47	5.43
	5.54	5.57	5.45	4.94	5.52	5.18	5.43
	5.37	5.43	5.24	4.95	5.62	5.43	5.52
	5.42	5.57	5.47	5.06	5.76	5.33	5.52
$\bar{x}_j$	5.48	5.57	5.47	4.80	5.59	5.43	5.48
s_j	0.11	0.093	0.16	0.23	0.11	0.15	0.05

supposing also that we know that the formulation is homogeneous, we would then send a sample to each of the participating laboratories and ask them to carry out 6 replicate measurements with the procedure under investigation. Column 1 would then give the results of laboratory 1, and so on.

It is not necessarily evident that the formulation is homogeneous and to test this we could then carry out an experiment that could yield exactly the same type data as in Table 6.1b. This would consist of taking samples from the lot at different locations (top, bottom, etc.) and analyzing each of them 6 times. Each column would then give the replicate determinations for one sample. If the total variance were significantly larger than the variance in one column, this could be attributed

to sample inhomogeneity. Thus the factor studied would be the effect of the sample. In the next section we will use this example to introduce the theory and some computational details.

In both cases there are 7 columns, i.e., the factor is studied at 7 *levels*.

6.1.2 Estimating sources of variance and their significance

The data shown in Table 6.1 are presented in a more general fashion in Table 6.2. Since it is our intention to investigate effects on the variance in this data table, we should first estimate the variance. Let us start by supposing that the lot investigated is homogeneous, so that the only source of variation is that due to measurement uncertainties. In other words, the precision of the analytical determination is the sole factor that determines the variance in the table. In this case, we can reason that the variance can, for example, be estimated from the first column

$$s_1^2 = \sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 / (n_1 - 1) \tag{6.3}$$

This means that the variance is determined using the replicate analysis of sample 1. This can also be done using the data of the second column (sample 2), etc. Eventually, this would yield k estimates of the variance of the data. This is not very satisfying: we would really want to obtain a single estimate and use all the data to do so. Supposing still that the batch is homogeneous and that therefore the variance is not affected by analyzing portions from different samples, the columns should have the same population mean, μ, and variance, σ^2. The column means $\bar{x}_k$ and variances s_k^2 are then separate estimates of these population parameters. To obtain one estimate of the mean, we can use the *grand mean* $\bar{x}$, i.e. the mean of all results, and pool the variances to obtain one single estimate of σ^2 (see Section 2.1.4.4).

TABLE 6.2

One-way ANOVA layout

	Sample 1	Sample 2	... Sample j ...	Sample k
	x_{11}	x_{12}	x_{1j}	x_{1k}
	x_{21}	x_{22}	x_{2j}	x_{2k}
	.			
	x_{i1}	x_{i2}	x_{ij}	x_{ik}
	.			
	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_j j}$	$x_{n_k k}$
Mean	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_j$	$\bar{x}_k$
Variance	s_1^2	s_2^2	s_j^2	s_k^2

Grand mean: $\bar{x}$.

This implies the assumption that all variances come from the same population.

$$\sigma_1^2 = \sigma_2^2 = \sigma_j^2 \ldots \sigma_k^2 = \sigma^2$$

This should be noted, because it is far from evident that this assumption is always true in practical situations (see also Section 6.2).

The pooled variance s_p^2 can then be used to estimate σ^2; because of the larger number of data used, it is a better estimate than the separate estimates s_j^2

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + \ldots + (n_k - 1)\, s_k^2}{n_1 + \ldots + n_k - k}$$

$$= \sum_{j=1}^{k} (n_j - 1)\, s_j^2 \Big/ \sum_{j=1}^{k} (n_j - 1) \qquad (6.4)$$

In this section we will consider that all n_j are equal, $n_1 = n_2 = \ldots n_k$. This is done here for ease of computation and is not a requirement of the technique. From Section 6.1.3 on, we will no longer include this restriction.

A second possibility to estimate σ^2 is to obtain it from the variance of the column means, $s_{\bar{x}}^2$, which is given by:

$$s_{\bar{x}}^2 = \sum_{j=1}^{k} (\bar{x}_j - \bar{\bar{x}})^2 / (k - 1) \qquad (6.5)$$

where $\bar{\bar{x}}$ is the grand mean, which here is also the mean of the k column means $\bar{x}_j$. As usual, one considers that $s_{\bar{x}}^2$ estimates $\sigma_{\bar{x}}^2$ and since there are n_j data in each column $\sigma_{\bar{x}}^2 = \sigma^2/n_j$ or $\sigma^2 = n_j\, \sigma_{\bar{x}}^2$. It follows that $n_j\, s_{\bar{x}}^2$ estimates σ^2. A second estimate of σ^2 is therefore given by:

$$n_j\, s_{\bar{x}}^2 = n_j \sum_{j=1}^{k} (\bar{x}_j - \bar{\bar{x}})^2 / (k - 1) \qquad (6.6)$$

The two estimates of σ^2, s_p^2 of eq. (6.4) and $n_j\, s_{\bar{x}}^2$ of eq. (6.6) are equal only if the material is homogeneous. If it is heterogeneous then the two will estimate different quantities.

The pooled variance s_p^2 is not affected by heterogeneity, since it is determined exclusively by the precision of the analytical determination. Expressed in a more general way, it describes variance of the data within each column, i.e. the *within-column variance*. Since this is not affected by heterogeneity, we must still consider that s_p^2 estimates σ^2.

The variance of the column means $s_{\bar{x}}^2$ describes the *between-column variance*. It no longer estimates only σ^2/n_j, but the additional component σ_a^2 must be added, where σ_a^2 estimates the additional variance due to heterogeneity. Therefore, $n_j s_{\bar{x}}^2$ estimates

$$\sigma^2 + n_j\, \sigma_a^2$$

These considerations allow us to write down a hypothesis and test it. Indeed, if the material is homogeneous, then s_p^2 and $n_j s_{\bar{x}}^2$ both estimate σ^2, or

H_0: $\sigma_p^2 = n_j \sigma_{\bar{x}}^2$ or H_0: $\sigma_a^2 = 0$

If the material is not homogeneous then $n_j s_{\bar{x}}^2$ estimates $\sigma^2 + n_j \sigma_a^2$ while s_p^2 estimates σ^2. In other words, $n_j s_{\bar{x}}^2$ estimates a larger variance than s_p^2 and

H_1: $\sigma_p^2 < n_j \sigma_{\bar{x}}^2$ or H_1: $\sigma_a^2 > 0$

Variances can be compared by using an F-test (see Section 5.4) and, in view of the way H_1 is formulated, this F-test must be one-sided.

6.1.3 Breaking up total variance in its components

The way ANOVA was explained above shows some of the basic assumptions of ANOVA, its philosophy, and the way the eventual hypothesis test (one-sided F-test) is carried out. The actual computations can be understood and carried out more easily by considering ANOVA as a splitting-up of the total variance in its components. The total variance is given by

$$s_T^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 / (n - 1) \tag{6.7}$$

where $n = \sum_{j=1}^{k} n_j$

In words, the total variance is the sum of the squared differences between each of the data x_{ij} and the grand mean $\bar{x}$, divided by $n - 1$ degrees of freedom where n is the total number of data in the table. For reasons of computational convenience, let us first work with the sums of squares, SS

$$SS_T = \sum_{j} \sum_{i} (x_{ij} - \bar{x})^2 \tag{6.8}$$

and introduce the degrees of freedom at a later stage of the computations. SS_T is the sum of squared differences of each individual observation from the grand mean. In some texts SS_T is the sum of squares of the data and the SS_T, as used here, is then called the corrected sum of squares (where 'corrected' denotes corrected for the mean) and represented as SS_{corr}. We will not follow this practice.

Since

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

it follows that

$$(x_{ij} - \bar{x})^2 = (x_{ij} - \bar{x}_j)^2 + (\bar{x}_j - \bar{x})^2 + 2(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x}) \tag{6.9}$$

To obtain SS_T in eq. (6.8), we sum over rows (i) and columns (j). The last term of eq. (6.9) becomes zero, since differences from a mean cancel out when they are summed. The result is therefore

$$SS_T = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2 + \sum_j n_j(\bar{x}_j - \bar{x})^2 \tag{6.10}$$

or

$$SS_T = SS_R + SS_A \tag{6.11}$$

where

$$SS_R = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2 \tag{6.12}$$

and

$$SS_A = \sum_j n_j (\bar{x}_j - \bar{x})^2 \tag{6.13}$$

SS_R is called the *residual sum of squares.* The term "residual" is explained later in this section. SS_A is the sum of squares due to the effect of the factor studied (this factor, called A, is here the composition heterogeneity among samples). It is also sometimes called $SS_{treatment}$ in general (ANOVA was frequently used first in agronomy, where the effects were agricultural treatments) or it refers in some way to the reason of the effect. Here we could write $SS_{heterogeneity}$. Finally, SS_{within} (*within-column sum of squares)* can be written for SS_R, because it has to do with variance within columns and $SS_{between}$ (*between-column sum of squares*) for SS_A because it is linked to variance between columns in the table.

To obtain estimates of variance from the sums of squares we divide by the number of degrees of freedom. In general this is written as

$$MS = SS/df$$

where MS or *mean square* is a variance estimate and df is the number of degrees of freedom. Applied to SS_R and SS_A, this yields

$$MS_A = SS_A/(k - 1) \tag{6.14}$$

It estimates

$$\sigma^2 + \sigma_a^2 \frac{n - (\sum n_j^2 / n)}{k - 1}$$

For equal n_j, this can be shown to be equal to

$$\sigma^2 + n_j \sigma_a^2$$

Furthermore:

$$MS_R = SS_R/(n-k) \tag{6.15}$$

This estimates σ^2. The number of degrees of freedom $n - k$ can easily be understood by referring to Section 6.1.2. MS_R is equal to s_p^2 as given by eq. (6.4) and we can verify that the denominator is indeed equal to $kn_j - k = n - k$. In practice, we can derive the number of degrees of freedom also by reasoning that the number of degrees of freedom for SS_T is $(n - 1)$, that $(k - 1)$ of these are used up by SS_A and that the rest $(n - 1) - (k - 1) = n - k$ is available for SS_R. This helps us to understand the reason for the term "residual". The residual sum of squares is the total sum of squares minus the sum of squares due to a specific factor ($SS_R = SS_T - SS_A$) and the residual degrees of freedom are those that are not used up by this specific factor: $df_R = df_T - df_A$.

Since MS_R and MS_A respectively estimate the σ^2 and $\sigma^2 + n_j \sigma_a^2$ of the preceding section, this also means that we can carry out the hypothesis test as described in that section, i.e., by an F test.

$$F = \frac{MS_A}{MS_R} = \frac{SS_A/(k-1)}{SS_R/(n-k)} \tag{6.16}$$

and this F ratio must then be compared with the tabulated F for $k - 1$ and $n - k$ degrees of freedom (Table 5.3). It should be remembered (see Section 6.1.2) that this is a one-sided test.

6.1.4 Random and fixed effect models

When explaining ANOVA in Sections 6.1.2 and 6.1.3, we have applied a so-called *random effect model*. There is a second type of model called the *fixed effect model*. These two different models rarely have an effect on the set-up of the experiment or on the ANOVA table (see next section) and the first hypothesis test to be carried out, namely the F-test. The purpose, however, of the ANOVA is different as are some of the operations or tests carried out after the F-test. This requires some additional explanation.

In eq. (6.2) an additive model — also called linear model — was defined, in which each single result can be divided in several components. One of these was described as the effect of the factor (in eq. (6.2) this was the pretreatment method). A more precise definition of the additive model is now required. There are, in fact, two definitions.

The first possibility is to consider the effect of the factor as a fixed deviation of the mean of group j from the grand mean. This would be the case for the example given in Table 6.1a in which the effect of different pretreatments is studied. Each result for pretreatment method j would then consist of $\mu + a_j$, the mean and the effect of the pretreatment method on the one hand and the randomly distributed error or residual e_{ij} on the other. This is called *Model I ANOVA* or a *fixed effect*

model. Strictly speaking, this has an effect on the mathematics because MS_A now estimates

$$\sigma^2 + \frac{\sum_j n_j a_j^2}{k-1}$$

and the hypothesis to be tested can be stated as:

H_0: $a_1 = a_2 = \ldots a_k = 0$

H_1: $a_j \neq 0$ for a least one j

This has no computational consequences for the ANOVA as such as the test is still performed as an F-test on the ratio MS_A/MS_R.

In the case of a fixed effect model, it can be concluded that at least one column mean is different from the others (in the example, at least one pretreatment method is different from the others). We might then be interested in knowing which means are significantly different from the others. How this is done is described in Section 6.3.

The model to be applied to the homogeneity problem of Table 6.1b is called a *Model II ANOVA* or a *random effect model*. We are not interested in a specific effect due to a certain column, but a general effect on all columns and that effect is considered to be normally distributed. To distinguish between model I and model II, we sometimes use different symbols for the effects. For instance, we could use the lower case letter a for model I and the capital letter A for model II. This yields

$$x_{ij} = \mu + a_j + e_{ij} \qquad (6.17\text{I})$$

$$x_{ij} = \mu + A_j + e_{ij} \qquad (6.17\text{II})$$

where a_j is the fixed effect of model I and A_j is the normally distributed variable with mean 0 and variance σ_A^2 of model II. As already explained in the preceding section, MS_A estimates for model II

$$\sigma^2 + \sigma_A^2 \frac{n - \sum n_j^2/n}{k-1}$$

or, for equal n_j, $\sigma^2 + n_j \sigma_A^2$

Since the effect on the column means is random there is no sense in trying to determine which column mean is significantly different from another. We should consider that inhomogeneity of the samples adds variance to the variance due to the determination and, in this case, we might like to determine how large the added variance component is. This will be described under Section 6.4.

The difference between the models is not always evident. In the example of the intercomparison of laboratories, we might focus on differences between the specific laboratories taking part (proficiency testing), which would then be a fixed

effect model. On the other hand, we could consider the laboratories as representative for a population of sufficiently proficient laboratories. The within-column variance describes repeatability, the overall variance the reproducibility and the between-column variance the added variance component due to between-laboratory variance (see also Chapters 13 and 14). This model would be a random effects model.

This may seem complex to the first-time user. Fortunately the distinction is often important only from a philosophical point of view and may be disregarded at a first reading. As stated earlier, the ANOVA table and the hypothesis test with the F-tables is exactly the same in both cases. However, for a deeper understanding, the philosophy behind the statistics is important and the difference between the two models should therefore be included in a more thorough study.

6.1.5 The ANOVA table

The computational scheme of Section 6.1.3 can be summarized in an ANOVA table. These tables, whether they concern a one-way experiment as defined in the previous section or multi-way layouts (see e.g. Section 6.5) always have similar formats. They consist of up to five columns: the first column gives the source of the variation, the second and third the degrees of freedom and sums of squares (not necessarily in that order), the fourth the mean square and the fifth the F values. Under the table is often written the critical F-values that have to be compared with the experimental values in the fifth column and the conclusion (the effect is significant or not at a certain level). Sometimes p-values are given in a sixth column. This then yields the general layout of an ANOVA Table (Table 6.3).

For the data of Table 6.1 this yields Table 6.4. The between-group variance is significant at the level $\alpha < 0.001$ since $22.97 > 4.92$. If, as in Table 6.1b, we considered these data to be the data of a homogeneity experiment, i.e. a Model II ANOVA, then our first conclusion would be that the material is not homogeneous. We might then continue with the techniques described in Section 6.4 and try to determine how much of the variance in the data is due to this effect. If we considered the data to be those of Table 6.1a, i.e. a comparison of pretreatment

TABLE 6.3

One-way ANOVA table

Source	Degrees of freedom	Sum of squares	Mean square	F
Between columns (A)	$k-1$	SS_A	$SS_A/(k-1)$	MS_A/MS_R
Within columns (residual)	$n-k$	SS_R	$SS_R/(n-k)$	
Total	$n-1$	SS_T		

$F_{0.05;k-1,n-k}$ = ..., conclusion about significance of A: ...

TABLE 6.4
ANOVA of the data of Table 6.1

Source	Degrees of freedom	Sum of squares	Mean square	F
Between columns	6	2.6765	0.4461	22.9709
Within columns	35	0.6797	0.0194	
Total	41	3.3562		

$F_{0.05;(6,35)} = 2.38$, $F_{0.001;(6,35)} = 4.92$.

methods, we would now decide that at least one method gives results different from the others and turn to the methods described under Section 6.3 to obtain more detailed conclusions.

6.2 Assumptions

Because within-column variances are pooled to estimate MS_R (Section 6.1.2), we assume that these variances are equal. In other words, we assume *homogeneity of variance* or *homoscedasticity*. When this assumption is violated and the variances are not equal, we conclude that there is *heteroscedasticity*. Wrongly assuming homoscedasticity can lead to serious errors and therefore tests that allow us to verify this assumption are required.

In some contexts (e.g., method validation, see Chapter 13), the emphasis is on deciding whether the variance in one of the columns is higher than in the other columns, rather than on investigating that all variances are equal. In other words, we suspect the column with highest variance to have a significantly higher variance than all the others. This is only another way of saying that there is heteroscedasticity, and therefore tests for heteroscedasticity can also be applied for this type of application.

In our view, ANOVA should always be preceded by visual inspection of the data before any test is carried out. Figure 6.1 provides such an analysis. Inspection of the plot immediately indicates that it is probable that SZC proves to be different from the other pretreatment methods. A particular powerful aid is the box plot (Chapter 12). This gives immediate visual indication of whether a violation of the assumptions is to be feared. At the same time it will permit us to assess the occurrence of differences between means (Section 6.3), violations of the normality assumption within columns and, in the case of two-way ANOVA, the occurrence of interaction (Section 6.6).

Rapid tests can be carried out with the use of ranges [2]. One test is based on the comparison of the highest within-column range, w_{max}, with the sum of all ranges.

Another is based on the comparison of the highest range to the lowest [3], w_{max}/w_{min}. The idea of comparing largest to smallest dispersion is also used in *Hartley's test* [4] which compares the highest variance within columns to the lowest, s^2_{max}/s^2_{min}.

Cochran's criterion is based on comparing s^2_{max} with all the other variances. Because it is recommended by ISO [5], we will describe it here in somewhat more detail. It is given by:

$$C = \frac{s^2_{max}}{\sum_j s_j^2} \tag{6.18}$$

C is then compared to critical values (see Table 6.5). It should be noted that this criterion requires that all columns contain the same number of results (n_j). If the n_j do not differ too much, the test can still be carried out. ISO [5] then recommends the most frequent n_j value is used. For $n_j = 2$, we can replace s in eq. (6.18) by w, the range, so that

$$C = \frac{w^2_{max}}{\sum_j w_j^2}$$

Applied to the data of Table 6.1, Cochran's test would yield the following results:

$s_1^2 = 0.011$ $\quad s_2^2 = 0.008$ $\quad s_3^2 = 0.027$ $\quad s_4^2 = 0.052$
$s_5^2 = 0.012$ $\quad s_6^2 = 0.024$ $\quad s_7^2 = 0.002$

Thus $s^2_{max} = 0.052$ and $\sum s_j^2 = 0.135$ so that $C = 0.052/0.135 = 0.382$.

The critical value for C for $n_j = 6$ and $k = 7$ is 0.397. It follows that the data are considered to be homoscedastic.

A test which is often found in books on applied statistics is *Bartlett's test*. It has been shown [6] that this test is very sensitive to departures from normality within columns, so that finding a significant result often indicates non-normality rather than heteroscedasticity. This test will therefore not be discussed here. Other possibilities are the *log-ANOVA* or *Scheffé-Box* test [6].

When differences in variance have been found, several possibilities exist. In some contexts (e.g. method-performance testing) we can decide that the data from laboratories with too high a variance should be eliminated (see further Chapter 14). In many cases, we cannot reject data and must resort to methods that allow us to restore homoscedasticity. There are two ways of doing this: one is by transforming the variables (appropriate transformations are discussed in Chapter 8 on regression where the same problem occurs); the other is to apply weights (weighted ANOVA).

As seen in Section 6.1.1, a second assumption is that the e_{ij} within a column are

TABLE 6.5

Critical values for Cochran's C at the 5% level of significance

k	$n_j = 2$	$n_j = 3$	$n_j = 4$	$n_j = 5$	$n_j = 6$
2	–	0.975	0.939	0.906	0.877
3	0.967	0.871	0.798	0.746	0.707
4	0.906	0.768	0.684	0.629	0.590
5	0.841	0.684	0.598	0.544	0.506
6	0.781	0.616	0.532	0.480	0.445
7	0.727	0.561	0.480	0.431	0.397
8	0.680	0.516	0.438	0.391	0.360
9	0.638	0.478	0.403	0.358	0.329
10	0.602	0.445	0.373	0.331	0.303
11	0.570	0.417	0.348	0.308	0.281
12	0.541	0.392	0.326	0.288	0.262
13	0.515	0.371	0.307	0.271	0.243
14	0.492	0.352	0.291	0.255	0.223
15	0.471	0.335	0.276	0.242	0.220
16	0.452	0.319	0.262	0.230	0.208
17	0.434	0.305	0.250	0.219	0.198
18	0.418	0.293	0.240	0.209	0.189
19	0.403	0.281	0.230	0.200	0.181
20	0.389	0.270	0.220	0.192	0.174
21	0.377	0.261	0.212	0.185	0.167
22	0.365	0.252	0.204	0.178	0.160
23	0.354	0.243	0.197	0.172	0.155
24	0.343	0.235	0.191	0.166	0.149
25	0.334	0.228	0.185	0.160	0.144
26	0.325	0.221	0.179	0.155	0.140
27	0.316	0.215	0.173	0.150	0.135
28	0.308	0.209	0.168	0.146	0.131
29	0.300	0.203	0.164	0.142	0.127
30	0.293	0.198	0.159	0.138	0.124

normally distributed, which means that the data within one column should also be normally distributed. How to test normality is described in Chapter 3. In many cases, knowledge about the process investigated will give us a good reason to accept that the underlying distribution is normal and we will be concerned more about the occurrence of outliers, which can also be considered as a deviation from normality. Tests for outliers are discussed in Chapter 5.

To avoid problems with the assumptions of normality and homoscedasticity, it is possible to carry out robust ANOVA. In Chapter 12 the best known such method for ANOVA — the ANOVA by ranks — will be described (Section 12.1.4.4) as well as a randomization method (Section 12.4).

A1	A2	A3	A4	A5	A6
B1	B2	B3	B4	B5	B6

Fig. 6.2. Non-randomized layout of plots on a field for a fertilizer test.

There is another assumption about the e_{ij}, namely that they are independent. This is written in all statistics books, but often not explained. Let us therefore give an example. 'Independent' means that the individual e_{ij} are randomly distributed and not influenced by an external factor. Suppose for instance that we want to compare two fertilizers, A and B, by treating small plots of a field with the fertilizers in a layout shown by Fig. 6.2. It may be that, due to some factor such as irrigation or exposure to the sun, the yield of the crop in series A and/or B changes in a specific order (for instance in the north–south direction A > B or in the east–west direction 1 > 2 > 3 > 4 > 5 > 6). The errors for A and for 1,2 and 3 will then probably be positive, and those for the others negative. They are not independent (they depend in this case on location). Usually we will not test for independence, but rather make sure that independence is achieved by proper randomization (see further).

Conclusions can be biased by *uncontrolled factors,* i.e. factors that have not been taken into account. Suppose we need to compare several industrial extraction methods. At first sight, a logical experimental set-up could be the following. Extraction procedure A is first carried out six times on day 1, then the same for procedure B on day 2, etc. However, suppose that temperature is not controlled but that, unknown to the operator, it does influence the extraction yield. Then, if the temperature is different on the successive days, the effect of the extraction procedure will be confused or *confounded* with that of the (uncontrolled) factor temperature (or days). How to avoid this depends on practical considerations, but in this instance and supposing there are only three extraction methods, we could carry out 2 extractions with A, 2 with B and 2 with C on day 1, repeat this on days 2 and 3 and analyze all samples in a random order afterwards. Indeed, it is always possible that some drift would occur during the analysis. If we were analysing first all A, then all B and finally all C samples, it could be that a difference in the results could be introduced by the order of analysis. For this reason, the order of the 18 determinations should be randomized.

In doing this we have applied the two main principles to avoid bias due to uncontrolled variables. We have applied *planned grouping* or *blocking out* to the extraction step and *randomization* to the measurement step. These are very important principles for experimental design and will be discussed again in Section 6.9 and Chapter 22.

An assumption, which is important in the random effect models, is that the effect of the factor (if any) should indeed be random and normally distributed. This can

TABLE 6.6

ANOVA of the data of Table 6.1 after deleting the SZC procedure or sample 4

Source	Degrees of freedom	Sum of squares	Mean square	F
Between groups	5	0.1284	0.0257	1.8285
Within groups	30	0.4214	0.0140	
Total	35	0.5498		

$F_{0.05;5,30} = 2.53$.

be checked in the way described in Chapter 3. Applied to the sampling problem and the data of Table 6.1b, this means we could check whether the $\bar{x}_k$ are normally distributed or we could apply an outlier test. The data for sample 4 are then found to be outliers. How to detect outliers is discussed in Chapter 5. If the experimenter feels justified in eliminating the outlier, he or she can then continue work with the other 6 samples. The result is described in Table 6.6.

As in other hypothesis tests, the sample size determines whether it will be possible to demonstrate a certain difference (if it exists), which is considered important by the investigator. How to compute the sample size for a given one-way ANOVA problem is described for instance in Ref. [6].

6.3 Fixed effect models: testing differences between means of columns

When the null hypothesis has been rejected, in the fixed effect model it is considered that at least one column has a mean value different from the others. We would then like to know which one(s). For instance, for the data of Table 6.1a we would wonder which pretreatment methods give different (higher, lower) results compared with the other methods.

As already stated in Section 6.2, ANOVA should always be accompanied by a visual analysis of the data such as that shown in Fig. 6.1. The box plot (see Chapter 12) can also be recommended for such an analysis. It immediately singles out those columns for which it is most likely that differences exist and it may well be that further statistical analysis is no longer needed or can be considered as necessary only to confirm what one has seen.

The first obvious way is to use the appropriate t-test to compare all means with each other. It was explained in Chapter 5 that this requires an adjustment of the probability levels (the Bonferroni procedure). We might try to avoid this by selecting the groups with the highest and lowest means and carrying out t-tests to compare them. However, although this involves explicitly only one or a few t-tests, it really means one has surveyed all means and compared them implicitly to each

other. Therefore, the same Bonferroni correction should be applied, whether one actually carries out the tests only for a few pairs or for all of them.

Many different methods have been described in the statistical literature that were specifically designed for the comparison of several means. Appealing because of its simplicity is the *Least Significant Difference (LSD)* method. One tries to define a difference between two means that, when exceeded, indicates that these two means are significantly different. Any pair of means for which $\bar{x}_j - \bar{x}_h > \text{LSD}$ is then considered different.

In Chapter 5 it was seen that for the comparison of two means with an independent t-test and for small sample sizes, we apply

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s^2[(1/n_1) + (1/n_2)]}}$$

The denominator of this equation contains the pooled variance due to measurement error. In ANOVA, this is estimated by the MS_R and we can therefore write

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{MS_R\,[(1/n_1) + (1/n_2)]}}$$

For equal sample size n_j this simplifies to

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{MS_R\,(2/n_j)}}$$

We can then test each $\bar{x}_1 - \bar{x}_2$ against

$$\text{LSD} = t\sqrt{MS_R(2/n_j)} \tag{6.19}$$

t is obtained from a t table at the appropriate level of confidence α (usually 0.05) and degrees of freedom (that for MS_R).

Consider the example of Table 6.1a for which the ANOVA was carried out in Table 6.4. Since the number of degrees of freedom for MS_R is 35, we can consider that $t = 1.96$ at $\alpha = 0.05$. LSD is then given by

$$\text{LSD} = 1.96\sqrt{0.0194(2/6)} = 0.158$$

The only differences between two means larger than 0.158 are those between SZC and all the others. The difference between LTA and ZZF is also marginally larger (0.160), but because we know that the LSD method tends to select too many significant differences, this is not considered enough. This is due to the fact that we implicitly compare all means with each other without correction of α, i.e. we apply a non-simultaneous approach. As explained earlier in this section, this would really require an adjustment of the probabilities.

More rigorous, but also more cumbersome tests, have been described in the literature, e.g., the *Scheffé method*, the *Student–Newman–Keuls method*, the *Tukey–Kramer method*, the *T-method*. As the LSD, they are all based in some way on defining a minimum distance, which consists of a product of a critical value, based on a statistical distribution (e.g. a t or F distribution) and a standard deviation, derived from the MS_R. Some of the tests are valid only for equal n_j and for some of them a simultaneous approach is used, while for others it is a non-simultaneous one. Discussions can be found in several books and, more briefly, in a review article by Stahl [7].

A special purpose test is the *Dunnett test*, where one compares one mean (usually the first) with all the other means. It is applied, for instance, in the following situation. A control value of a certain variable has been determined at the beginning of the experiment. Some treatment is applied and at different points in time the values of the variable are measured again. The question is then: from which moment on is there a significant difference?

Iterative procedures are also possible. We can eliminate the column which is considered to be the most probably different and carry out the ANOVA on the remaining data. As an example, SZC is eliminated from Table 6.1. The ANOVA on the remaining data is given in Table 6.6. No significant effect is obtained. Therefore, we conclude that the significant difference noted earlier was indeed due to SZC and only to SZC.

6.4 Random effect models: variance components

As pointed out in Section 6.1.4, when the effect of a factor is random it makes no sense to try and determine which column mean is responsible for the significance of the effect as was done in the preceding section. However, the effect does add variance and we can determine how much. This is useful in, e.g., the study of the precision of analytical methods, since it is possible to determine how much of the total variance is due to each step. In the same way, it can be used in SPC to determine what could be the effect of better control of a certain step on the total variance.

Let us consider again the example of Section 6.1.2. We know how to determine the variance due to the measurement, σ^2: it is estimated by MS_R. For equal n_j, MS_A estimates $\sigma^2 + n_j \sigma_A^2$, where σ_A^2 is the variance due to the sample heterogeneity. We can estimate the variance due to heterogeneity, s_A^2, as

$$s_A^2 = (MS_A - MS_R) / n_j \tag{6.20}$$

Since we concluded in Section 6.3 that sample 4 is an outlier, it was eliminated and the ANOVA of Table 6.6 on the remaining samples shows that the effect of

the samples is not significant. Since the effect of the samples is not significant, we would decide not to make the calculations of eq. (6.20). In some cases we have no interest in the test, but merely want the best estimate possible of the effect. This is provided by s_A. Applied to the example of Table 6.6, this means that $s_A^2 = (0.0257 - 0.0140)/6 = 0.00195$ and $s_A = 0.044$ while $s_R^2 = 0.0140$ and $s_R = 0.1183$. The standard deviation due to composition heterogeneity explains only a small part of the total standard deviation.

It should be added here that the determination of variance components is often carried out in more complex cases by nested ANOVA (see Section 6.11).

6.5 Two-way and multi-way ANOVA

Let us return to the example of Table 6.1. Because matrix effects may occur, we should ask whether a matrix modifier should be added before the determination takes place. To investigate this we can set up an experiment with two factors. The first factor is the pretreatment and, to keep the example simple, 3 instead of 6 types of pretreatment are considered. In other words, the first factor is studied at three levels. The second factor is the matrix modifier. This is studied at two levels, namely with a certain amount of modifier and without modifier. This yields Table 6.7.

TABLE 6.7

Effect of pretreatment and matrix modification on the determination of Fe by AAS (hypothetical data derived from Table 6.1)

Matrix modification	Pretreatment		
	Dry	Micro	ZZC
Without	5.59	5.67	5.75
	5.59	5.67	5.47
	5.37	5.55	5.43
	5.54	5.57	5.45
	5.37	5.43	5.24
	5.42	5.57	5.47
With	5.90	5.90	5.81
	5.75	6.01	5.90
	6.07	5.85	5.81
	5.90	5.54	5.81
	6.01	5.81	5.90
	6.06	5.70	5.90

TABLE 6.8

Two-way ANOVA design (without replication)

	Factor B						
Factor A	1	2	...	j	...	k	Means factor A
1	x_{11}	x_{12}		x_{1j}		x_{1k}	$\bar{x}_{1.}$
2	x_{21}	x_{22}		x_{2j}		x_{2k}	$\bar{x}_{2.}$
.							
.							
h	x_{h1}	x_{h2}		x_{hj}		x_{hk}	$\bar{x}_{h.}$
.							
.							
l	x_{l1}	x_{l2}		x_{lj}		x_{lk}	$\bar{x}_{l.}$
Means factor B	$\bar{x}_{.1}$	$\bar{x}_{.2}$		$\bar{x}_{.j}$		$\bar{x}_{.k}$	Grand mean: $\bar{x}$

Tables such as Table 6.7 are called *two-way* tables or designs and the data analysis by ANOVA is a *two-way ANOVA* because the data are subject to a double classification. In this example the data are classified once according to the pretreatment of the data and once according to the matrix modification. In general, this leads to a design such as that in Table 6.8. In the specific example of Table 6.7 factor A would be the pretreatment and factor B the matrix modification. To simplify the description of the computation, we will first discuss the case where there is no replication. By 'replication' we mean that more than one result is obtained in each *cell* of the ANOVA table. For instance, in Table 6.7 there are 6 replicates in each cell.

The grand mean, the mean of all the data of Table 6.8, is given by:

$$\bar{x} = \sum_h \sum_j x_{hj} / lk \qquad h = 1 \text{ to } l, j = 1 \text{ to } k$$

There are l levels of factor A and the mean at each of those levels is given by $\bar{x}_{1.}, \bar{x}_{2.}, \ldots \bar{x}_{h.}, \bar{x}_{l.}$ and

$$\bar{x}_{h.} = \sum_j x_{hj} / k$$

Similarly, there are k levels of factor B and the mean at each level is given by

$$\bar{x}_{.j} = \sum_h x_{hj} / l$$

The total sum of squares, SS_T, is obtained in a similar way as described earlier in eq. (6.8) and it is broken down in a similar way as in Section 6.1.3 into components due to the different factors and the residual.

$$SS_T = \sum_h \sum_j (x_{hj} - \bar{x})^2 = SS_A + SS_B + SS_R \tag{6.21}$$

The sum of squares due to factor A, SS_A, is given below, together with the number of degrees of freedom df_A and the mean square $MS_A = SS_A/df_A$.

$$SS_A = k \sum_h (\bar{x}_{h.} - \bar{x})^2 \tag{6.22}$$

$$df_A = l - 1$$

$$MS_A = SS_A/(l - 1) \tag{6.23}$$

The sum of squares due to factor B, SS_B, is given by:

$$SS_B = l \sum_j (\bar{x}_j - \bar{x})^2 \tag{6.24}$$

$$df_B = k - 1$$

$$MS_B = SS_B/(k - 1) \tag{6.25}$$

The equations for SS_R and MS_R are less easy to understand and, in practice, they can always be determined as

$$SS_R = SS_T - SS_A - SS_B \tag{6.26}$$

and

$$df_R = df_T - (k - 1) - (l - 1) \tag{6.27}$$

$$MS_R = SS_R/df_R \tag{6.28}$$

In words, the residual sum of squares is the total sum of squares minus the sum of squares for each of the factors and the number of degrees of freedom for the residual term is equal to the total number of degrees of freedom (i.e. the total number of data, kl, minus 1) minus the number of degrees of freedom used for the other sources of variance, factor A and factor B.

It can be shown that this is equal to:

$$SS_R = \sum_h \sum_j (x - \bar{x}_{h.} - \bar{x}_j + \bar{x})^2$$

$$df_R = (k - 1)(l - 1) \tag{6.29}$$

and

$$MS_R = SS_R/(k - 1)(l - 1) \tag{6.30}$$

It is useful at this stage to note that by breaking down the sum of squares as described above, one assumes the linear model:

$$x_{hj} = \mu + a_h + b_j + e_{hj} \tag{6.31}$$

where x_{hj} is the value in cell hj, a_h is the effect of the hth level of factor a and b_j the effect of the jth level of factor b and e_{hj} is the random error of the observation in cell hj. As for the one-way ANOVA, one can make a distinction between fixed effects and random effects. Our example is a fixed effects model. For two-way ANOVA it is possible that one factor, say a, is fixed and the other random. Following the convention of eqs. (6.17) we could then write

$$x_{hj} = \mu + a_h + B_j + e_{hj}$$

This is then called a *mixed effect model.*

Table 6.8 is constituted of a grid of data. Each of these data forms a cell. Until now we have assumed that each cell contains only one numerical result. It is however possible that there are more. In fact, in our example there are 3×2 cells each containing 6 replicates. We will see later that replicates are not required when one computes a two-way ANOVA with only the main effects but that they are required when one wants to estimate also interaction effects (see Section 6.6). When replicates are present, we should then write x_{hji} for the ith replicate in cell hj ($i = 1$ to n_j). It is, in fact, not necessary that all cells contain the same amount of replicates; large differences, however, should be avoided. The computations are summarized in a two-way ANOVA table (Table 6.9) similar in construction to the one-way Table 6.3. For the example of Table 6.7, this yields Table 6.10. We conclude that the pretreatment has no significant effect but that the effect of the modifier is very clear.

It is possible to investigate more than two factors by ANOVA. In our AAS example we could ask if changing the atomization temperature from 2300 to 2400°C has an effect. This could then lead us to carry out experiments at all combinations of the three types of pretreatment, the two types of modifier and the two levels of temperature. The table would be a three-way table and the ANOVA would be a three-way ANOVA. ANOVA applications for more than two factors are often called *multi-way ANOVA.*

TABLE 6.9

Two-way ANOVA table

Source	Degrees of freedom	Sum of squares	Mean square	F
Main effects	$df_A + df_B$	$SS_A + SS_B$		
Factor 1 (A)	$l - 1$	SS_A	$SS_A/(l-1)$	MS_A/MS_R
Factor 2 (B)	$k - 1$	SS_B	$SS_B/(k-1)$	MS_B/MS_R
Residual	$t - [(k-1) + (l-1)] = r$	SS_R	SS_R/r	
Total	$n_j kl - 1 = t$	SS_T		

TABLE 6.10

Two-way ANOVA table of the data of Table 6.7

Source	Degrees of freedom	Sum of squares	Mean square	F
Main effects	3	1.183		
Pretreatment	2	0.016	0.008	0.49
Modifier	1	1.166	1.166	68.85
Residual	32	0.542	0.017	
Total	35	1.725	0.049	

Significance: pretreatment, $p = 0.61$ (NS); modifier, $p < 0.001$.

6.6 Interaction

In many experimental systems, the effect of one factor depends on the level of the other. This is called *interaction*. In the example given to explain the two-way ANOVA the effect of the modifier might depend on the medium in which the Fe is dissolved and therefore on the other factor we studied, namely the dissolution procedure. One would conclude that factor A (pretreatment) interacts with factor B (the modifier).

The interaction influences the variation found in the data table. The way ANOVA treats this is to consider the interaction as an additional source of variance, next to the main effects of factor A and factor B. More precisely, in the linear model of eq. (6.31), one adds an additional term and takes into account replicates

$$x_{hji} = \mu + a_h + b_j + (ab)_{hj} + e_{hji} \tag{6.32}$$

The cross term $(ab)_{hj}$ describes the interaction. The number of degrees of freedom for the interaction, df_{AB}, is equal to the product of the degrees of freedom for the interacting factors. For the two-way lay out of Table 6.8 one would then obtain:

$$df_{AB} = df_A \cdot df_B = (k-1)(l-1) \tag{6.33}$$

Equation (6.33) yields exactly the same number of degrees of freedom for the residual in eq. (6.29) which was obtained without replication. As df_R is always equal to df_T minus the degrees of freedom used up for the other sources of variance, df_R would then be given by:

$$df_R = df_T - (k-1) - (l-1) - (k-1)(l-1) = 0 \tag{6.34}$$

There are no degrees of freedom for the residual left when there is no replication. Therefore, to test all effects, including the interaction effect, it is necessary to

TABLE 6.11

Two-way ANOVA table with interaction for the data of Table 6.7

Source	Degrees of freedom	Sum of squares	Mean square	F
Main effects	3	1.183		
Pretreatment	2	0.016	0.008	0.55
Modifier	1	1.166	1.166	77.74
Interaction	2	0.092	0.046	3.07
Residual	30	0.450	0.015	
Total	35	1.725		

Significance: pretreatment, $p = 0.474$ (NS); modifier, $p = 0.000$; interaction, $p = 0.076$ (NS).

replicate measurements. As 6 replicates were obtained in Table 6.7, we can estimate and test the interaction effect between pretreatment and modifier.

We will not go further into the details of the computation. The principles, however, are the same. One computes a total sum of squares and breaks it up in sums of squares for each source of variance. The same SS-terms as in Tables 6.9 and 6.10 are then included with, additionally, an $SS_{interaction}$. In the ANOVA table with interaction (Table 6.11), one writes down first the effects due specifically and solely to a certain factor (the *main effects*). This is then followed with the interaction term. The SS_R is, as always, equal to SS_T minus all SS-terms due to the effects, i.e. the main and the interaction effects.

To obtain the mean square, we again divide SS by df. For the interaction:

$$MS_{AB} = SS_{AB}/df_{AB}$$

By applying this to the data of Table 6.7, Table 6.11 is obtained. There is a very significant effect of the modifier, the pretreatment is not significant and neither is the interaction. One would be tempted to add "of course", because, since there is only one significant factor, one could reason that there can be no interaction between two factors. In fact, it is possible that an interaction exists and that the pure effects on their own are not significant. However, this is rare and such a result should be viewed with suspicion. A possible artefact when the significant factor is very significant is that some of the variance due to it may be partitioned into the interaction which may then be computed as significant.

It should be stressed that in two-way ANOVA the same assumptions are made as in one-way ANOVA (normality and homoscedasticity of all cells, etc.). Because the number of data in each cell is often small and the number of cells to be investigated relatively large, one often is not able (or willing) to test these assumptions. One should, however, be aware that these assumptions are made and that large deviations can invalidate the data analysis. In particular much attention should be paid to randomization and blocking issues (see Section 6.2).

Again, as in the preceding section, the considerations about two-way ANOVA can be generalized to multi-way ANOVA.

6.7 Incorporation of interaction in the residual

When the interaction is not significant it can be concluded that there was no reason to include the interaction term in the linear model as is the case in eq. (6.32), so that one should fall back on the linear model of eq. (6.31). This also means that the computation of the interaction SS is no longer required, nor the degrees of freedom reserved for it. Of course, we could then start the ANOVA all over again using Table 6.9 as a model ANOVA. However, there is a much easier way. The calculation of SS_A, SS_B, df_A and df_B in Table 6.11 is unaffected by whether we take interaction into account or not. Without interaction, we would write

$$SS_{R(without)} = SS_T - SS_A - SS_B \tag{6.35}$$

With interaction

$$SS_{R(with)} = SS_T - SS_A - SS_B - SS_{AB} \tag{6.36}$$

Therefore

$$SS_{R(without)} = SS_{R(with)} + SS_{AB} \tag{6.37}$$

In practice, this means the following. Suppose we have computed an ANOVA table with interaction and concluded that the interaction is not significant. We decide therefore to compute the ANOVA table without interaction. Then we can obtain the SS_R by simply summing the SS_R of the previous table (i.e. the one in which interaction was taken into account) with the sum of squares of the interaction term SS_{AB}. In the same way, it is easy to demonstrate that

$$df_{R(without)} = df_{R(with)} + df_{AB} \tag{6.38}$$

In other words, having computed an ANOVA table with interaction and having found that the interaction was not significant, we can obtain the residual sum of squares that would have been obtained if interaction had not been considered by pooling the sums of squares and the degrees of freedom as described above. Having obtained in this way the results that would have been obtained if no interaction had been included, we can then proceed to obtain the MS without having to compute everything again.

Let us apply this to Table 6.11. The sum of squares without interaction is obtained by adding 0.450 (residual sum of squares when interaction taken into account) + 0.092 (sum of squares due to interaction) and the degrees of freedom by adding 30 and 2. The result was already given in Table 6.10. Because the

conclusions were very clear in this case, this has no real influence on the results. In some cases, particularly when the number of degrees of freedom is small, this incorporation is useful. Most computer programs include the possibility of doing this as a matter of course and will even propose it automatically. It should be noted that some statisticians disagree on whether pooling is indeed always acceptable. For more guidance on this matter the reader should refer to Ref. [8].

6.8 Experimental design and modelling

ANOVA is one of the most important statistical techniques in chemometrics. We will apply it repeatedly in later chapters. In Chapters 20 to 25, for instance, techniques of experimental design are discussed. One of the main applications is to decide which factors have an influence on the properties of a process or a product. In Chapter 22 we will discuss an example in which the effect of four main factors and the interactions between each pair of main factors is investigated. Multi-way ANOVA is one of the main tools that is used to decide which factors and which interactions are significant.

Starting with Chapter 8, we will discuss the very important subject of modelling. Suppose we have developed a simple model $y = b_0 + b_1 x$. To do this we have applied regression on a set of replicate y-values, i.e. for certain levels of x we have measured y a few times and obtained the straight-line regression model for these data. We can then predict for each x the value of y we should have obtained. The measured y values will not be exactly equal to the predicted y values. We will then wonder whether we can interpret the variance around the regression line in terms of the model and the variance due to replicate measurements. If this is not the case, an additional source of variance must be present. This will then be due to the fact that the model is not correct (for instance, it is quadratic instead of linear) so that the variance around the straight-line model is larger than could be expected on random variation alone. ANOVA is applied to decide whether this is indeed the case (see Section 8.2.2.2).

6.9 Blocking

Let us go back to the extraction example of Section 6.2. It was decided there that three extraction procedures A, B and C would be carried out twice each on days 1, 2 and 3. The reason was that an inter-day effect was feared. If the ANOVA is carried out, one will therefore consider not only an SS(extraction) and an SS(residual), but also an SS(blocks) (or SS(days) in this case). In general, blocking will therefore lead to an additional factor or factors in the analysis. The block effect is

rarely tested, but the variance due to it — for reasons similar to those explained in the following section — must be filtered away. The construction of the blocks and designs such as Latin squares, which apply the blocking principle, are discussed further in Chapter 24.

6.10 Repeated testing by ANOVA

Let us suppose that we test a treatment for high blood pressure on a group of 10 patients. The blood pressure is measured at the start of the treatment, after 3 weeks and after 6 months. If only two measurements were carried out on each patient (e.g. start and 6 months), then we would be able to carry out a paired *t*-test (one-sided because the alternative hypothesis H_1 would be that the blood pressure at the start is higher than after 6 months). Since there are more than two measurements we must carry out an ANOVA in the case that we want to carry out a single hypothesis test. This is sometimes called *repeated testing* by ANOVA. Repeated testing is the ANOVA equivalent of a paired *t*-test. In the same way that a paired *t*-test would be applied if we had obtained two measurements (at different times or with different techniques) for a set of individuals, samples, etc., we apply repeated testing by ANOVA when there are three or more such measurements for each individual. Conceptually, the experiment is one-way (we want to test one factor, the times, techniques, etc.), but statistically it is a two-way ANOVA.

At first sight, it may appear strange that this should be a two-way ANOVA since we are really interested in only one factor — the effects of time of treatment. However, the total variance in the data is made up by the following components: the measurement error, the effect of time, and also the difference between persons since the persons in the study will not have the same blood pressure. The measurement error is estimated by MS_R and the other two by MS_{time} and MS_{person}. The test is an *F*-test, comparing MS_{time} to MS_R. We could carry out a test on the effect of persons by looking at MS_{person}/MS_R but that would not be useful, since we know that this effect must exist. However, it is necessary to isolate SS_{person}, so that $SS_R = SS_T - SS_{time} - SS_{person}$. Not doing this would be equivalent to writing

$$SS_R = SS_T - SS_{time}$$

and would result in a gross overestimation of the SS_R.

This type of ANOVA is often applied without replication. In this case the effect of interaction cannot be measured. Let us again suppose that we are interested in comparing *k* pretreatment methods; to do this we now select six homogeneous samples with different (but unknown) concentration (instead of six replicates of the same sample as in Table 6.1) and analyze a portion of it once with each of the methods. As we will see in Chapter 13 such a set-up could be applied when the

expected concentration range of the analyte is larger than a few percent. The six samples are analyzed only once according to each pretreatment procedure. Since they have different concentrations, we will need a two-way ANOVA, with pretreatment and sample as factors. Only the pretreatment will be tested, since we know that the concentration adds to the variance. We might ask here the following additional question: is the effect of the pretreatment the same for all samples (at all levels of concentration)? This can be restated as: is there an interaction between samples and pretreatment? Because there was no replication, it is not possible to carry out the test on the interaction. One necessarily assumes there is no interaction.

6.11 Nested ANOVA

Let us suppose we want to analyze the effect of using different instruments and different analysts on the variance of the data obtained. We could make a design, where each analyst performs a few replicate determinations on each of the instruments. In the simplest case (only two instruments and two analysts), we could make the following combinations:

Instrument A — Analyst 1
Instrument A — Analyst 2
Instrument B — Analyst 1
Instrument B — Analyst 2

By carrying out replicated experiments of each combination, we could estimate the effect of the analysts, of the instruments and of the interaction analyst × instrument using two-way ANOVA. This would be a simple example of the ANOVA methods described in Section 6.6.

Now let us suppose that we would like to do something similar with laboratories and analysts as factors. It would not be practical to move analysts from laboratory 1 to laboratory 2: we need another design. This could be the following:

	Analyst 1
Laboratory A	
	Analyst 2
	Analyst 3
Laboratory B	
	Analyst 4

This design is constructed in a hierarchical way. The first effect to be considered is the laboratory and, within each laboratory, the analysts. One of the consequences is that we now cannot determine an interaction between analysts and laboratories.

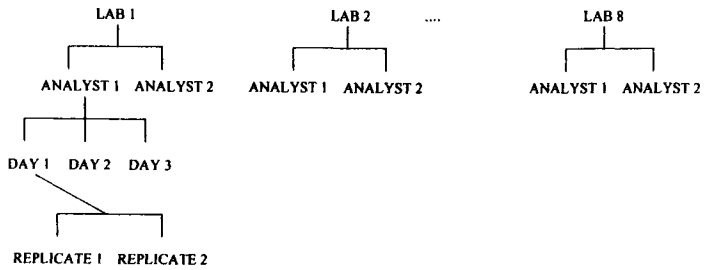


Fig. 6.3. Nested design to trace sources of variability in an interlaboratory study of an analytical method.

TABLE 6.12

ANOVA table for a nested design

Source of variation	df	SS	MS	F
Laboratories (A)	7	5.4248	0.7750	42.82
Analysts within laboratories (B)	8	0.1450	0.0181	1.31
Days within analysts (C)	32	0.4426	0.0138	3.00
Replicates within days (D)	48	0.2227	0.0046	
Total	95	6.2351		

Indeed, the analysts in laboratory A are not the same as in laboratory B. This type of ANOVA is called *hierarchical* or *nested* in contrast with the usual design as described above for instruments–analysts, which is called *crossed*.

A typical example of a hierarchical plan is given by Wernimont [9]. In eight laboratories ($a = 8$) two different analysts ($b = 2$) determined on three days ($c = 3$) the acetyl content of cellulose acetate in two replicates ($n = 2$). This yields the nested design of Fig. 6.3 and the ANOVA Table 6.12.

The sum of squares and the corresponding degrees of freedom in the latter table are obtained as:

$$SS_A = bcn \sum_{i=1}^{a} (\bar{x}_i - \bar{x})^2 \qquad df = a - 1 \tag{6.39}$$

$$SS_B = cn \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{x}_{ij} - \bar{x}_i)^2 \qquad df = a(b - 1) \tag{6.40}$$

$$SS_C = n \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} (\bar{x}_{ijk} - \bar{x}_{ij})^2 \qquad df = ab(c - 1) \tag{6.41}$$

$$SS_D = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \sum_{l=1}^{n} (x_{ijkl} - \bar{x}_{ijk})^2 \qquad df = abc(n-1) \tag{6.42}$$

where:
x_{ijkl} is the value obtained in the ith laboratory by the jth analyst on the kth day for the lth replicate;
$\bar{x}_{ijk}$ is the mean value for the ith laboratory by the jth analyst on the kth day;
$\bar{x}_{ij}$ is the mean value for the ith laboratory by the jth analyst;
$\bar{x}_i$ is the mean value for the ith laboratory;
$\bar{x}$ is the grand mean.

It should be noted that in this case the F-value is obtained by dividing the MS of a factor with the one exactly below. Thus F for laboratories = 0.7750/0.0181 = 42.82 with 7 and 8 degrees of freedom.

The factor laboratories and the factor days are significant. However, this is not the main interest: we would like to know the extent to which each of the factors contributes to the total precision. With σ_D^2 the variance due to the replicates, σ_C^2 that due to the days, σ_B^2 the contribution of the analysts and σ_A^2 that of the laboratories, the mean squares MS_A, MS_B, MS_C and MS_D may be shown to estimate:

$MS_D = SS_D/(abc(n-1))$ estimates σ_D^2

$MS_C = SS_C/(ab(c-1))$ estimates $\sigma_D^2 + n\,\sigma_C^2$

$MS_B = SS_B/(a(b-1))$ estimates $\sigma_D^2 + n\,\sigma_C^2 + cn\,\sigma_B^2$

$MS_A = SS_A/(a-1)$ estimates $\sigma_D^2 + n\,\sigma_C^2 + cn\,\sigma_B^2 + bcn\,\sigma_A^2$

Therefore

$0.0046 = s_D^2$

$0.0138 = s_D^2 + 2s_C^2$

$0.0181 = s_D^2 + 2s_C^2 + 6s_B^2$

$0.7750 = s_D^2 + 2s_C^2 + 6s_B^2 + 12s_A^2$

This can be solved to yield

$s_D = 0.068$ $\quad s_C = 0.068$ $\quad s_B = 0.027$ $\quad s_A = 0.251$

The contribution of the laboratories is by far the largest and that of the analysts the smallest. If we want to obtain better overall reproducibility, the reason for the large variance due to the laboratories must be investigated.

References

1. K. Godelaine, Selectie van ontsluitingsmethoden voor atomaire absorptie analyse van tabletten. Vrije Universiteit Brussel, 1991.
2. W.J. Youden and E.H. Steiner, Statistical Manual of the Association of Official Analytical Chemists, Washington, DC, 1975.
3. C. Lang-Michaut, Pratique des tests statistiques. Dunod, 1990.
4. H.O. Hartley, The maximum F-ratio as a short cut test for heterogeneity of variances. Biometrika, 37 (1950) 308–312.
5. ISO Standard 5725-1986 (E) Precision of test methods – Determination of repeatability and reproducibility for a standard test method for inter-laboratory tests and ISO Standard 5725-2: 1994 (E), Accuracy (trueness and precision of measurement methods and results, Part 2 (1994).
6. R.R. Sokal and F.J. Rohlf, Biometry, 2nd edn. W.H. Freeman, New York, 1981, p. 403.
7. L. Ståhle and S. Wold, Analysis of variance (ANOVA). Chemom. Intell. Lab. Syst., 6 (1989) 259–272.
8. T.A. Bancroft, Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance, Biometrics, 20 (1964) 427–442 or to Sokal p. 284.
9. G. Wernimont, Design and interpretation of interlaboratory studies of test methods. Anal. Chem., 23 (1951) 1572–1576.

Chapter 7

Control Charts

7.1 Quality control

The control chart for industrial product control was developed by Shewhart in 1931 [1] and is the basis of quality control (QC) in statistical process control (SPC) (Chapter 2). The main objective of SPC is to investigate whether a process is in a state of *statistical control.* This requires that characteristics such as central location and dispersion (or in other words, systematic and random errors) do not change perceptibly. Sometimes this requirement is relaxed in order to make sure that tolerance limits are respected. To achieve this purpose, one selects sets of n individual objects or samples from the product line and measures them. Statistics describing the set of n measurements such as the mean for central location and the range for dispersion are plotted as a function of time. One can then observe changes or trends in those statistics.

In analytical chemical QC the purpose is to monitor the performance of a measurement method. The practical question is then whether the method still yields the same result for some (reference) sample (often called check or QC sample). "The same" can then be translated as estimating the same mean value with the same precision. Essentially, this also means that one verifies that the method is in a state of statistical control. It should be noted that the term quality control in analytical chemistry is sometimes used in a wider sense as "all activities undertaken to ensure the required freedom from error of analytical results" [2]. In this chapter, we will take the more restricted view of verifying that the method is in a state of statistical control. The QC is then carried out to ascertain that the method is still sufficiently precise and free of bias.

7.2 Mean and range charts

7.2.1 Mean charts

7.2.1.1 Setting up a mean chart

The principle of the mean chart is shown in Fig. 7.1. The solid line depicts the mean value, $\bar{x}_T$, which is often called the *centre line, CL* and the broken lines are

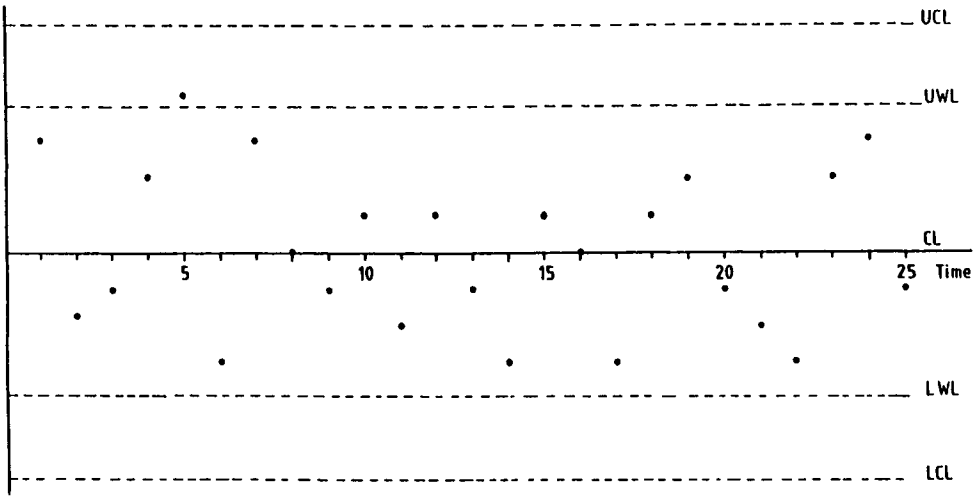


Fig. 7.1. Mean chart. UCL: upper control (or action) limit, UWL: upper warning limit, LWL: lower warning limit, LCL: lower control (or action) limit.

limits at 1.96 S and 3.09 S around the mean. S is a standard deviation. As explained later, there are different ways to define it and therefore we have preferred not to use s as a symbol. The lines at ±1.96 S are called *warning lines* or also the *lower warning limit (LWL)* and *upper warning limit (UWL)* and those at ±3.09 S *action lines* or the *lower control limit (LCL)* and *upper control limit (UCL)*. If the process is under control then the warning lines include 95% of all values and the action lines 99.8%. Finding values outside the warning lines is interpreted as a warning that the process may be getting out of control and outside the action lines as a sign for immediate action to bring the process back under control. This is only a rough indication of how QC charts are interpreted. We will consider this interpretation in more detail later, but let us first study how a QC chart is set up. The first step is the determination of $\bar{x}_T$ and S. They are determined on N sets of n individuals (Table 7.1) before the control procedure starts. Such a set is sometimes called a *training set*.

$\bar{x}_T$ is the estimate of the mean value of the process (or measurement) and is computed as follows

$$\bar{x}_T = \sum \bar{x}_i / N \tag{7.1}$$

where $\bar{x}_i$ is the mean of the ith set of n individual measurements ($i = 1,N$).

There is much more variation in the computation of S. The simplest situation, which often occurs in analytical QC, is the one where $n = 1$. This means that one obtains N individual results for a QC material, with N at least equal to 10, but preferably more (20 is usually considered the acceptable number). S is then simply the standard deviation, s, on individual results of eq. (2.2).

TABLE 7.1

Composition of training set for QC

Measurement j within set	Number of set i				
	1	2	3	i	N
1	x_{11}	x_{12}	x_{13}	x_{1i}	x_{1N}
2					
.					
j	x_{j1}	x_{j2}	x_{j3}	x_{ji}	x_{jN}
.					
n	x_{n1}	x_{n2}	x_{n3}	x_{ni}	x_{nN}
Mean of set	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_i$	$\bar{x}_N$
Standard deviation of set	$s_{\bar{x}_1}$	$s_{\bar{x}_2}$	$s_{\bar{x}_3}$	$s_{\bar{x}_i}$	$s_{\bar{x}_N}$

$$S = \sqrt{\frac{\sum (x_i - \bar{x}_T)^2}{N-1}} \tag{7.2}$$

When $n > 1$, one obtains S by averaging in some way the standard deviations of the N groups of n data. The most evident procedure (see for instance Ref. [3]) is to obtain first the standard error on the mean in each group

$$s_{\bar{x}_i} = \sqrt{\frac{\sum (x_{ij} - \bar{x}_i)^2}{(n-1)\,n}}$$

and to average the variances (since variances are additive).

$$S = \sqrt{\frac{\sum s_{\bar{x}_i}^2}{N}} \tag{7.3}$$

Another procedure simply averages the $s_{\bar{x}_i}$ and then divides the average by a factor, which is often called C_4, and can be found in tables [4] (see Table 7.2).

$$S = \frac{\sum s_{\bar{x}_i}}{NC_4} \tag{7.4}$$

Another variant uses ranges instead of standard deviations to compute the lines, using eq. (2.9):

$$S = \bar{R}/d_n$$

where d_n is Hartley's constant (Section 2.1.4.5). Taking into account that $s_{\bar{x}} = s/\sqrt{n}$ this yields the following limits:

TABLE 7.2

Constants for the determination of S in function of n (for the meaning of the constants, see text). Adapted from Ref. [4]

n	C_4
2	0.798
3	0.886
4	0.921
5	0.940
6	0.951
7	0.959
8	0.965
9	0.969
10	0.973

$$\text{warning lines at } \bar{x}_T \pm \frac{1.96}{d_n\sqrt{n}} \cdot \bar{R}$$
$$\text{action lines at } \bar{x}_T \pm \frac{3.09}{d_n\sqrt{n}} \cdot \bar{R} \tag{7.5}$$

This is often rewritten as

$$\text{warning lines at } \bar{x}_T \pm A'_{0.025}\,\bar{R}$$
$$\text{action lines at } \bar{x}_T \pm A'_{0.001}\,\bar{R} \tag{7.6}$$

where $A'_{0.025}$ and $A'_{0.001}$ are the constants (see Table 7.3) needed to compute the warning lines and action lines, respectively, from $\bar{R}$. The 0.001 and 0.025 refer to the probabilities that a point of a process under control would be higher than the upper lines. They can be understood as follows. If the process is under control and the process errors are normally distributed, then the $\pm 1.96\ S$ lines include 95% of all values that can be expected, 5% should fall outside, i.e. 2.5% should exceed the upper warning limit and 2.5% should be lower than the lower warning limit. In the same way, there is a probability of 0.2% to find a point outside the action limits, i.e. 0.1% to find it higher than the upper action limit and 0.1% lower than the lower action limit. The occurrence of such a point is sufficiently rare to stop the process and reset it.

There are different variants of these mean charts. For instance, instead of drawing lines at 1.96 S and 3.09 S, one often draws them at 2 S and 3 S, so that 95.5% and 99.7% are then included within the warning and action lines respectively.

TABLE 7.3

Constants for the determination of warning and action limits in the mean chart in function of n (for the meaning of the constants, see text). Adapted from Ref. [4]

n	$A'_{0.025}$	$A'_{0.001}$
2	1.229	1.937
3	0.668	1.054
4	0.476	0.750
5	0.377	0.594
6	0.316	0.498
7	0.274	0.432
8	0.244	0.384
10	0.202	0.317

One should note that these charts are based on the assumption of normality. Since we work with means of sets of data, it is probable that the means are normally distributed. Still, it is to be preferred to check this hypothesis. The most frequently occurring problem is that of outliers in the training set. One can verify whether certain of the N groups have an outlying variance using the Cochran test, explained in Section 6.2, and outlying $\bar{x}_i$ can be detected, using for instance the Grubbs test (Section 5.5.2). The reader should remember that, as stated in Section 5.5, when outliers are detected they should not simply be removed, but one should investigate why they occur. In a QC context, this is certainly needed, since they can indicate an instability of the process.

In all cases the estimates $\bar{x}_T$ and S should be representative for the source of error monitored. For instance, when the measurement will be monitored in the routine phase with one measurement/day, then the $n = 20$ training values should be obtained over 20 days, so that the random error includes the between-day component.

Control charts can be updated by incorporating new results in the estimation of $\bar{x}_T$ and S. A typical procedure is as follows. Each time after having plotted e.g. 30 new points on the QC chart, test whether the S_{new}, i.e. the S value for the set of 30 new points is consistent with the S value used until then. One often considers S as a given value, so that one applies the χ^2 test of Section 5.4.2.

7.2.1.2 Application of the mean chart

The occurrence of a point outside the warning limits is by itself not enough to declare the process out of control. However, since the probability of finding a point outside one of two warning limits is only 2.5%, that of two successive points outside the same warning limit is 1/1600, so that when this occurs it is an indication that the process should be inspected and brought under control (reset).

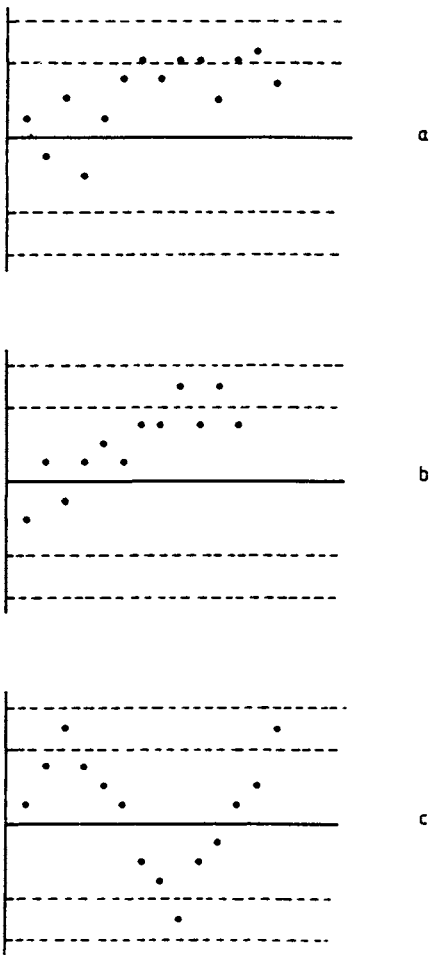


Fig. 7.2. Effects that can be detected with a mean chart: (a) shift or bias; (b) drift; (c) cyclical change.

Mean charts can help to detect the following effects (see Fig. 7.2):

- occurrence of a bias: when consecutive values distribute themselves on one side of the mean value, but remain at a constant level, the trend is called a *shift* (of the mean) (Fig. 7.2a);
- occurrence of a progressively decreasing or increasing trend (*drift* — see also Section 2.6) (Fig. 7.2b);
- cyclical or periodical changes (Fig. 7.2c).

So far, we have seen two rules for taking action based on a mean chart. They are:

1. One point is outside the action limits
2. Two consecutive points are outside the warning limits.

Used in this way the mean charts are very good at detecting either large biases or strongly increased random fluctuations. They are not very good at finding small shifts or slow drifts. For this reason one sometimes adds additional rules such as the following:

3. Seven consecutive points are situated on the same side of the CL or 10 out of 11 consecutive points are found on the same side.

4. Seven consecutive points show an increase (or a decrease).

Rules such as 3 and 4 are based on probability considerations. If the process is under control the probability that a point is situated above or below the line is 0.5 for each of the two possibilities. The probability that 2 consecutive points are on the same side (say, above the line) is $p = (0.5)^2 = 0.25$. Seven consecutive points above the line has a probability $p = (0.5)^7 = 0.007$.

To have exactly 10 points out of 11 above the line, $p = 11(0.5)^{10} \cdot 0.5 = 0.0054$ and for 11 points out of 11, $p = (0.5)^{11} = 0.00049$. For at least 10 points out of 11 on either side of the line p is given by $2(0.0054 + 0.00049) = 0.0118$ or 1/85.

A well known set of rules is known as the *Western Electric rules* [5]. These divide the chart in zones with 7 lines, the LCL and UCL at ± 3 S, the warning lines at ± 2 S, additional lines at ± 1 S which we will call the 1 S lines and the CL. The process is considered out of control or the process has changed when there are:

1. One point outside UCL or LCL.
2. Nine points in a row on one side of the CL.
3. Six decreasing (or six increasing) points in a row.
4. Fourteen points in a row, alternating down and up.
5. Two out of three points outside UWL or LWL.
6. Four out of five points outside the 1 S line on the same side of the CL.
7. Fifteen points in a row within the two 1 S lines.
8. Eight points in a row beyond either of the two 1 S lines.

Let us investigate how good a control chart is at detecting a certain shift. This is determined by the *average run length*, ARL. This is the average number of sets of measurements to be carried out before one detects a given shift (L_1) or the average number of sets of measurements to be carried out before a false alarm is given, i.e. a warning or an action alarm when in fact the process is still under control (L_0). Clearly L_1 is connected to the β-error (Chapter 4) and should be as small as possible, L_0 is related to the α-error and should be large. The probability that a set of n data will fall outside the action limits when the true mean of the process is unchanged is 0.002. It can be shown [6] that if the probability that any sample will fall outside any limits considered is p, that the average number of such samples that will be measured before this happens once, is equal to $1/p$. Therefore, for the above situation $L_0 = 500$. On average, it takes a run of 500 samples before a false alarm will occur.

It is slightly more complex to compute L_1. It must be computed for a change that is considered significant. Again, we can refer to the β-error. In Section 4.8 we

defined β-errors for a given δ (a bias that was considered sufficiently important to necessitate its detection). The same applies here: we must specify what change we want to detect. Let us start by taking a simple example. We suppose that the mean has changed by $3\sigma/\sqrt{n}$. Instead of μ, it has become $\mu + 3\sigma/\sqrt{n}$. What is the probability that we will find a value above the action limit when such a shift occurs? It is now equally probable that values below and above the action limit will be obtained. For this situation $p = 0.5$ and $L_1 = 2$. It will take us on average 2 samples to detect that the mean has increased by $3\sigma/\sqrt{n}$. Let us now carry out the same calculations for a shift of $+\sigma\sqrt{n}$. The mean has become $\mu + \sigma\sqrt{n}$, meaning that the upper action limit is $2\sigma/\sqrt{n}$ away, i.e. 2 standard deviate units. The tail area above $z = 2$ contains 2.28% of all cases or $p = 0.0228$ and $L_1 = 1/p = 44$. This means that we will have to wait on average for 44 time units before the shift is detected. This illustrates that the mean chart is not very sensitive to small shifts.

The situation can be improved by combining rules. Let us consider the "two consecutive points outside warning limits" rule for the same shift (of $+\sigma\sqrt{n}$). This limit is $z = 1$ away from the new mean corresponding to $p = 0.159$. Two consecutive points have a $p = (0.159)^2 = 0.0253$. For a shift of $\sigma\sqrt{n}$ the L_1 for the "two consecutive points outside warning limits" is therefore $1/0.0253 \approx 40$. The combination of the two rules (i.e. "one point outside action limits" and "two consecutive points outside warning limits") yields $L_1 = 24$ [6]. One should not forget that, at the same time, L_0 decreases. As is usual with α and β errors, one has to look for a good compromise.

Small biases or drifts can be detected more easily with methods such as the CUSUM chart (see Section 7.4.2) and periodical changes with autocorrelation charts (see Section 7.5 and Chapter 20).

7.2.2 Range charts

Although some information about the spread of the process can be obtained from the mean chart, it is preferable to control a direct measure of it, such as the range. The range is plotted in function of time in the same way as the mean in the mean chart and, in SPC, one often combines the two plots on the same page using the same time axis. The range chart is shown in Fig. 7.3 for the data of Table 2.6. As for the mean chart, one determines warning and action limits. The distribution of $\overline{R}$ is skewed [4], so that one needs different constants for the upper and lower lines. Since a decreasing spread is no problem for the process, one often uses only the upper limits. When all lines are drawn, this then requires four constants:

Upper action line at $D_{0.001}\,\overline{R}$
Upper warning line at $D_{0.025}\,\overline{R}$
Lower warning line at $D_{0.975}\,\overline{R}$
Lower action line at $D_{0.999}\,\overline{R}$

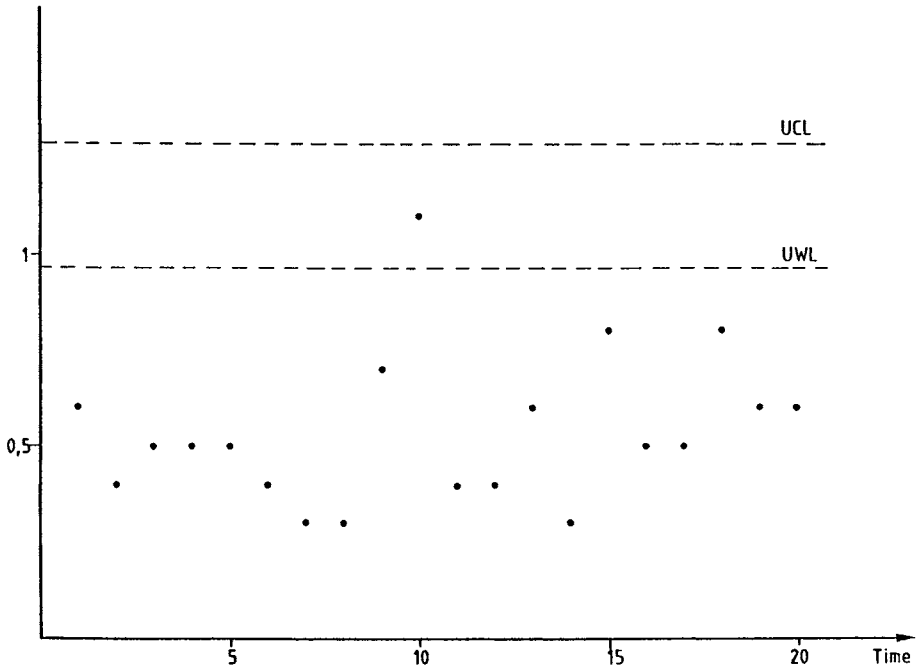


Fig. 7.3. Range chart for the data from Table 2.6. It is assumed that one knows from the training period that $\bar{R} = 0.5$.

As for the constant A' of eq. (7.6), D depends on the sample size n. Some values are given in Table 7.4. For $n = 4$ and supposing that in the training period a value of $\bar{R} = 0.5$ was obtained, one derives UCL = 1.28, UWL = 0.96.

TABLE 7.4

Constants for the determination of warning and action limits in the range chart in function of n (meaning of symbols, see text). Adapted from Ref. [4]

n	$D_{0.001}$	$D_{0.025}$	$D_{0.975}$	$D_{0.999}$
2	4.12	2.81	0.04	0.00
3	2.98	2.17	0.18	0.04
4	2.57	1.93	0.29	0.10
5	2.34	1.81	0.37	0.16
6	2.21	1.72	0.42	0.21
7	2.11	1.66	0.46	0.26
8	2.04	1.62	0.50	0.29
10	1.93	1.56	0.54	0.35

7.2.3 Other charts for central location and spread

The mean chart is by far the most often used one. A variant is the mean chart for unequal sample sizes (i.e. n varies). An alternative that is sometimes used is the *median chart*. It is obtained by plotting all data for each time point and ringing the middle one. Tables for warning and action lines can be found in the specialized literature (e.g., Ref. [4]). In Chapter 12 it is shown that the median is a robust measure of central location and that this is sometimes an advantage. Here it is rather a disadvantage, because it means that the median is not sensitive towards extreme values. For instance the following two sets of data have the same median

2, 3, 4, 5, 6
2, 3, 4, 5, 16

The occurrence of the value 16 would indicate a problem in the process. It would be detected by the mean, but not with the median.

An alternative for the range chart is the standard deviation chart. If a large enough training set is used for that purpose, we can consider that the standard deviation σ of the process is known. During control, sets of n measurements will be obtained and for each of these sets, the sample standard deviations, can be obtained. The question will then be to determine whether s is still compatible with σ. The upper warning line will then be drawn at a value of

$$F = s^2/\sigma^2$$

such that if the process is still under control it will be exceeded only in 2.5% of all cases. Therefore the upper warning line must be drawn at $F = F_{0.025;n-1,\infty}$ (see Chapter 5). In the same way the upper action line will be at $F = F_{0.001;n-1,\infty}$. For instance, for $n = 5$, $F_{0.025;4,\infty} = 2.79$. If σ were 0.1, then any $s > 0.167$ $(= \sqrt{2.79 \times 0.01})$ would be outside the warning lines. Again, it is unusual to draw lower warning and action lines.

7.2.4 Charts for the analytical laboratory

In the analytical laboratory, one often uses quality control only to detect biases. These biases are due to changes in the way in which the laboratory performs the method, or to changes in instruments or reagents. In other words, one tries to detect changes in lab bias (see Chapter 13). When a change in bias is detected the first step is often to check the calibration step. In those cases, for instance where one does not calibrate frequently, the first action will be to re-calibrate and investigate whether this corrects the problem. Charts for precision (and more specifically, repeatability — see Chapter 13) are often restricted to duplicates.

Different types of quality-control samples can be analyzed, such as standard solutions, synthetic materials obtained for instance by spiking the matrix with known amounts of analyte, reference materials and certified reference materials. *Reference materials* are real analytical test materials, that are homogeneous and stable and *certified reference materials* (CRMs) are reference materials that are accompanied by a certificate usually giving an estimate of the concentration of the analyte and the confidence in that estimate [7].

Mean and range charts have been described in Sections 7.2.1 and 7.2.2. The data are expressed as concentrations which means that for methods that require calibration, the errors due to the calibration step have been incorporated. However, there are some special applications:

– The blank chart. This is a special case of the mean chart. The data are now measurements of blanks and the mean is often the mean of one blank, measured at the beginning of a run, and a second blank, measured at the end of the run.

– The recovery chart. This is used when matrix effects are considered possible. The reference material is then usually a spiked matrix material and results are expressed as percentage recoveries.

7.3 Charts for attributes

So far, we have made charts for continuous variables. It is possible to do this also for discrete variables, usually for the number of defects. Discrete variables are described by other probability distributions than the normal distribution and, since the charts we have applied so far are based on the latter distribution, we cannot use these charts for discrete variables. The distributions needed will be described in Chapter 15 as will the related control charts. For now, we will only note that we will make a distinction between two types of situations:

(a) A certain number of objects, e.g. one hundred stoppers, are sampled to determine how many are defective towards a given response (e.g. do not fit on a certain bottle). The result is a ratio: number defectives/number sampled. This is described by the binomial distribution (see Section 15.2).

(b) A certain domain (e.g. 1 m^2 of paint sprayed on a car) is investigated and the number of defects over that area are counted. This is not a ratio and is described by the Poisson distribution (see Section 15.4).

7.4 Moving average and related charts

7.4.1 Moving average and range charts

When observations are made at regular time intervals, the resulting series of observations is called a *time series.* The analysis of time series is discussed further

in Chapter 20 in the context of the characterization of processes, and in Chapter 40 where its use for signal processing is described. Time series are applied in general to separate long-term effects (in signal processing, the signal) from random effects (in signal processing, the noise). In quality control, too, we want to separate systematic effects, such as a shift of the mean, from random effects due to the imprecision of the production and/or measurement. One of the simplest techniques applied in the analysis of time series is the use of *moving averages*. For a series of control measurements $x_1, x_2, \ldots$, we define the moving averages as

$$\frac{x_1 + x_2 + \ldots x_n}{n}, \quad \frac{x_2 + x_3 + \ldots x_{n+1}}{n}, \quad \frac{x_3 + x_4 + \ldots + x_{n+2}}{n}$$

In signal processing it is usual to select an odd value for n and replace the central point in the window of n values by the moving average (see Chapter 40). In QC we plot the moving average at the ends of the window, i.e. at point n, $n+1$, $n+2$, successively. We can then easily use even values of n if we want to do so. Consider for example Table 7.5: the moving average for $n = 4$ is computed. The first value is plotted at time $t = 4$. As can be seen from Fig. 7.4, moving averages have the effect of reducing random variations, so that systematic effects can be more easily observed. The action and warning lines are determined in the conventional way, described by eqs. (7.2) and (7.3), i.e. mean and standard deviation or range are determined from historical data or from a training set.

Some of the rules for detecting the out-of-control situations of Section 7.2.1 cannot be applied. For instance, one cannot apply rules such as "two consecutive points outside the warning limits". Indeed, when one point has been found outside the warning limits, it is quite probable that the next one will also be, because the points are not independent: they use in part the same values.

TABLE 7.5

Moving averages of order $n = 4$

t	Measured value	Moving average
1	12	
2	6	
3	18	
4	11	11.75
5	4	9.75
6	16	12.25
7	22	13.25
8	17	14.75
9	28	20.75
10	18	21.25
11	30	23.25

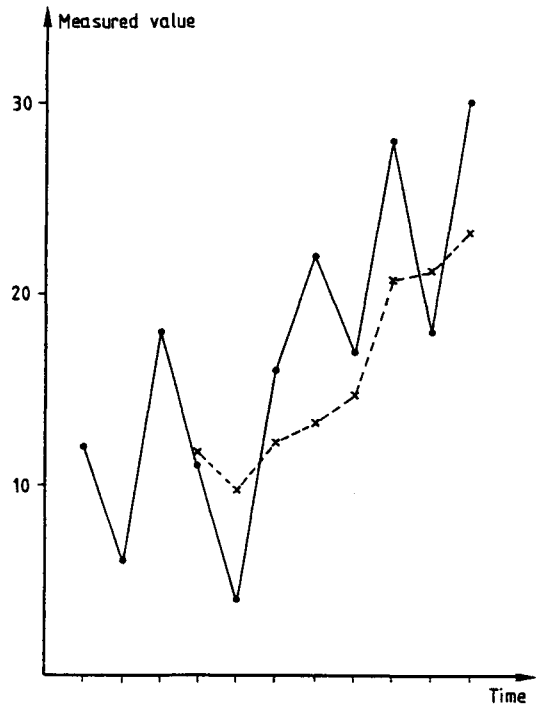


Fig. 7.4. Moving average. Chart displaying the results of Table 7.5.

The principle for the moving average of the observations has been explained, and the same principle can be applied to the ranges.

7.4.2 The cumulative sum (CUSUM) chart

For a series of measurements $x_1,...,x_t$ we determine the *cumulative sum of differences* (CUSUM) between the observed value and the target value $\bar{x}_T$

$$C_1 = x_1 - \bar{x}_T$$

$$C_2 = (x_2 - \bar{x}_T) + (x_1 - \bar{x}_T) = C_1 + (x_2 - \bar{x}_T)$$

$$C_v = \sum_{i=1}^{v} (x_i - \bar{x}_T) = C_{v-1} + (x_v - \bar{x}_T) \qquad (7.7)$$

These values are displayed on a chart such as that in Fig. 7.5 for the data of Table 7.6. The data describe a process for which, during the training phase, values of $\bar{x}_T = 100$, $S = 4$ were derived with $n = 1$. If the deviations from $\bar{x}_T$ are random, then the C values oscillate around the zero line. If a trend occurs, the distance from zero will

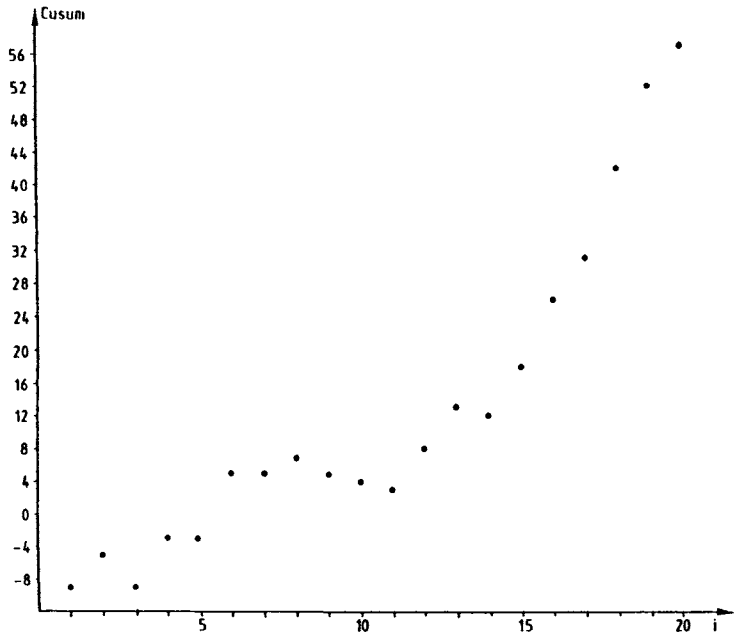


Fig. 7.5. CUSUM chart for the data of Table 7.6.

TABLE 7.6

Example of a CUSUM and an EWMA

Data point	x_i	$x_i - \bar{x}_T$	C_i	x_i^*
1	91	–9	–9	95.5
2	104	+4	–5	99.75
3	96	–4	–9	97.87
4	106	+6	–3	101.94
5	100	0	–3	100.97
6	108	+8	+5	104.48
7	99	–1	+4	101.74
8	103	+3	+7	102.37
9	98	–2	+5	100.19
10	99	–1	+4	99.59
11	99	–1	+3	99.30
12	105	+5	+8	102.15
13	105	+5	+13	103.57
14	99	–1	+12	101.29
15	106	+6	+18	103.64
16	108	+8	+26	105.82
17	105	+5	+31	105.41
18	111	+11	+42	108.21
19	110	+10	+52	109.10
20	105	+5	+57	107.05

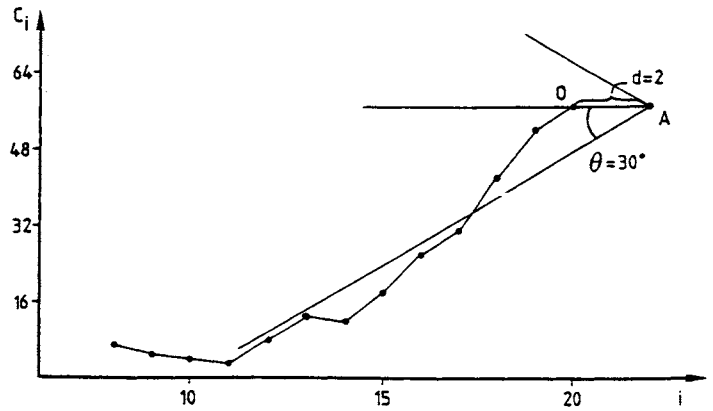


Fig. 7.6. The V-mask method applied to the data of Table 7.6.

gradually increase. Using simple visual observation we would reach this conclusion at around point 15. With the classical mean chart we would detect the shift at time 19 (two points in a row outside warning limits) or with rule 6 of Western Electric (see Section 7.2.1.2) at time 18.

To be able to use the so-called *V-mask* method (see Fig. 7.6) for the interpretation of the CUSUM chart, it is necessary to decide on the scaling of the axes and in particular of the C-axis. The recommended scaling factor is $2\ S/\sqrt{n}$. This means that 1 unit on the $\bar{x}$-axis is equal to $2\ S/\sqrt{n}$ units on the C-axis. When we wish to evaluate a possible trend at time t, we place the mask so that C_t coincides with the point 0. This point is placed at a distance d of the apex A of the V_{mask}. This is called the *lead distance*. When the CUSUM line cuts one of the limits of the mask, then the trend is considered significant and it starts where the mask cuts the line connecting the C values. This is considered equivalent to crossing an action line.

Of course, the detection of the trend depends on the selection of d and of θ, the angle of the V. For the selection of these parameters, we refer to the specialized literature [4,8]. We can base this selection on the average run length ARL (see Section 7.2.1.2), i.e. we can require an L_1-value smaller than a given number (e.g. 5) for a specified shift (e.g. $\pm\ S/\sqrt{n}$) and an L_0-value larger than another number (e.g. 300). Typical values are $d = 2$, $\theta = 30°$; $d = 8$, $\theta = 15°$. The former is used in Fig. 7.6 for time 20. This means that the apex of the mask is at time 22. The line of points crosses the V-mask, so that point 20 is considered to be out of control. In fact, the earliest indication would be received at point 18. In that case point 14 is just outside the mask as the reader can verify by putting the mask at the height of point 18 with its apex at 20.

The V-mask is useful for rapid visual decision making. However, we can also base decisions on a confidence interval. Let us define a gradient or slope in the CUSUM plot. From time v to time $v + m$ the average gradient G is given by

$$G = (C_{v+m} - C_v)/m = \left[\sum_{j=1}^{m} (x_{v+j} - \bar{x}_T) \right] /m \tag{7.8}$$

$$= \left[\left(\sum_{j=1}^{m} x_{v+j} \right) /m \right] - \bar{x}_T \text{ (where } j \text{ is the number of points starting from v)} \tag{7.9}$$

= mean value in time interval – mean value to be controlled.

Reformulated in this way, we can see that the CUSUM chart is really a variant of the moving average method. The CUSUM chart detects no drift when the gradient is equal to zero or, in statistical terminology, when it is not significantly different from zero. We now see that this is equivalent to stating that the mean value in the interval should not differ significantly from the mean value to be controlled. The confidence interval

$$\left(\sum_{j=1}^{m} x_{v+j} \right) /m \pm z_{0.025} S/\sqrt{m} \tag{7.10}$$

is constructed and we verify if $\bar{x}_T$ is inside it. If the result at each time point is itself the average of n measurements, then the square root term becomes $\sqrt{m\,n}$. Let us apply this to the data of Table 7.6. The confidence interval from point 10 (v =10) to point 20 (j = 10) is given by

$$1053/10 \pm 1.96 \times 4/\sqrt{10} = 105.3 \pm 2.5$$

The value of $\bar{x}_T$ is not inside the confidence interval. This confirms that there was a meaningful gradient and that, in this interval, the process was out of control.

7.4.3 Exponentially weighted moving average charts

In signal processing (see Chapter 40), we will see weighted moving averages, where less weight is given to the extreme values in the window moving over the data (i.e. the values closest to the boundaries of the window) than to the central point, with the philosophy that the extreme points are less important to the value of the average of the window. In QC weighted moving averages are also used, but the philosophy is different: it is the last point in the window which is the most important as this describes best the actual state of the process. Weights must therefore be given such that they diminish as they are more distant from the last point. This is what is done with the exponentially weighted average charts [9].

Suppose a certain process has $\bar{x}_T$ = 100 and the observed successive values are for t = 1, ..., 5 respectively 102, 97, 103, 98, 101, then one computes the *exponentially weighted moving average (EWMA)* as follows

$$\text{EWMA} = x^*_{t+1} = x^*_t + \lambda e_t$$

$$= x^*_t + \lambda(x_t - x^*_t) \tag{7.11}$$

This estimate can be re-written as a weighted mean of the last observed x-value and the previous estimate

$$x^*_{t+1} = \lambda x_t + (1 - \lambda)\, x^*_t \tag{7.12}$$

In eq. (7.11) x_t is the value observed at time t and x^*_{t+1} the value at time $t + 1$ predicted from x_t. To make the prediction one needs λ, a constant ($0 < \lambda < 1$) and e_t the difference of the observed value x_t and x^*_t, the value predicted from x_{t-1}.

Suppose $\lambda = 0.5$. To initiate the EWMA we put $x^*_t(t = 1) = \bar{x}_T$, so that

$x^*_2 = 100 + 0.5\,(102 - 100) = 101$

$x^*_3 = 101 + 0.5\,(97 - 101) = 99$

$x^*_4 = 99 + 0.5\,(103 - 99) = 101$

$x^*_5 = 101 + 0.5\,(98 - 101) = 99.5$

$x^*_6 = 99.5 + 0.5\,(101 - 99.5) = 100.25$

The value of λ is chosen by the experimenter. Its value is often chosen to be 0.1 or 0.2 and is sometimes optimized experimentally. Least squares procedures have been described. It has been shown (see Chapter 40) that the eq. (7.11) can be re-written as

$$x^*_{t+1} = \sum_{i=0}^{t} w_i\, x_i \tag{7.13}$$

where w_i are weights ($\sum w_i = 1$) given by

$$w_i = \lambda(1 - \lambda)^{t-i} \tag{7.14}$$

We can see that the nearest point has the highest weight. For instance, for $\lambda = 0.1$ $w_t = 0.1$ and $w_{t-10} = 0.035$. The most recent x, x_t, has a weight of 0.1, while x_{t-10} has a weight of 0.035, i.e. 3 times less than x_t. For $\lambda = 0.3$ it would be about 35 times less. In practice, the value of λ should be optimized and experience shows that it is often situated between 0.1 and 0.3. For ease of calculation, we will compute the EWMA for the data of Table 7.6 with $\lambda = 0.5$.

The procedure consists of plotting on the same chart x^*_t and x_t. If there is no trend or random variation, then x^*_{t+1} can be forecast perfectly from x^*_t and x_t. Since random variation occurs, usually $x^*_{t+1} \neq x_{t+1}$ and $e_{t+1} \neq 0$. There is an error in the forecast of x^*_{t+1} from x_t, so that e is called the forecast error. When there is no trend, e will oscillate around zero, so that x^*_t and x_t do not differ much and the difference will

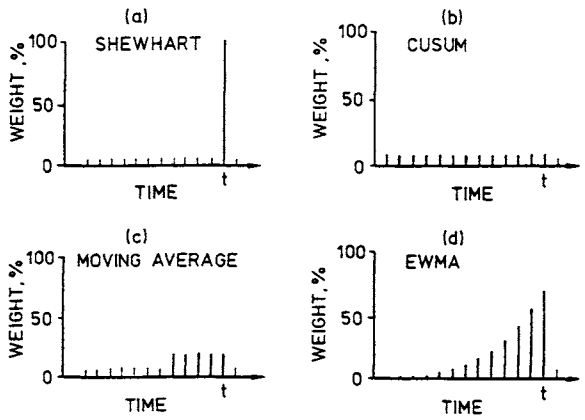


Fig. 7.7. Comparison of weights given to historical data in the control process for the Shewhart, CUSUM, Moving Average and EWMA charts (adapted from Ref. [9]).

not be systematic. When a trend occurs, this will no longer be the case. Suppose the trend is positive, then e_t will also usually be positive and the EWMA will increase. The control limits are given by Hunter [9] as

$$\bar{x}_T \pm 3\,[\lambda/(2-\lambda)]^{0.5}\,S \tag{7.15}$$

For our example, this becomes $\bar{x}_T \pm 1.73\,S$, so that UCL = 106.92. Point 18 crosses the UCL line (see Table 7.6) so that a trend is detected in this point with the (not optimized) $\lambda = 0.5$.

In Fig. 7.7 the difference between the approaches of different charts are shown. In the Shewhart chart a decision is based on the past point: if that point exceeds an action limit, the process is considered out of control whatever happened before. In the CUSUM methods all the points up to the last one are taken into account, since they all influence the CUSUM and have equal weight. As Hunter puts it, the CUSUM has an elephant-like memory. In the moving average, an intermediate solution is adopted: the last k points are used with equal weights, k being the number of points that are used to obtain the moving average. The EWMA is another intermediate solution: the last point is most important, but use is also made of the points before that.

The EWMA is the basis of what is called the *proportional, integral, differential (PID) control equation*. It is given by

$$x^*_{t+1} = x^*_t + \lambda_1\, e_t + \lambda_2 \sum e_t + \lambda_3\, \Delta\, e_t \tag{7.16}$$

The two first terms are the EWMA of eq. (7.11). The third term takes into account the sum of e_t-values and will detect steady drifts away from the target value. The

fourth one is given by the difference between the two last obtained values, e_t and e_{t-1}:

$$\Delta e_t = e_t - e_{t-1}$$

The three terms are weighted to obtain optimal prediction. The name PID is due to the fact that term two is *proportional* to e_t, term three is related to the sum of e_t's (*integral*) and term four to a *difference*. This development of EWMA is more than a quality control equation: it can be considered as a tool for dynamic control and is a member of a class of time series called *ARIMA models* (Autoregressive, integrated, moving average models) [10] (see also Chapter 20).

7.5 Further developments

In Section 7.2.3 we introduced the median chart. We stated that the median is robust towards extreme signal values and that it will be introduced further in Chapter 12 on robust methods. Other methods of this type have been described for QC purposes and the runs test is given in Chapter 12 as an example.

In this chapter, we have discussed situations where only one characteristic is controlled. Suppose that we carry out a quality control of a chromatographic measurement and the quantities of two substances are monitored. With what we have seen so far, we would need two charts, one for each substance. However, the results of the two charts may be related. If something goes wrong with the injection, e.g. a smaller amount is injected, then the results for both substances will be affected in the same way. In other words, the observations in the two charts can be (cor)related. To take this into account we would like a chart which monitors the results as a whole, i.e. in a single chart. How to do this with *multivariate control charts* is explained in Chapter 20.

Another special situation is when cyclical variations occur. In such a case, when point t is for instance high, it is more probable that points $t - 1$ and $t + 1$ are also high than that they are low. Points close to each other have (cor)related values, i.e. they are autocorrelated. This can be taken into account with the use of *autocorrelation charts* (see Chapter 20).

References

1. W. Shewhart, The Economic Control of Quality of Manufactured Products, D. Van Nostrand, New York, 1931.
2. Analytical Methods Committee, Principles of data quality control in chemical analysis. Analyst, 114 (1989) 1497–1503.

3. E. Mullins, Introduction to control charts in the analytical laboratory. Analyst, 119 (1994) 369–375.
4. J.S. Oakland, Statistical Process Control. Wiley, New York, 1990.
5. Western Electric Corp., Statistical Quality Control Handbook. AT&T Technologics, Indianapolis, IN, 1956.
6. D. McCormick and A. Roach, Measurements, Statistics and Computation. Wiley, Chichester, 1987, p. 435.
7. ISO Guide 30: 1992. Terms and definitions used in connection with reference materials.
8. W. Funk, V. Dammann and G. Donnevert, Qualitätssicherung in der Analytischen Chemie. VCH, Weinheim, 1992.
9. J.S. Hunter, The exponentially weighted moving average, J. Qual. Techn., 18 (1986) 203–210.
10. G.E.P. Box and J.M. Jenkins, Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, CA, 1976.

Chapter 8

Straight Line Regression and Calibration

8.1 Introduction

In many situations information about two or more associated variables is obtained in order to study their relationship. Depending on the nature of the variables, the investigation is carried out either by regression analysis or by correlation analysis.

Regression analysis is used to study the relationship between two or more variables. This relationship is expressed as a mathematical function which can also be used for predicting one variable from knowledge of the other(s). To study the dependence of a random variable (the *dependent variable* or the *response variable*) on a variable which is controlled by the experimenter, either because its values are exactly known or can be preselected (the *independent* or the *prediction variable*), *Model I regression* techniques are appropriate. They assume that the independent variable is not subject to error. An important application of Model I regression analysis is *calibration* where an instrumental response is related to the known concentration of the analyte in calibration standards. The final aim of the regression analysis is then to use the mathematical expression, relating the response and the concentration, to predict the concentration of unknown samples. If both variables are subject to error, *Model II regression* techniques, which take into account the error associated with both variables, must be applied. This is the case, for example, in method-comparison studies where there are measurement errors in both methods.

In other applications regression analysis provides a means of simplifying experimental data in order to facilitate their interpretation. The data are represented by an appropriate mathematical model. In the process of *model building*, emphasis is then placed on discovering those independent variables that best explain the variation in the random variable. As an example, consider a chromatographic system. To understand the retention behaviour of the system the retention time (the random variable) can be studied as a function of different system variables (e.g. pH, methanol content of the eluent).

A good theoretical knowledge of the system under study is necessary to construct a model in which the regression parameters have a physical meaning. Empirical models, which do not have a clearly interpretable scientific meaning, are

most often used. Nevertheless such models are very useful, e.g. for prediction purposes, if they provide an adequate description of the data.

The process under study might be understood so well that information about the form of the relationship is available beforehand. The primary goal of regression analysis is then to estimate the regression coefficients which have a well-defined meaning.

In the following sections we will deal only with straight line regression between two variables. Different applications from measurement science are used to introduce and illustrate the method. Multiple and polynomial regression in which several independent variables are involved are discussed in Chapter 10 and nonlinear regression in Chapter 11. Robust regression is explained in Chapter 12 and fuzzy regression in Chapter 19. In Chapter 35 multivariate regression, which studies the association of several response variables with several independent variables, is described.

Correlation analysis is appropriate for studying the degree of association between two random variables: for example, the concentrations of As and Sb in rainwater samples collected at different locations near a copper smelter. The problem here is to find a quantitative measure for the relationship between both concentrations. Correlation analysis is discussed in Section 8.3 of this chapter.

8.2 Straight line regression

8.2.1 Estimation of the regression parameters

The use of a calibration line for determining the concentration of an analyte in a sample is an important application of straight line regression. The variable y then represents the response measurements and the x variable the concentration of the standard solutions. The errors made in preparing the standards are most often negligible in comparison with the measuring errors. Therefore, the assumption that the x variable is exactly known and consequently has no error is justified in calibration. The x variable is then the independent and y the dependent variable. The calibration function can be obtained by fitting an adequate mathematical model through the experimental data.

If we assume that the true relationship between the response and the concentration is a straight line, the model which describes this relationship is:

$$\eta = \beta_0 + \beta_1 x \tag{8.1}$$

η represents the true response; β_0 and β_1 are the *model parameters*, they are the *intercept* and the *slope* of the true but unknown regression line, respectively (see Fig. 8.1a).

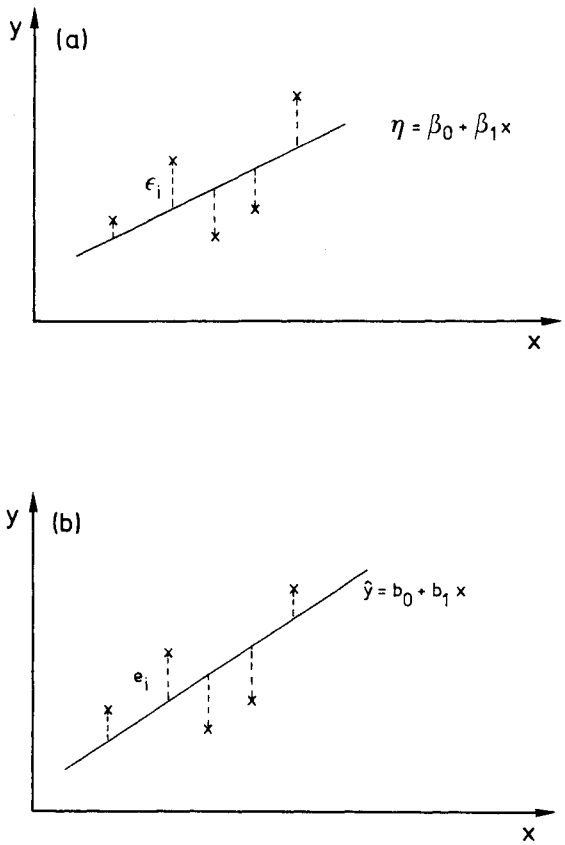


Fig. 8.1. Straight line regression. (a) The true regression line; (b) the estimated regression line.

For any given concentration the true response value is unknown but we have measurement values, y_i, which, due to the fact that these measurements are subject to error, will differ from the true response. Each measurement can therefore be represented as:

$$y_i = \eta_i + \varepsilon_i$$

or

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

This means that each observation is composed of a component which is determined by the model and a component ε_i which represents the difference between the observed response y_i and the true response η_i.

The model parameters, β_0 and β_1, are unknown. However, one can use the information provided by the measurements to obtain estimates, b_0 and b_1, of β_0 and

β_1, respectively. These estimates, b_0 and b_1, are calculated in such a way that the estimated line (Fig. 8.1b)

$$\hat{y} = b_0 + b_1 x \tag{8.2}$$

fits the n experimental points as well as possible. The estimated line, which is an estimate of the true but unknown line, is also called the *least-squares line* when the estimation is performed by the least-squares method. The line fitted by least squares is the one that minimizes the sum of squares of the residuals. The *residual* e_i is the deviation of the measurement y_i from its value predicted by the regression line $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i \tag{8.3}$$

Therefore the least-squares method minimizes R, the sum of the squared residuals:

$$R = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

where $\sum$ in this and all subsequent expressions is the reduced notation of $\sum_{i=1}^{n}$, unless otherwise stated and n is the total number of observation pairs. Differentiating this expression with respect to b_0 and b_1 and setting the results equal to zero provides two simultaneous equations which can be solved for intercept b_0 and slope b_1:

$$\frac{\delta R}{\delta b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\delta R}{\delta b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

This is equivalent to:

$$\sum y_i - n b_0 - b_1 \sum x_i = 0$$

$$\sum x_i y_i - b_0 \sum x_i - b_1 \sum x_i^2 = 0$$

which are the *normal equations* from which the following expressions for the least-squares estimates, b_0 and b_1, can be obtained:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{8.4}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{8.5}$$

with $\bar{y} = (\sum y_i)/n$ the mean of all y_i, and $\bar{x} = (\sum x_i)/n$ the mean of all x_i.

An important statistic in regression analysis is the *residual variance* s_e^2:

$$s_e^2 = \frac{\Sigma(e_i)^2}{n-2} = \frac{\Sigma(y_i - \hat{y}_i)^2}{n-2} \tag{8.6}$$

For the calculation of this variance we divide by $n - 2$ and not as usual by $n - 1$, because the residuals result from a fitted straight line for which two parameters, b_0 and b_1, need to be estimated. This is a measure of the spread of the measurements around the fitted regression line. Consequently it represents the variance in the response which cannot be accounted for by the regression line. Since it is the variance which remains unexplained after x has been taken into account, it is also called the *variance of y given x* and the symbol $(s_{y\cdot x})^2$ is sometimes used. If the model is correct, s_e^2 is an estimate of the variance of the measurements σ^2, also called *pure experimental error.*

For hand calculations the sum of the squared residuals, $\Sigma(y_i - \hat{y}_i)^2$, in eq. (8.6) can be obtained from:

$$\Sigma(y_i - \hat{y}_i)^2 = \Sigma\,(y_i - \bar{y})^2 - \frac{(\Sigma\,(x_i - \bar{x})\,(y_i - \bar{y}))^2}{\Sigma\,(x_i - \bar{x})^2} \tag{8.7}$$

Example 1:

As an example consider the following calibration data (see also Fig. 8.2) for the determination of quinine, according due to Miller and Miller [1]. The response y_i represents the fluorescence intensity (I) in arbitrary units.

i	1	2	3	4	5	6
x_i (ng/ml)	0	10	20	30	40	50
$y_i(I)$	4.0	21.2	44.6	61.8	78.0	105.2

$n = 6$

$\Sigma x_i = 150 \qquad \bar{x} = 25$

$\Sigma y_i = 314.8 \qquad \bar{y} = 52.4667$

$$\Sigma(x_i - \bar{x})^2 = (0-25)^2 + (10-25)^2 + \ldots + (50-25)^2 = 1750$$

$$\Sigma(x_i - \bar{x})\,(y_i - \bar{y}) = (0-25)\,(4.0-52.4667) + (10-25)\,(21.2-52.4667) + \ldots + (50-25)\,(105.2-52.4667)$$

$$= 3468$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})\,(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{3468}{1750} = 1.9817$$

$$b_0 = \bar{y} - b_1\,\bar{x} = 52.4667 - (1.9817 \times 25) = 2.9242$$

Therefore:

$$\hat{y} = 2.924 + 1.982x$$

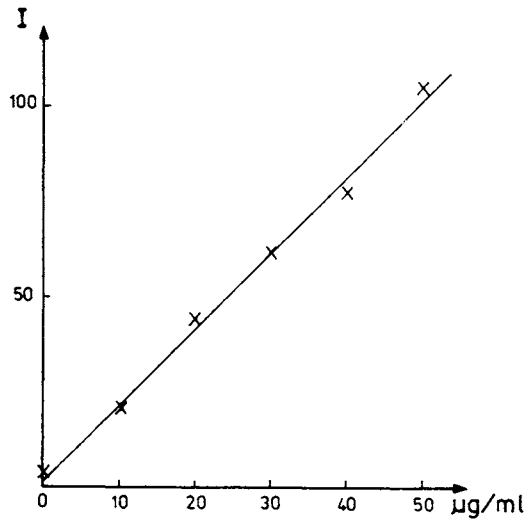


Fig. 8.2. Calibration line for the fluorimetric determination of quinine.

From this fitted line the residual variance s_e^2 can be calculated as follows:

x_i	y_i	$\hat{y}_i$	$e_i = (y_i - \hat{y}_i)$	e_i^2
0	4.0	2.92	1.08	1.1664
10	21.2	22.74	–1.54	2.3716
20	44.6	42.56	2.04	4.1616
30	61.8	62.38	–0.58	0.3364
40	78.0	82.19	–4.19	17.5561
50	105.2	102.01	3.19	10.1761
			$\Sigma e_i = 0$	$\Sigma e_i^2 = 35.7682$
			$\bar{e}_i = 0$	

$$s_e^2 = \frac{\Sigma(y_i - \hat{y}_i)^2}{n-2} = \frac{35.7682}{4} = 8.94$$

In the least-squares method the following assumptions concerning the residuals are made:

(i) for each x_i the residuals e_i are from a population that is normally distributed with mean zero;

(ii) the e_i are independent (see Section 6.2);

(iii) they all have the same variance σ^2. Consequently it is assumed that for each specific x_i the responses y_i are normally distributed with a mean $\eta_i = \beta_0 + \beta_1 x_i$ and a constant variance σ^2. This is shown in Fig. 8.3. For a calibration experiment, the

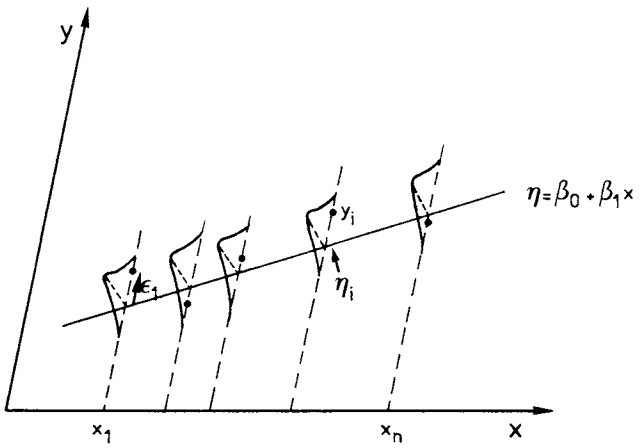


Fig. 8.3. Assumptions concerning the residuals.

latter condition means that the precision of the measurements is independent of the concentration. This condition of uniform variance is called the condition of homoscedasticity (see also Chapter 6).

In many situations it is reasonable to assume that the distribution of measurement errors is normal. Very often, the overall error is a sum of several smaller independently distributed errors. For example the error in flame atomic absorption spectrometry is caused by noise from several sources: the photomultiplier detector, fluctuations in the light source, the instrument electronics, the flame, etc. Whatever the probability distribution of these component errors is, their sum tends to be approximately normal. This is an illustration of the central limit theorem.

On the other hand, the condition of homoscedasticity is certainly not always fulfilled. It is frequently observed that the standard deviation of y, s_y, depends on the value of y or x. In calibration, for example, heteroscedasticity (non-constant variance) may occur with lines that cover a large concentration range. Land et al. [2] illustrate this with several examples from HPLC. In a plasma assay of a haemoglobin-O_2 affinity modifier the variance changed by a factor of 700 over the 0.2–80.0 µg/ml range. Often s_y is proportional to y or x resulting in a constant relative standard deviation (RSD). An example from inductively coupled plasma (ICP) calibration is given in Table 8.1.

To check homoscedasticity, replicate measurements are necessary. Past experience of similar measurements can however be used. In the example given in Table 8.1 the information necessary to check homoscedasticity was obtained from single calibration experiments performed once a week for 14 consecutive weeks. This explains the relatively high RSD value but it indicates that the ICP-Pb measurements are heteroscedastic and that they show a constant relative standard deviation.

TABLE 8.1

ICP–Pb calibration data [3]. Heteroscedasticity with constant relative standard deviation (RSD)

n	x ($\mu g \cdot ml^{-1}$)	$\bar{y}$ (I)	s_y	RSD (%)
14	0.5	0.75	0.164	22
14	1.0	1.49	0.263	18
14	5.0	7.24	1.533	21
14	10.0	14.39	3.096	22
14	50.0	72.17	17.350	24

Knowledge of the *variance function* — the way the variance varies with y or x — is useful to find solutions for the heteroscedasticity problem (see Section 8.2.3).

It is important to note that these assumptions are especially important for prediction purposes to establish confidence intervals for, or tests of hypothesis about, estimated parameters (see Section 8.2.4 and 8.2.5). These intervals, being based on t- and F-distributions, assume that the condition of normality is fulfilled. Moreover, the way they are constructed also assumes the condition of homoscedasticity to be fulfilled.

The regression procedure involves several steps:

1. Selection of a model. Here we have selected the straight line model $\eta = \beta_0 + \beta_1 x$.

2. Establishment of the experimental design which means the choice of the *experimental domain* (in Example 1, this ranges from 0 to 50 ng/ml), the repartition of the x variable over that domain, the number of measurements, etc. The influence of the design of the experiments on the precision of the estimated regression parameters is discussed in Section 8.2.4, and is also treated in Chapter 24 on experimental design.

3. Estimation of the parameters of the model. Here this means estimation of β_0 and β_1 by computing b_0 and b_1 by means of the least-squares method. Other regression methods, which may be useful if departures from the assumption of normality or homoscedasticity occur, are described in Chapter 12.

More complex regression methods for estimating regression parameters when both variables y and x are subject to error are illustrated in Section 8.2.11.

4. Validation of the model. Validation of the model is important to verify that the model selected is the correct one (for instance, is the model really a straight line or are the data better described by a curved line) and to check the assumptions. In the next section it is shown that analysis of the residuals and analysis of variance (ANOVA) are useful for validation purposes.

5. Computation of confidence intervals. In Sections 8.2.4 and 8.2.5 confidence intervals for the regression parameters β_0 and β_1 and for the true line as well as

confidence intervals for the true values of y and x predicted from the regression equation are calculated.

8.2.2 Validation of the model

As already mentioned, validation is necessary (i) to verify that the chosen model adequately describes the relationship between the two variables x and y, or in other words that there is no lack of fit, and (ii) to check the assumptions of normality and constant variance of the residuals. The assumption of independence is generally not tested since this can most often be controlled by a proper experimental set-up. It will be shown how an examination of the residuals and the analysis of variance can be used for validation purposes.

8.2.2.1 Analysis of the residuals

The residuals ($e_i = y_i - \hat{y}_i$) can provide valuable information concerning the assumptions made as well as concerning the goodness (or lack) of fit of the model.

To check the normality of the distribution of the residuals (or also the distribution of the responses y for each specific x_i, see Fig. 8.3) one could apply the techniques to check the normality of data described in Section 3.8. Usually, however, one does not have enough replicate measurements to do this. However, as explained earlier, it can generally be assumed that measurement errors are approximately normally distributed.

Section 6.2 describes how to investigate homoscedasticity. These tests require replicate measurements (to estimate the variance of the response at the different x_i values) which are not always available. As mentioned in the previous section, past experience of similar experiments can then be useful.

Useful information can also be obtained from a residuals plot where the residuals e_i are plotted against $\hat{y}_i$ or against x_i. It is recommended that such a plot is obtained whenever one needs to validate the model. Since no tests are involved, some experience may be necessary for the interpretation of these plots. Some examples are given in Fig. 8.4. Figure 8.4a indicates no abnormality: the residuals are randomly scattered within a horizontal band with a number of positive residuals which is approximately equal to the number of negative residuals. Moreover, a random sequence of positive and negative residuals is obtained. Figure 8.4b indicates that the condition of homoscedasticity is not fulfilled: the scatter of the residuals increases with $\hat{y}$. This indicates that the precision of the measurements over the concentration range considered is not constant. The U-shaped residuals plot in Fig. 8.4c is the result of fitting a straight line to data which are better represented by a curve. There is a lack of fit with the straight line model.

As an example, consider the Ca calibration line obtained from flame atomic absorption spectrometry shown in Fig. 8.5a. An unusual pattern of positive and

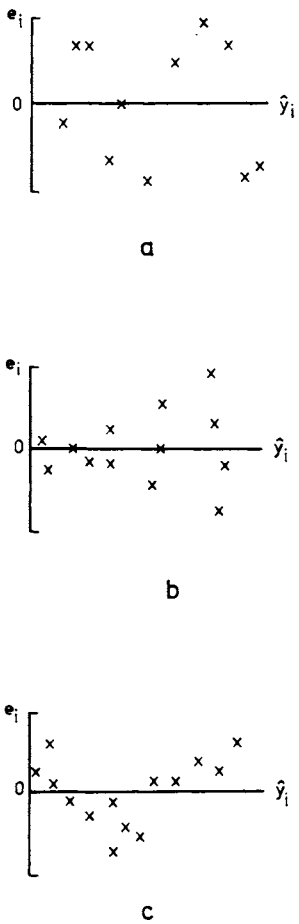


Fig. 8.4. Examples of residual plots.

negative residuals is observed from the residuals plot in Fig. 8.5b: the 19 residuals are arranged in 5 groups (called runs) of respectively 6 negative, 9 positive, 1 negative, 1 positive and 2 negative residuals. The probability that such an arrangement of 19 residuals in 5 runs of positive and negative residuals is random can be shown to be less than 5% (see Section 12.1.4.6). Therefore a non-random arrangement has been detected which has to be attributed here to a (small) deviation of linearity of the Ca calibration line in the low concentration range.

8.2.2.2 Analysis of variance

Analysis of variance (ANOVA) can be used to detect lack of fit in a regression, in order to verify whether the model chosen is the correct one. Therefore replicate measurements are needed.

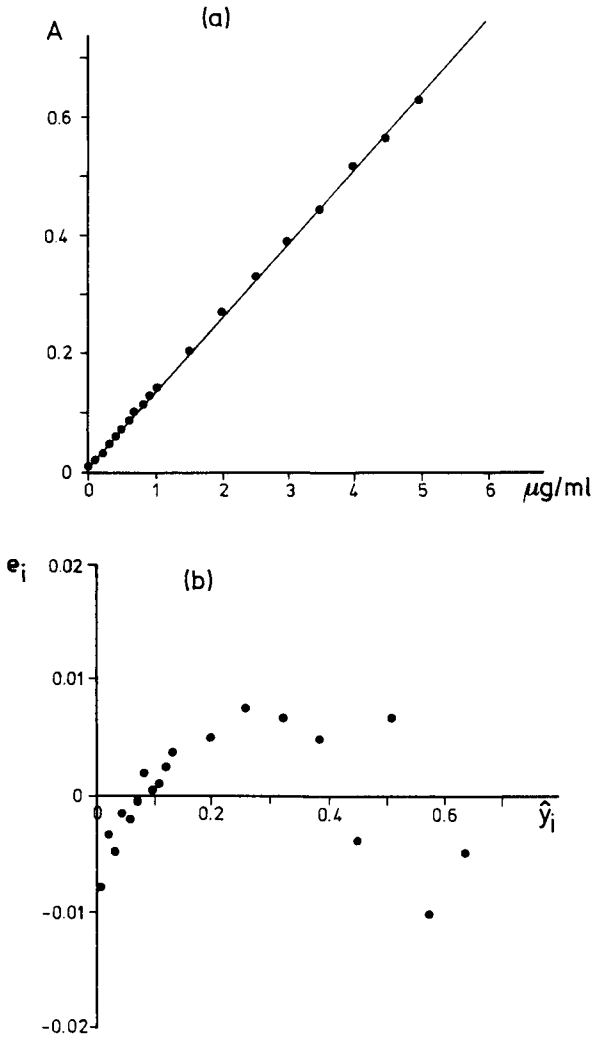


Fig. 8.5. (a) A Ca calibration line obtained from flame atomic absorption spectrometry; (b) The corresponding residual plot.

The total variation of the y values about the mean value, $\bar{y}$, as described by the total sum of squares, SS_T, is then given by:

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \tag{8.8}$$

with:

y_{ij} one of the n_i replicate measurements at x_i,
n_i, the number of replicate measurements made at x_i,
$\sum_{i=1}^{k} n_i = n$, the total number of observations, including all replicate measurements,
k, the number of levels, i.e. different x values,
$\bar{y}$, the mean of all the observations (grand mean).

By an analysis of variance this total sum of squares is split into different sources of variation. Consider first the deviation of the jth response value at x_i, y_{ij}, from the grand mean, $\bar{y}$. This can also be written as:

$$(y_{ij} - \bar{y}) = \underbrace{(y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)}_{\text{residual}} + (\hat{y}_i - \bar{y}) \tag{8.9}$$

with: $\bar{y}_i$ the mean value of the replicates y_{ij} at x_i,
$\hat{y}_i$ the value of y at x_i estimated by the regression function. All replicates at x_i have the same estimated value $\hat{y}_i$.

In this way the deviation has been decomposed into three parts which are represented graphically in Fig. 8.6 and which can be interpreted as follows:

$(\hat{y}_i - \bar{y})$: the deviation of the estimated response at x_i from the grand mean. This quantity depends on the existence of a regression between x and y. It becomes zero if y does not change with x. It is therefore useful to test the significance of the regression (see further).

$(y_{ij} - \hat{y}_i)$: the residual which can be written as $(y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$ with

$(y_{ij} - \bar{y}_i)$: the deviation of an individual observation at x_i from the mean of the observations at x_i. This quantity is independent of the mathematical model chosen;

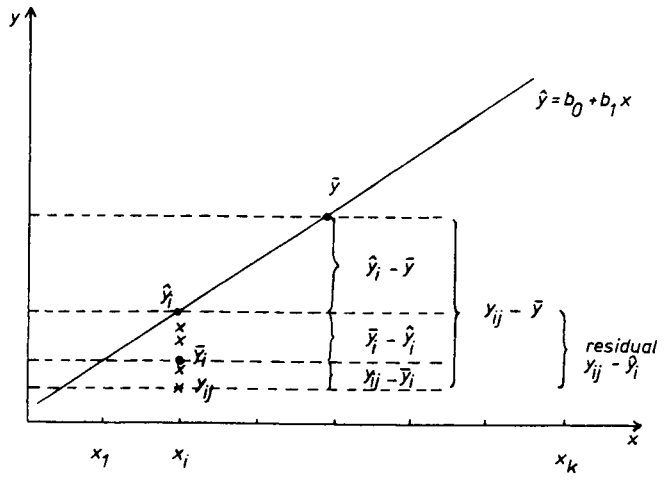


Fig. 8.6. Decomposition of the deviation of y_{ij} from the grand mean $\bar{y}$ into different components.

it only depends on the measurement error. It is important for the estimation of the pure experimental error.

$(\bar{y}_i - \hat{y}_i)$: the deviation of the mean response at x_i from the estimated response at x_i. This quantity depends on the mathematical model chosen. If the model chosen is not the correct one this deviation contains the bias (lack of fit). On the other hand, if the model is adequate, this deviation can be explained in terms of the experimental error. It can then also be used for the estimation of the measurement error (see further).

Squaring both sides of eq. (8.9) and summation over i and j, to include all measurements, yields the total sum of squares SS_T of eq. (8.8):

$$SS_T = \sum_{i}^{k} \sum_{j}^{n_i} (y_{ij} - \bar{y})^2$$

$$= \underbrace{\sum_{i}^{k} \sum_{j}^{n_i} (y_{ij} - \bar{y}_i)^2}_{SS_{PE}} + \underbrace{\sum_{i}^{k} n_i (\bar{y}_i - \hat{y}_i)^2}_{SS_{LOF}} + \underbrace{\sum_{i}^{k} n_i (\hat{y}_i - \bar{y})^2}_{SS_{REG}} \qquad (8.10)$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{SS_R}$$

The sums of cross products cancelled out by summation over j and i.

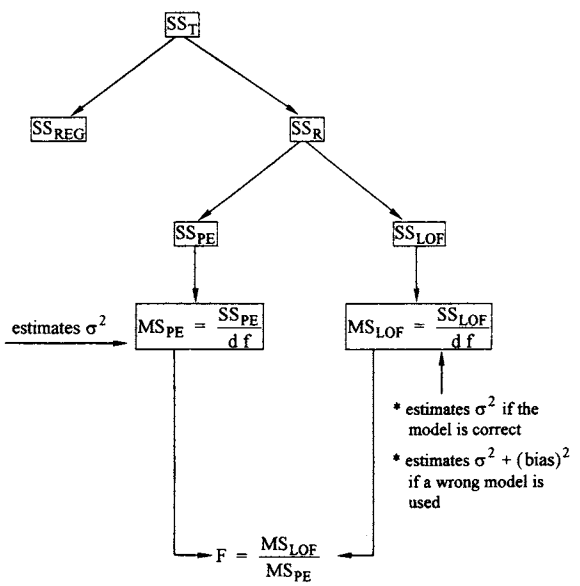


Fig. 8.7. Analysis of variance. Breakdown of sum of squares.

In this way the total variation of the y values about $\bar{y}$ has been separated into two main components (see also Fig. 8.7) namely:

SS_{REG}: the variation which can be ascribed to the regression line i.e. to the fact that y changes with x. It is called the *sum of squares due to regression.* The term *sum of squares due to slope* is also used.

SS_R: the residual variation which measures the variation which cannot be explained by the regression line. This is called the *residual sum of squares* or the *about line sum of squares.* When replicate measurements are available SS_R can be separated into:

- a component which measures the variation due to pure experimental uncertainty. This is the *pure error sum of squares*, SS_{PE}.
- a component which measures the variation of the group means, $\bar{y}_i$, about the regression line. This is called SS_{LOF}, the *sum of squares due to lack-of-fit.*

All this can be arranged in an ANOVA table (Table 8.2) in which the mean squares, MS, are as always obtained by dividing the sums of squares, SS, by their corresponding degrees of freedom (df). MS_{PE} is an estimate of σ^2, the pure error and MS_{LOF} is an estimate of σ^2 if the model chosen is the correct one. It estimates σ^2 + (the bias)^2 if the model is inadequate.

The lack-of-fit test is a one-sided test that is performed by comparing the ratio $F = MS_{LOF}/MS_{PE}$ with the F-distribution at $(k-2)$ and $(n-k)$ degrees of freedom. If this ratio is significant at the chosen significance level (MS_{LOF} significantly larger than MS_{PE}) one concludes that the model is inadequate since the variation of the group means about the line cannot be explained in terms of the pure experimental uncertainty. In this case an examination of the residuals plot can be helpful to adapt the model. If MS_{LOF} and MS_{PE} are comparable, the model is justified and both mean squares are independent estimates of σ^2. Consequently the pooled estimate of σ^2, $MS_R = s_e^2$, is used in all subsequent calculations.

The ANOVA table as represented here also allows us to check the significance of the regression, in other words to check whether a significant amount of the variation of y can be explained by the variation in the independent variable x. For

TABLE 8.2

Analysis of variance of simple regression model with replicate observations

Source of variation	SS	df	MS	F
Regression	SS_{REG}	1	MS_{REG}	MS_{REG}/MS_R
Residual	SS_R	$n-2$	MS_R	
Lack of fit	SS_{LOF}	$k-2$	MS_{LOF}	MS_{LOF}/MS_{PE}
Pure error	SS_{PE}	$n-k$	MS_{PE}	
Total	SS_T	$n-1$		

example, is there a significant effect of the amount of fertilizer (= x) on the yield of wheat (= y)? This is tested by comparing the mean square due to regression, MS_{REG}, with the residual mean square MS_R by means of an F-test. This yields the same conclusion as testing the hypothesis H_0: $\beta_1 = 0$ by means of the confidence interval for the slope or by means of a t-test which will be discussed in Section 8.2.4.1. In a calibration experiment, testing the significance of regression is not relevant because calibration is, by definition, based on the fact that the response of the instrument changes with the concentration of the standard solutions, and thus that there is regression between response and concentration.

Example 2:

As an example of testing lack of fit, consider the data of Table 8.3 which could be the result of a calibration experiment. They are also represented in Fig. 8.8. The different sums of squares necessary to construct the ANOVA table are:

TABLE 8.3

Calibration data for testing lack of fit

x_i	0	1	2	3	4	5
y_{ij}	0.00	0.98	2.10	3.16	3.68	4.15
		0.90	2.20	3.22	3.72	4.27
$\sum y_{ij}$	0.00	1.88	4.30	6.38	7.40	8.42
$\bar{y}_i$	0.00	0.94	2.15	3.19	3.70	4.21
$\hat{y}_i$	0.265	1.114	1.963	2.812	3.661	4.509
$k = 6$	$n = \sum n_i = 11$		$\hat{y} = 0.265 + 0.849x$		$\bar{y} = 2.58$	

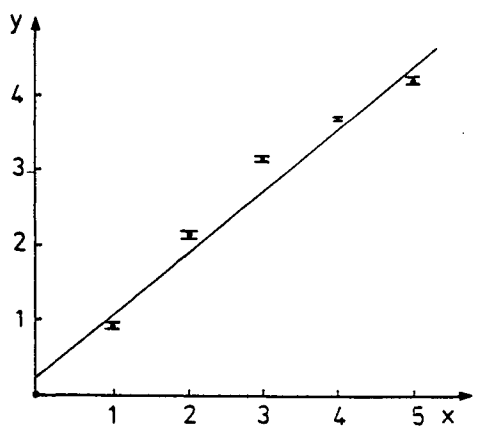


Fig. 8.8. Calibration data of Table 8.3.

TABLE 8.4
ANOVA table

Source of variation	SS	df	MS	F
Due to regression	20.31	1	20.31	
Residual	0.680	9		
Means about line = lack of fit	0.662	4	0.166	46.11
Within group = pure error	0.018	5	0.0036	
Total	20.99	10		

$$SS_T = \sum \sum (y_{ij} - \bar{y})^2$$
$$= (0 - 2.58)^2 + (0.98 - 2.58)^2 + \ldots + (4.27 - 2.58)^2$$
$$= \underline{20.99}$$

$$SS_{PE} = \sum \sum (y_{ij} - \bar{y}_i)^2$$
$$= (0.98 - 0.94)^2 + (0.90 - 0.94)^2 + (2.10 - 2.15)^2 + \ldots + (4.27 - 4.21)^2$$
$$= \underline{0.018}$$

$$SS_{REG} = \sum n_i (\hat{y}_i - \bar{y})^2$$
$$= (0.265 - 2.58)^2 + 2(1.114 - 2.58)^2 + \ldots + 2\,(4.509 - 2.58)^2$$
$$= \underline{20.31}$$

$$SS_{LOF} = 20.99 - 0.018 - 20.31 = \underline{0.662}$$

This yields the ANOVA Table 8.4. Since $F = MS_{LOF}/MS_{PE} = 46.11$ is much larger than $F_{0.05;4,5} = 5.19$, the lack of fit term is highly significant and consequently the straight line model is not adequate to describe the relationship between y and x. If it were necessary to test the significance of the regression, this would not be done until an improved regression model had been found. Note that a residuals plot can be useful to adapt the model. For our example the plot of Fig. 8.8 suggests that the addition of a quadratic term in x might improve the model.

8.2.3 Heteroscedasticity

If the condition of homoscedasticity is violated, the simple least-squares procedure cannot be used without reducing the reliability of the estimations. The

problem of non-constant variance (heteroscedasticity) can be solved either by a transformation of the variables or by using a weighted least-squares procedure.

8.2.3.1 Transformation

The transformation to be used depends on the variance function, i.e. the way the variance of the y values, s_y^2, changes as a function of the response. If the variance is proportional to y, a square root transformation will give a constant variance.

$$\sqrt{y} = b_0 + b_1 \sqrt{x}$$

If the variance is proportional to y^2, which means that s_y is proportional to y, i.e. the relative standard deviation (RSD) is constant, a log transformation can be used:

$$\log y = b_0 + b_1 \log x$$

In our ICP example from Table 8.1 the RSD was found to be constant and indeed the standard deviation of the log transformed responses becomes constant:

x_i	0.5	1.0	5.0	10.0	50.0
$\bar{y}_i$	0.75	1.49	7.24	14.39	72.17
s_{y_i}	0.164	0.263	1.533	3.096	17.350
$s_{\log y_i}$	0.098	0.078	0.092	0.093	0.103

Both the y and x variables are transformed to avoid straight line graphs becoming curved after square root or logarithmic transformation. It should be realized that the transformation carried out to stabilize the variance does not necessarily preserve the straight line relationship. Log–log transformation leads to a straight line only when the intercept is zero or near to zero, which is usually true in calibration. A log transformation has also been recommended for bioanalytical methods using chromatographic procedures [4].

8.2.3.2 Weighted least squares

In weighted least-squares regression the problem of heteroscedasticity is overcome by introducing weighting factors inversely proportional to the variance:

$$w_i = 1/s_{y_i}^2$$

In this way the most importance is given to the most precise observations. This means that we want the calculated line to pass more closely to these points than to the less precise points. The slope and the intercept are then given by:

$$b_1 = \frac{\sum w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum w_i (x_i - \bar{x}_w)^2} \tag{8.11}$$

$$b_0 = \bar{y}_w - b_1 \bar{x}_w \tag{8.12}$$

with $\bar{x}_w = \dfrac{\sum w_i x_i}{\sum w_i}$

$$\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$$

The use of weighted least squares requires information on the errors occurring at different concentration levels. This information must be gained experimentally from a large number of replicate measurements or can be obtained from the variance function relating the variance of the measurements, s_y^2, to y. If the latter is known, the variances s_y^2 can be estimated from this functional relationship. All this is cumbersome and probably explains why the weighted least-squares procedure is less used than it should be.

Example 3:

The data are the same as in Example 1 but here information concerning the precision of the measurements is available, since for each concentration 5 responses have been obtained.

x_i	0	10	20	30	40	50
y_i	4	22	44	60	75	104
	3	20	46	63	81	109
	4	21	45	60	79	107
	5	22	44	63	78	101
	4	21	44	63	77	105
$\bar{y}_i$	4.0	21.2	44.6	61.8	78.0	105.2
s_i	0.71	0.84	0.89	1.64	2.24	3.03
s_i^2	0.50	0.70	0.80	2.69	5.02	9.18

Application of the Cochran test (see Section 6.2.1) to compare the different variances, s_i^2, confirms the presumption of non-constant variance. The computations needed to obtain the weighted regression line are summarized in Table 8.5.

The weighted regression equation is $\hat{y} = 3.481 + 1.964x$. This is very similar to the unweighted regression equation from Example 1, indicating that both lines will yield similar results when used to predict a concentration. However, as will be shown in Section 8.2.5.2 the differences become evident in the prediction errors.

Davidian and Haaland [5] describe an approach to dealing with heteroscedastic data when the variance function is not exactly known. The generalized least squares and variance function estimation (GLS-VFE) method allows the user to postulate a variance model, to estimate the unknown parameters and to use this information to provide more efficient estimates of the regression parameters.

TABLE 8.5

Computations for the weighted regression line of Example 3

x_i	$\bar{y}_i$	s_i	$w_i = 1/s_i^2$	$(x_i - \bar{x}_w)$	$(y_i - \bar{y}_w)$	$w_i(x_i - \bar{x}_w)^2$	$w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w)$
0	4.0	0.71	1.984	−11.97	−22.99	284.269	545.978
10	21.2	0.84	1.417	−1.97	−5.79	5.499	16.163
20	44.6	0.89	1.262	8.03	17.61	81.375	178.457
30	61.8	1.64	0.372	18.03	34.81	120.930	233.476
40	78.0	2.24	0.199	28.03	51.01	156.350	284.532
50	105.2	3.03	0.109	38.03	78.21	157.645	324.202
						$\Sigma = 806.069$	$\Sigma = 1582.808$

$$\bar{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{63.980}{5.343} = 11.97 \qquad \bar{y}_w = \frac{\Sigma w_i y_i}{\Sigma w_i} = \frac{144.240}{5.343} = 26.99$$

$$b_1 = \frac{\Sigma w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\Sigma w_i (x_i - \bar{x}_w)^2} \qquad b_0 = \bar{y}_w - b_1 \bar{x}_w = 3.481$$

$$= \frac{1582.808}{806.069} = 1.964$$

8.2.4 Confidence intervals and hypothesis tests

Once it has been established that the estimated straight line, $\hat{y} = b_0 + b_1 x$, adequately describes the experimental points, it is important to know how precise the estimated parameters (est. par.) b_1, b_0 and $\hat{y}$ are. This is necessary to compute confidence intervals (CI) for the true slope, β_1, intercept, β_0, and response, η. These 100 $(1 - \alpha)$% confidence intervals take the following general form:

100 $(1 - \alpha)$% CI for the true parameter = est. par. $\pm\, t_{\alpha/2;n-2} \cdot s_{\text{est.par.}}$

with $s_{\text{est.par.}}$ the standard deviation of the estimated parameter.

The difference with the confidence interval around the mean calculated in Chapter 3 is in the t-value used. In straight line regression analysis a value of t with $n - 2$, instead of $n - 1$, degrees of freedom is used because the fitted line is based on the estimation of two parameters. As explained in Section 4.3 these confidence intervals can be used to carry out hypothesis tests.

8.2.4.1 Confidence interval for the intercept and the slope

To determine the confidence intervals for the slope and the intercept we need the standard deviations of b_0 and b_1. It can be shown [6] that these can be estimated by:

$$s_{b_0} = s_e \sqrt{\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}} \tag{8.13}$$

$$s_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{8.14}$$

The 95% two-sided confidence intervals for intercept and slope respectively are then calculated as follows:

95% CI for β_0: $b_0 \pm t_{0.025;n-2}\, s_{b_0}$

95% CI for β_1: $b_1 \pm t_{0.025;n-2}\, s_{b_1}$

with $t_{0.025;n-2}$ the value of t with $n-2$ degrees of freedom (see Section 3.7 and Table 3.4). This means that there is 95% probability that the true intercept and slope fall within the limits specified by the confidence interval for β_0 and β_1, respectively.

As an alternative to answer tests of significance concerning β_0 and β_1, t-tests can be applied. To test the hypothesis that the intercept is equal to a specified value, β_0^* e.g. zero, (H_0: $\beta_0 = \beta_0^*$ versus H_1: $\beta_0 \neq \beta_0^*$), the following t is calculated:

$$|t| = \frac{|b_0 - \beta_0^*|}{s_e \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}} = \frac{|b_0 - 0|}{s_e \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}} \tag{8.15}$$

The latter expression is used to test whether the true line might pass through the origin. The calculated t-value is compared with the value of the distribution with $n-2$ degrees of freedom at the chosen significance level. If $|t|$ exceeds the tabulated t we conclude that the intercept is significantly different from zero.

In a similar way the hypothesis that the slope is equal to a specified value, β_1^*, is tested by calculating:

$$|t| = \frac{|b_1 - \beta_1^*|}{s_e \sqrt{\sum (x_i - \bar{x})^2}} \tag{8.16}$$

Testing the hypothesis that the slope is zero (H_0: $\beta_1 = \beta_1^* = 0$ versus H_1: $\beta_1 \neq 0$) is an alternative to testing the significance of the regression by an F-test in the analysis of variance (see end of Section 8.2.2.2).

Example 4:

The following results were obtained for a Tl calibration line by means of graphite furnace AAS. The evaluation of the absorbance signals was in peak area (As). Each measurement was blank-corrected which means that the absorbance measured for the blank has been substracted from each absorbance measurement.

(a) Calculate the confidence limits for the slope and the intercept.

(b) Was the blank correction performed correctly?

x_i (ng/ml)	20	40	60	80	100
y_i (As)	0.038	0.089	0.136	0.186	0.232

The least squares line is:

$\hat{y} = -0.0093 + 0.002425x$

Further:

$\bar{x} = 60$ $\quad \sum x_i^2 = 22000$ $\quad n = 5$ $\quad t_{0.025;3} = 3.18$

$\sum(x_i - \bar{x})^2 = 4000$ $\quad s_e = 0.00145$

(a) The confidence interval for the intercept β_0 is given by:

$$-0.0093 \pm 3.18 \times 0.00145 \sqrt{\frac{22000}{5 \times 4000}}$$

-0.0093 ± 0.0048

Thus CI = [–0.0141; –0.0045]. We can state with 95% confidence that the true intercept, β_0, lies in the interval –0.0141 to –0.0045.

The confidence interval for the slope β_1 is given by:

$0.002425 \pm 3.18 \times 0.00145 / \sqrt{4000}$

0.002425 ± 0.000073

Thus CI = [0.002352; 0.002498]. We can state with 95% confidence that the true slope, β_1, lies in the interval 0.002352 to 0.002498.

(b) Since the measurements were blank-corrected the intercept should be zero. Therefore the hypothesis to be tested is:

$H_0: \beta_0 = 0$ versus $H_1: \beta_0 \neq 0$

This hypothesis can be tested by means of the confidence interval for the intercept or by means of a *t*-test. Since the confidence interval for β_0, calculated in (a) does not include zero, the null hypothesis has to be rejected and consequently one concludes that the intercept is significantly different from zero. This means that the blank value used to correct the absorbances is not representative for the standard solutions and/or the measurement process. Since the blank correction results in a negative intercept the blank absorbance was overestimated.

Of course we come to the same conclusion when the null hypothesis that $\beta_0 = 0$ is tested by means of a t-test (eq. (8.15)):

$$|t| = \frac{|-0.0093|}{0.00145\sqrt{\frac{22000}{5\times4000}}} = 6.11$$

Since at the 5% level of significance the calculated absolute value of t (6.11) is larger than the tabulated t with 3 degrees of freedom (3.18) we again conclude that the intercept is not equal to zero.

An important application of these confidence intervals is as follows: when an analytical chemist develops a new method for the determination of a particular analyte he can validate his method by analyzing spiked blank samples (see Section 13.5.4). If the validation has to be performed at different analyte concentrations, regression analysis can be used. By considering the measured concentration as the y variable and the *added* concentration as the x variable, the slope and the intercept of the regression line can be calculated. In an ideal situation where exactly the same results are obtained, the slope of the regression line should be 1 and the intercept should be 0. This will never occur in practice: even if systematic errors are absent, the presence of random error leads to a scatter of the points around the least-squares line and to small deviations of the estimated slope and intercept from 1 and 0, respectively. A calculated slope that is significantly different from 1 indicates that a proportional systematic error (for instance a matrix effect) is present. A calculated intercept that is significantly different from 0 reveals the presence of a constant systematic error (for instance an incorrect blank correction). The confidence intervals for slope and intercept can serve to carry out these tests of significance.

Example 5:

Consider the data from Table 8.6 which have been adapted from Mannino [7]. In the original article x represents the concentration of Pb in fruit juices measured by flameless AAS and y represents the concentration of Pb measured by a potentiometric method. This situation in which both the x and y variable are subject to error will be treated in Section 8.2.11. In this example we use the same data but consider x as being the concentration of an analyte added and y the concentration of the analyte measured. Consequently x is supposed to be known without error.

From $\hat{y} = 3.87 + 0.963x$, $s_e = 10.56$, $s_{b_0} = 6.64$, $s_{b_1} = 0.0357$ and $t_{0.025;8} = 2.31$, the 95% confidence intervals for the slope and the intercept can be calculated:

95% CI for β_0: 3.87 ± 15.34 (−11.47, 19.21)

Since 0 is included in that interval we have no reason to reject the null hypothesis that the intercept is 0; this means that there is no evidence for a constant systematic error.

TABLE 8.6

Concentration of an analyte added to a sample versus the concentrations measured (adapted from Manino [7])

Sample	Added x	Found y
1	35	35
2	75	70
3	75	80
4	80	80
5	125	120
6	205	200
7	205	220
8	215	200
9	240	250
10	350	330

95% CI for β_1: 0.963 ± 0.083 (0.880, 1.046)

Since 1 is included in that interval we have no reason to reject the null hypothesis that the slope is 1; this means that there is no evidence for a proportional systematic error.

8.2.4.2 Joint confidence region for slope and intercept

In the above example the confidence intervals are computed separately for β_0 and β_1. They specify ranges for the individual parameters irrespective of the value of the other parameter. However, the estimated slope and intercept, b_1 and b_0, are related: if for example in Fig. 8.1 another set of measurement points taken from the same population gives rise to an increased value of b_1, it is likely that this will lead to a decreased value of b_0. The estimates of slope and intercept are not independent and a value for one of the parameters automatically influences that for the other. Therefore, from the individual $100(1-\alpha)\%$ intervals, we cannot say with the same degree of confidence that the null hypotheses $\beta_0 = 0$ and $\beta_1 = 1$ are simultaneously acceptable. This is comparable with the multiple comparison problem discussed in Section 5.2.

If we wish to test the joint hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$, the use of a *joint hypothesis test* or a *joint confidence region* for slope and intercept is required. These take into account the correlation between the estimates (b_0, b_1).

The joint confidence region takes the form of an ellipse with as centre (b_0, b_1). All sets of (b_0, b_1) that fall within the ellipse are considered to be included in the joint confidence interval. The equation for this 95% joint ellipse is given by:

$$(\beta_0 - b_0)^2 + 2\bar{x}(\beta_0 - b_0)(\beta_1 - b_1) + (\sum x_i^2/n)(\beta_1 - b_1)^2 = 2F_{\alpha;2,n-2}\, s_e^2/n \qquad (8.17)$$

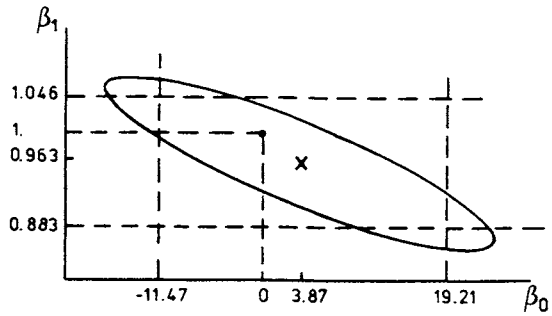


Fig. 8.9. Joint confidence region for β_0 and β_1 from Example 5. The individual confidence limits for β_0 and β_1 are also displayed.

with $F_{\alpha;2,n-2}$ the tabulated F with 2 and $n-2$ degrees of freedom and $\alpha = 0.05$.

This is shown in Fig. 8.9 for Example 5 of Section 8.2.4.1. The tilt of the ellipse with respect to the axes is a result of the negative correlation between b_0 and b_1. Since the point (0,1) lies within the joint 95% confidence region for β_0 and β_1 we can accept simultaneously the slope to be 1 and the intercept to be 0. The individual confidence limits for β_0 and β_1 are also displayed. From this figure it is obvious that individual and joint tests can differ in their results. For instance, the joint values $\beta_0 = -1$ and $\beta_1 = 0.89$ would be accepted as within the confidence interval if they are tested separately, but not if the joint interval is used.

The simultaneous hypothesis concerning slope and intercept can also be tested by an F-test which is a rearrangement of eq. (8.17):

$$F = \frac{(\beta_0 - b_0)^2 + 2\bar{x}(\beta_0 - b_0)(\beta_1 - b_1) + (\sum x_i^2/n)(\beta_1 - b_1)^2}{2s_e^2/n} \tag{8.18}$$

This F value is compared with the F-distribution with 2 and $n-2$ degrees of freedom at the chosen significance level. For example, to test the hypothesis that simultaneously the intercept is zero and the slope is one (H_0: $\beta_0 = 0$ and $\beta_1 = 1$ versus H_1: $\beta_0 \neq 0$ or $\beta_1 \neq 1$) in our previous example, β_0 and β_1 are replaced in eq. (8.18) by 0 and 1, respectively. Since for our example $b_0 = 3.87$, $b_1 = 0.963$, $s_e = 10.57$, $n = 10$, $\bar{x} = 160.5$ and $\sum x_i^2 = 344875$:

$$F = \frac{(-3.87)^2 + 2 \times 160.5(-3.87)(1 - 0.963) + (344875/10)(1 - 0.963)^2}{2(111.72/10)}$$

$$= 0.73$$

Since F is much smaller than the tabulated $F_{0.05;2,8}$ (= 4.46) the null hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$ is accepted.

8.2.4.3 Confidence interval for the true response at a given value of x

To know within what limits the true response η_0, at a particular value x_0 of x, may be expected to lie we need the confidence interval of a point on the true regression line. If $x = x_0$

$$\hat{y}_0 = b_0 + b_1 x_0$$

Using eq. (8.5) this can also be written as

$$\hat{y}_0 = \bar{y} + b_1(x_0 - \bar{x})$$

and it can be shown [6] that the standard deviation of $\hat{y}_0$ is given by:

$$s_{\hat{y}_0} = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{8.19}$$

where n again represents the total number of experimental points used to calculate the regression line.

The 95% confidence interval for a point on the true regression line is then given by:

$$\hat{y}_0 \pm t_{0.025;n-2}\, s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{8.20}$$

This expression should not be used repeatedly to calculate confidence intervals at several different x values in order to find confidence limits that apply to the whole regression line. The latter are obtained by replacing $t_{0.025;n-2}$ by $\sqrt{2F_{0.05;2,n-2}}$ in eq. (8.20). Thus, one takes into account the fact that the true line may have all combinations of values of β_0 and β_1 that lie within the joint confidence region described above. The confidence curves that apply to the whole regression line are two branches of a hyperbola, as represented in Fig. 8.10. The area between these two branches is called the *Working-Hotelling confidence band.*

From eq. (8.20) it can be seen that important terms affecting the width of the confidence band are $(x_0 - \bar{x})^2$ and $\sum(x_i - \bar{x})^2$. The first term reduces to zero when $x_0 = \bar{x}$ and increases as x_0 moves away from $\bar{x}$. Therefore, as can be seen in Fig. 8.10, the confidence intervals are smallest at the mean of the x values and increase away from $\bar{x}$. This means that the best predictions are made in the middle of the regression line. Consequently, extrapolation outside the experimental x-range should certainly be avoided. The term $\sum(x_i - \bar{x})^2$ depends on the design of the experiment, i.e. the repartition of the x_i with respect to $\bar{x}$. Theoretically, in a calibration experiment the smallest confidence intervals are therefore obtained if all standards are situated at both extremes of the calibration range. Then $\sum(x_i - \bar{x})^2$ is a maximum and the confidence interval becomes smaller. Unfortunately with such a repartition of the standards, checking linearity is impossible. Unless a straight line relationship between response and concentration has been

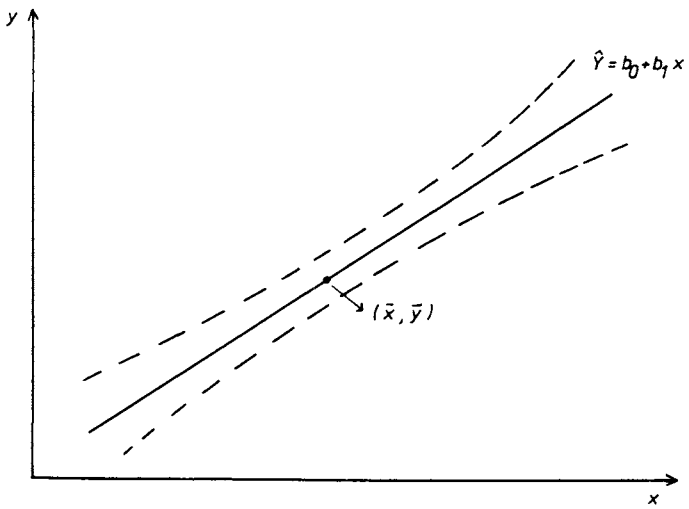


Fig. 8.10. A regression line with confidence limits.

shown the calibration points are therefore usually distributed over more than two x values. The fact that confidence intervals are smaller when the measurement points are situated at the extremes is also used in experimental design (see Chapter 22). The number of experimental points of course also has an effect because t, $1/n$ and $\sum(x_i - \bar{x})^2$ all depend on n.

8.2.5 Predictions made on the basis of the fitted line

The confidence intervals described in the previous section are based on measurements recorded to calculate the fitted line. However, estimation of the parameters and calculation of the confidence intervals for the true regression parameters is generally not the ultimate object of a regression analysis. Often the estimated line will be used in further experiments to predict the value of the y variable (and its associated error) corresponding to a particular value of the x variable or to predict the value of the x variable (and its associated error) from the value measured for the y variable. The corresponding confidence intervals are often called *prediction intervals*.

8.2.5.1 Prediction of new responses

A new individual observation y_0 at x_0 is distributed about η_0 with a variance σ^2. Therefore the uncertainty in y_0, predicted at x_0, is not only composed of the uncertainty of the regression line at x_0 (measured from eq. (8.19) by $s_e^2\,(1/n + (x_0 - \bar{x})^2/\sum(x_i - \bar{x})^2)$) but also of σ^2, the variability of the observation (estimated by s_e^2); these variances

being independent they may be added. Consequently the standard deviation of y_0 predicted at x_0 is:

$$s_{\hat{y}_0} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{8.21}$$

More generally for the prediction of $y_0 = \bar{y}_0$, the mean of m observations performed at x_0, the standard deviation is:

$$s_{\hat{y}_0} = s_e \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{8.22}$$

and the 95% confidence interval:

$$\hat{y}_0 \pm t_{0.025;n-2}\, s_e \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{8.23}$$

where $m = 1$ the confidence interval for a single observation predicted at x_0 is obtained since eq. (8.22) reduces to eq. (8.21). If $m = \infty$, the true mean and consequently a point on the true regression line is obtained: eq. (8.22) reduces to eq. (8.19).

Example 6:

Consider the calibration line calculated in example 1. The 95% confidence limits within which the intensity for a single blank sample ($m = 1$ and $x_0 = 0$) may be expected to lie are:

$$2.924 \pm 2.78 \times 2.99 \sqrt{1 + \frac{1}{6} + \frac{25^2}{1750}} =$$

$$2.924 \pm 10.261$$

This yields a confidence interval of –7.34 to 13.18.

The upper confidence limit at $x_0 = 0$ will play a role in the discussion of the detection limit in Section 13.2.5.

8.2.5.2 Prediction of x from y

In analytical chemistry this is the most important application of the calibration experiment. Indeed the calibration line is used to predict the concentration of an analyte in a sample, x_s, from measurements performed on the sample, $\bar{y}_s$:

$$\hat{x}_s = \frac{\bar{y}_s - b_0}{b_1} \tag{8.24}$$

with $\hat{x}_s$ the predicted concentration and $\bar{y}_s$ the mean of m determinations performed on the sample.

The precision of the estimate depends on the reliability of the fitted line (b_0 and b_1) but also on the precision of $\bar{y}_s$. The determination of the error in the predicted concentration is complex and generally [1,8] the following approximation is used:

$$s_{\hat{x}_s} = \frac{s_e}{b_1} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\bar{y}_s - \bar{y})^2}{b_1^2 \sum(x_i - \bar{x})^2}} \tag{8.25}$$

A different precision for the measurement of the sample and standard solutions can be taken into account in the following way:

$$s_{\hat{x}_s} = \frac{1}{b_1} \sqrt{\frac{s_s^2}{m} + s_e^2 \left(\frac{1}{n} + \frac{(\bar{y}_s - \bar{y})^2}{b_1^2 \sum(x_i - \bar{x})^2} \right)} \tag{8.26}$$

with s_s^2 an estimate of the variance of the sample measurement.

The 95% confidence interval for the true concentration is then:

$$\hat{x}_s \pm t_{0.025;n-2}\, s_{\hat{x}_s} \tag{8.27}$$

which means that there is 95% probability that the true concentration in the sample falls within the limits specified by the confidence interval.

The same calibration line is generally used to predict the concentration in several samples (repeated use of the calibration line). If eq. (8.27) is used to construct confidence intervals about different predicted x-values the probability that all intervals cover the true concentration would be smaller than 95%. This is due to the fact that the individual confidence statements are not independent since they are based on the same regression line. The problem is similar to the multiple comparison problem discussed in Section 5.3. If it is important to control the probability that all confidence intervals include the true concentration, the Bonferroni adjustment described in Section 5.3 can be used.

The factors that have an influence on the confidence limits and thus on the quality of the prediction are the same as those mentioned in Section 8.2.4.3. In addition the confidence limits can be narrowed by increasing the number of measurements, m, performed on the sample.

Example 7:

From the data in Example 1, calculate the confidence limits for the concentration of:

(a) a sample giving a response of 15 units in a single determination;

(b) a sample giving a response of 90 units in a single determination;

(c) a sample giving a mean response of 90 units from 5 separate determinations.

With the following data:

$\hat{y} = 2.924 + 1.982x$; $\bar{y} = 52.5$

$y_s = 15$ yields $x_s = 6.1$

and $(y_s - \bar{y})^2 = (15 - 52.5)^2 = 1406.25$;

$y_s = 90$ yields $x_s = 43.9$

and $(y_s - \bar{y})^2 = (90 - 52.5)^2 = 1406.25$.

Furthermore: $s_e = 2.99$, $\sum(x_i - \bar{x})^2 = 1750$, $t_{0.05;4} = 2.78$, $m = 1$ and $n = 6$.

The confidence limits can be calculated as follows:

(a) 95% confidence limits for the concentration of a sample giving a response of 15 units in a single determination:

$$6.1 \pm 2.78 \frac{2.99}{1.982}\sqrt{\frac{1}{1} + \frac{1}{6} + \frac{1406.25}{1.982^2 \times 1750}} = 6.1 \pm 4.9$$

(b) 95% confidence limits for the concentration of a sample giving a response of 90 units in a single determination:

$$43.9 \pm 2.78 \frac{2.99}{1.982}\sqrt{\frac{1}{1} + \frac{1}{6} + \frac{1406.25}{1.982^2 \times 1750}} = 43.9 \pm 4.9$$

(c) 95% confidence limits for the concentration of a sample giving a mean response of 90 units from 5 separate analyses:

$$43.9 \pm 2.78 \frac{2.99}{1.982}\sqrt{\frac{1}{5} + \frac{1}{6} + \frac{1406.25}{1.982^2 \times 1750}} = 43.9 \pm 3.2$$

Comparison of (a) and (b) confirms, as follows also from Fig. 8.10, that the uncertainty in the prediction of concentrations which are at a comparable distance from the mean concentration is similar. Comparison of (b) and (c) shows how increasing the number of measurements increases the precision of the prediction.

All confidence limits constructed up to now apply to homoscedastic data. As discussed in Section 8.2.3.2, a solution to the problem of heteroscedasticity is to introduce weighting factors which are inversely proportional to the corresponding variance. From Example 3 it was concluded that the weighting process does not have a large influence on the estimated regression equation. In fact, both with

weighted and unweighted regression, unbiased estimates of the regression coefficients are obtained but the variance of these estimates is smaller for the weighted regression procedure [6]. Consequently, in a calibration experiment the sample concentrations predicted by the weighted and unweighted regression line will be very similar. Let us now look at the effect of weighting on the uncertainty in the predicted concentration. The standard error of the predicted concentration is:

$$s_{\hat{x}_s} = \frac{s_e}{b_1} \sqrt{\frac{1}{w_s m} + \frac{1}{\sum w_i} + \frac{(\bar{y}_s - \bar{y}_w)^2 \sum w_i}{b_1^2 (\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2)}} \tag{8.28}$$

with

$$s_e = \sqrt{\frac{\sum w_i (y_i - \hat{y}_i)^2}{n-2}}$$

w_s the weighting factor applied for the sample measurement; $\bar{y}_w$ the weighted mean as defined for eqs. (8.11) and (8.12).

For the homoscedastic situation s_e^2 is an estimate of the common variance σ^2. This of course is not the case here.

Example 8:

From the weighted regression problem in Example 3 calculate the 95% confidence limits for a concentration of:

(a) a sample giving a response of 15 units in a single determination;

(b) a sample giving a response of 90 units in a single determination.

The weighting factors to be applied for both measurements can be obtained from a plot that relates y_i to s_i^2 for the standard solutions. From this plot, shown in Fig. 8.11, appropriate values for the variances at $y_i = 15$ and 90 seem to be 0.6 and 6.9 from which weighting factors respectively equal to 1.67 and 0.145 can be calculated. With the following data (see Table 8.5):

$\hat{y} = 3.481 + 1.964x$; $\bar{y}_w = 26.99$

$y_s = 15$ yields $x_s = 5.9$

and $(y_s - \bar{y}_w)^2 = (15 - 26.99)^2 = 143.76$

$w_s = 1.67$;

$y_s = 90$ yields $x_s = 44.1$

and $(y_s - \bar{y}_w)^2 = (90 - 26.99)^2 = 3970.26$

$w_s = 0.145$

Furthermore, $s_e = 1.921$, $m = 1$, $\sum w_i = 5.343$, $\sum w_i x_i^2 = 1572.2$ and $(\sum w_i x_i)^2 = 63.98^2$.

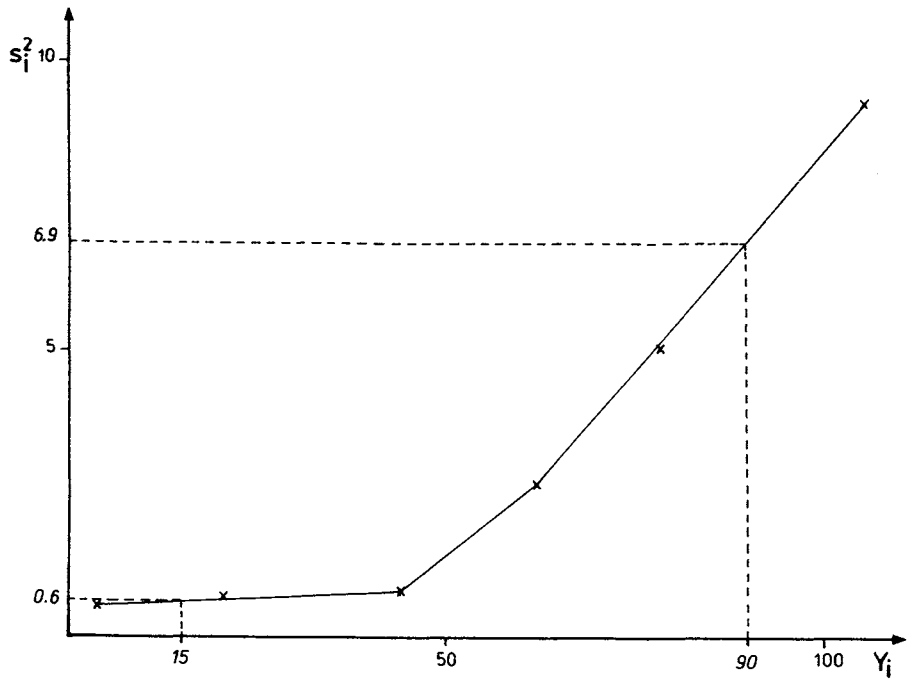


Fig. 8.11. Relation between y_i and s_i^2 for the data of Example 3. From this plot appropriate values for the variances at $y_i = 15$ and $y_i = 90$ are found to be 0.6 and 6.9, respectively.

The confidence limits can be calculated as follows:

(a) 95% confidence limits for the concentration of a sample giving a response of 15 units in a single determination:

$$5.9 \pm 2.78 \frac{1.921}{1.964} \sqrt{\frac{1}{1.67} + \frac{1}{5.343} + \frac{143.76 \times 5.343}{1.964^2\,(5.343 \times 1572.2 - 63.98^2)}} =$$

5.9 ± 2.5

(b) 95% confidence limits for the concentration of a sample giving a response of 90 units in a single determination:

$$44.1 \pm 2.78 \frac{1.921}{1.964} \sqrt{\frac{1}{0.145} + \frac{1}{5.343} + \frac{3970.26 \times 5.343}{1.964^2(5.343 \times 1572.2 - 63.98^2)}} =$$

44.1 ± 7.9

We should compare these results with those obtained in Example 7 in which the same data were treated by means of unweighted regression. Similar predicted concentrations are obtained but the difference between the unweighted and

weighted regression analysis becomes evident if the uncertainty of these predictions are compared. In the weighted regression situation the confidence interval increases with the concentration and this reflects the heteroscedasticity shown by the data. The fact that the confidence interval for the lowest concentration is smaller than with the unweighted regression procedure is a result of the higher weights given to the smallest concentration. The opposite holds for the highest concentration.

8.2.6 Outliers

Since least squares regression consists of minimizing the sum of the squared residuals, the presence of outliers (i.e. observations which are atypical for the rest of the data) can have a large influence on the least squares estimates.

Figures 8.12a–c illustrate different regression data sets with an outlier. Two regression outliers are present in Fig. 8.12d. In both Figs. 8.12a and 8.12b the outlying point is not representative for the linear model fitted by the rest of the data, while in Fig. 8.12c it is atypical because it is remote from the other observations.

According to Rousseeuw [9] the former are called *regression outliers* or influential points since they have a large influence on the regression parameters while the latter is a *leverage point*. This is a point for which the x-value is outlying with respect to the other x-values in the data set. Although leverage points can have a substantial impact on the regression coefficients, Fig. 8.12c shows that this is not necessarily always the case. In fact, in our example the outlying observation can be considered as a *good leverage point* since it fits the model described by the other data quite well. Moreover, it will have a beneficial effect on the confidence intervals of the different estimated parameters described in Sections 8.2.4 and 8.2.5. A *bad leverage point* is an outlier in the x-direction that has an influence on the regression parameters.

Several diagnostics have been proposed for the identification of regression outliers [9]. Some of these diagnostics will be discussed here. How to apply them in the multiple regression situation is shown in Section 10.9. The simplest one consists in a comparison of the absolute value of the *standardized residual* ($|e_i/s_e|$) with a cut-off value which is generally equal to 2 or 3. It is based on the fact that the probability for a residual to have a value as large as 2 or 3 times the residual standard deviation is very small (actually for a normal distribution this probability is 2.3 or 0.13%, respectively). For the different data sets illustrated in Fig. 8.12, Table 8.7 gives the standardized residuals for all data points. It is obvious that, based on the least-squares residuals, this diagnostic fails since for none of the outliers does $|e_i/s_e|$ exceed the value 2 used as cut-off value. The regression outliers are not detected because in order to minimize $\sum e_i^2$, they attract the regression line and inflate the residual standard deviation.

TABLE 8.7

Illustration of different outlier diagnostics

		e_i	$\lvert e_i/s_e \rvert$ (2)*	$CD^2_{(i)}$ (1)*	MD_i^2 (3.84)*	h_{ii} (0.67)*
Data set 1 (Fig. 8.12a)						
x	y					
0	0.0	0.90	0.51	0.30	1.79	0.52
1	1.1	0.30	0.17	0.01	0.64	0.30
2	2.0	–0.49	0.28	0.01	0.07	0.18
3	3.1	–1.08	0.61	0.05	0.07	0.18
4	3.8	–2.07	1.17	0.41	0.64	0.30
5	10.0	2.44	1.38	2.19	1.79	0.52
Data set 2 (Fig. 8.12b)						
x	y					
0	0.0	–0.70	0.22	0.06	1.79	0.52
1	1.1	–0.78	0.25	0.02	0.64	0.30
2	2.0	–1.07	0.34	0.02	0.07	0.18
3	10.0	5.74	1.81	0.44	0.07	0.18
4	3.8	–1.65	0.52	0.08	0.64	0.30
5	5.1	–1.54	0.48	0.27	1.79	0.52
Data set 3 (Fig. 8.12c)						
x	y					
0	0.0	–0.03	0.25	0.03	1.12	0.39
1	1.1	0.08	0.68	0.11	0.50	0.27
2	2.0	–0.01	0.08	0.00	0.12	0.19
3	3.1	0.10	0.84	0.09	0.00	0.17
4	3.8	–0.19	1.61	0.38	0.12	0.19
8	8.0	0.05	0.42	1.63	3.12	0.79
Data set 4 (Fig. 8.12d)						
x	y					
0	0.0	1.19	0.64	0.47	1.79	0.52
1	1.1	0.07	0.04	0.00	0.64	0.30
2	2.0	–1.26	0.67	0.06	0.07	0.18
3	3.1	–2.38	1.27	0.22	0.07	0.18
4	10.0	2.30	1.23	0.45	0.64	0.30
5	10.0	0.08	0.04	0.00	1.79	0.52

*Cut-off value.

Cook's squared distance, $CD^2_{(i)}$, measures the change in the regression coefficients that occurs if the ith observation is omitted from the data. It can be obtained as:

$$CD^2_{(i)} = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_j^{(i)})^2}{p\, s_e^2} \tag{8.29}$$

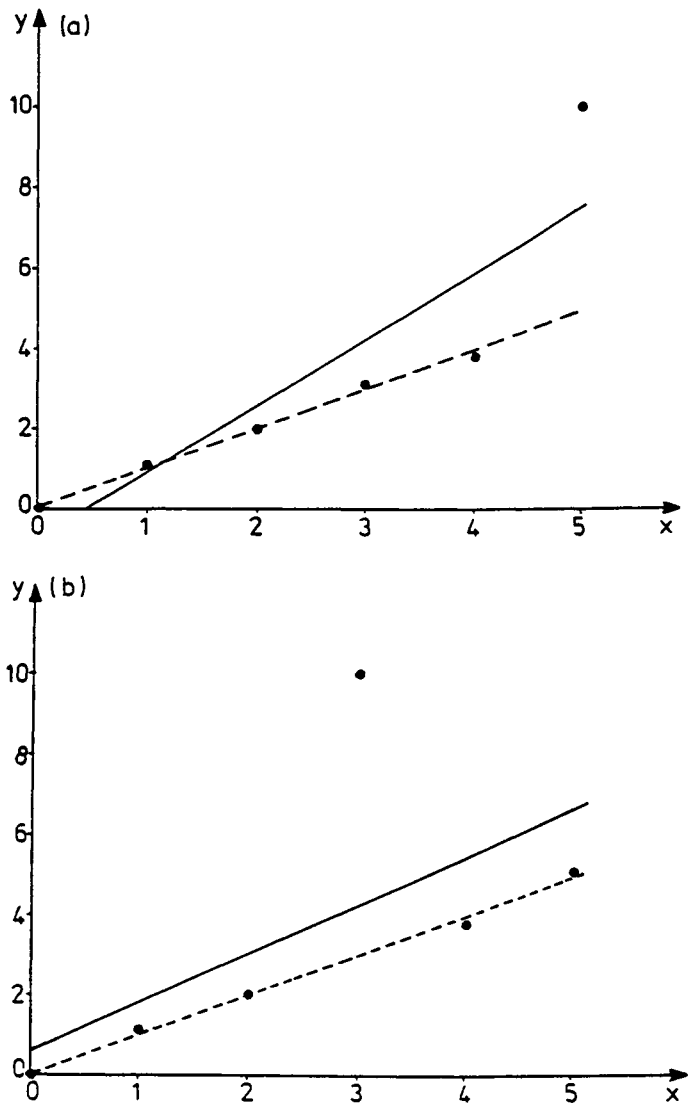


Fig. 8.12. Outliers in regression: a regression outlier (a) and (b); a leverage point (c); two regression outliers (d). The full line represents the regression line based on all data points, the broken line is the regression line obtained without the outlier(s) or leverage point. Data from Table 8.7.

with p the number of regression coefficients to be estimated (for a straight line with an intercept $p = 2$); $\hat{y}_j$ the predicted y-values from the regression equation obtained with all data points; s_e^2 the residual variance for the regression equation based on all data points; and $\hat{y}_j^{(i)}$ the predicted y-values from the regression equation obtained with observation i excluded from the data set.

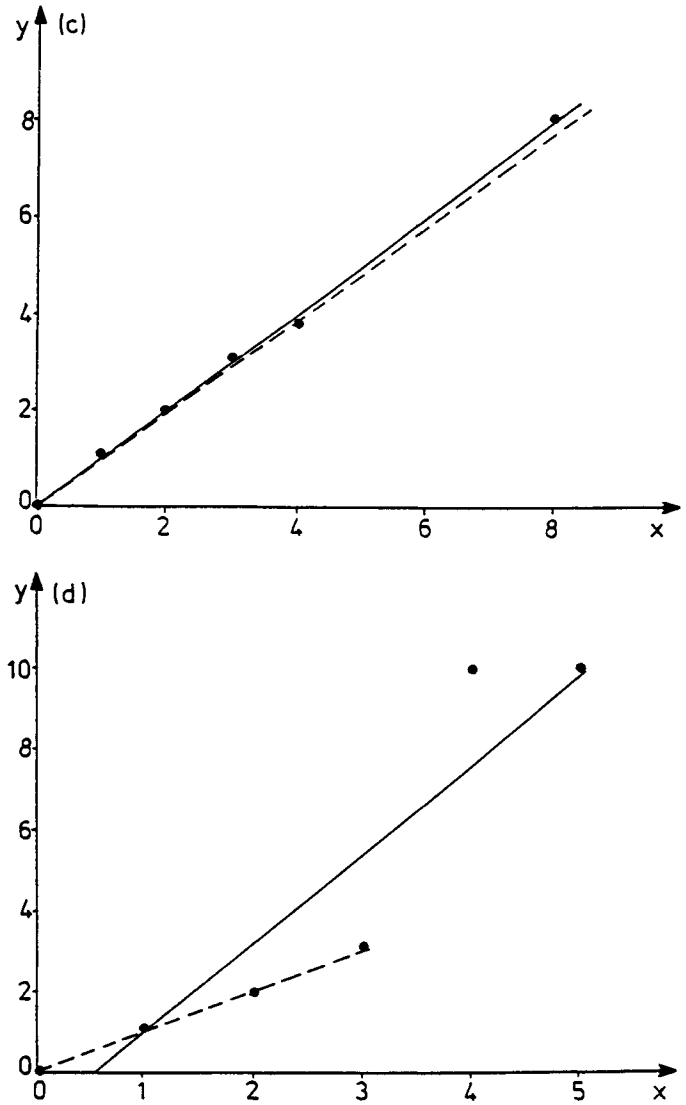


Fig. 8.12 continued.

A large value of $CD^2_{(i)}$ indicates that the ith observation has a large influence on the least squares estimators. Most authors indicate that a $CD^2_{(i)} = 1$ can be considered large. $CD^2_{(i)}$ values for all observations of the different data sets of Fig. 8.12 are also listed in Table 8.7. Cook's squared distance seems to be very sensitive for outliers at the extreme of the data range, but is insensitive to outliers situated in the middle of the range. Indeed, the outlier in data set 1 is detected whereas the one in data set 2 is not. As will be shown later in this section, this is due to the fact that

$CD_{(i)}^2$ also measures how far an x_i-value is from the rest of the x-values. Moreover, as illustrated with data set 4, it is much more difficult to diagnose more than one outlying observation because the influence of one point can be masked by another. Detection of one of the outliers hardly affects the regression line since there is another outlier with a similar effect.

Two related diagnostics for leverage points are the squared *Mahalanobis distance*, MD_i^2, and the *leverage*, h_{ii}, which are given by

$$MD_i^2 = (x_i - \bar{x})^2 / s_x^2 \tag{8.30}$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2} = \frac{1}{n} + \frac{MD_i^2}{n-1} \tag{8.31}$$

with s_x^2 the variance of the x-values. Both these diagnostics are most often applied for multivariate regression situations (see Section 10.9). The MD_i^2-values, which here are the square of the standardized value of x_i, are generally compared with tabulated chi-squared values with $p - 1$ degrees of freedom at the 5% significance level (see Table 5.4) while for the leverage a cut-off value equal to $2\ p/n$ (p representing the number of regression coefficients to be estimated and n the number of observations) is often used [9]. The leverage point in data set 3 has a high MD_i^2-value (without reaching significance) and because of eq. (8.31) also a high h_{ii}-value (exceeding the cut-off value 0.67). Since the squared Mahalanobis distance and the leverage are only based on the x-values (y-values are not taken into account) the same MD_i^2 and h_{ii} are obtained for the other data sets. Consequently the regression outliers are not detected by these diagnostics.

Let us now come back to the Cook's squared distance which is also equal to [9]:

$$CD_{(i)}^2 = \frac{1}{p} \frac{e_i^2\, h_{ii}}{s_e^2 (1 - h_{ii})^2} \tag{8.32}$$

From this expression it follows that $CD_{(i)}^2$ not only reflects how well the model fits y_i, as indicated by e_i, but also how far x_i is from the rest of the x-values, as indicated by h_{ii}. The outlier in data set 2 has a relatively small $CD_{(i)}^2$, despite its considerable influence on the regression line, because with $x = 3$ it is near the mean of the x-values ($\bar{x} = 2.5$). On the other hand, the large $CD_{(i)}^2$ for the leverage point in data set 3, which fits the model described by the other data quite well, can be ascribed to the fact that with $x = 8$ it is far away from the mean x-value ($\bar{x} = 3.16$).

From the above discussion it follows that different diagnostics must be considered to identify outlying observations. However, their interpretation is not straightforward. Another approach is to use robust regression methods which are less easily affected by outliers. These methods are introduced in Chapter 12.

8.2.7 Inverse regression

The main purpose of a calibration experiment is to predict the concentration, x, of a sample from some measurement, y, performed on that sample. Since x has to be inferred from y it is sometimes proposed to regress x directly on y since this is the way the regression equation will be used. Thus:

$$\hat{x} = b_0 + b_1 y$$

This inverse regression is included in some commercial analytical instruments because it facilitates the calculation of the concentration, especially from polynomial models. There has been considerable controversy about this method since the error-free x variable is fitted to the y variable which is subject to error. However, in multivariate calibration (see Chapter 36) inverse least squares is generally preferred to the classical approach.

8.2.8 Standard addition method

In analytical chemistry a calibration line cannot be used to determine an analyte in a sample when the sample matrix is known to interfere with the determination and matrix-matched standards (i.e., standards which have a composition similar to that of the sample) cannot be prepared. A possible solution to this problem is to apply the method of *standard additions* in which the sample is used for performing the calibration.

In the standard addition method small known concentrations of the analyte to be determined are added to aliquots of the unknown sample. These spiked samples as well as the unknown are measured. A typical plot of the added concentration as a function of the measured response is shown in Fig. 8.13. The least-squares regression line is obtained in the usual way and the amount of the analyte present in the sample, x_s, is estimated by extrapolating the line to the abscissa (y = 0). In the absence of absolute systematic errors the negative intercept on the concentration axis corresponds to $-\hat{x}_s$. Consequently $\hat{x}_s = b_0/b_1$.

For the homoscedastic situation the standard error of the predicted concentration, which depends on the reliability of b_0/b_1, can be approximated by [1]:

$$s_{\hat{x}_s} = \frac{s_e}{b_1}\sqrt{\frac{1}{n} + \frac{\overline{y}^2}{b_1^2 \sum (x_i - \overline{x})^2}} \tag{8.33}$$

from which the confidence interval for the concentration is obtained as in eq. (8.27).

For the weighted case the approximation of the standard error is as follows:

$$s_{\hat{x}_s} = \frac{s_e}{b_1}\sqrt{\frac{1}{\sum w_i} + \frac{\overline{y}_w^2}{b_1^2 (\sum w_i x_i^2 - \sum w_i \overline{x}_w^2)}} \tag{8.34}$$

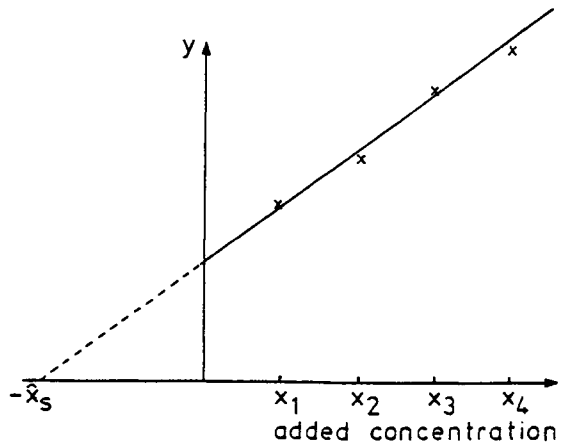


Fig. 8.13. Standard addition line.

in which $s_e^2 = (\sum w_i(y_i - \hat{y}_i)^2)/(n-2)$, $\bar{y}_w = \sum w_i y_i / \sum w_i$, $\bar{x}_w = \sum w_i x_i / \sum w_i$ and b_1 is given by eq. (8.11).

Since in the standard addition method calibration is performed in the sample matrix the technique can be applied in the presence of matrix interferences that introduce relative systematic errors. A major drawback of the method, however, is that it is based on an extrapolation and, as explained in Section 8.2.4.3, this adversely affects the precision. However, a useful application of the method is in the detection of matrix interferences that result in a relative systematic error. These can be revealed by a comparison of the slopes of the standard addition line and an aqueous calibration line. If the matrix does not interfere, we expect both lines to have the same slope. How to check this is explained in Section 8.2.9.

8.2.9 Comparison of the slopes of two regression lines

The comparison of the slopes of two regression lines (represented as b_{11} and b_{12}, respectively) can be performed by means of a *t*-test:

$$t = \frac{b_{11} - b_{12}}{\sqrt{s_{b_{11}}^2 + s_{b_{12}}^2}} \tag{8.35}$$

It follows from eq. (8.14) that:

$$s_{b11}^2 = \frac{s_{e1}^2}{\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2}$$

$$s^2_{b12} = \frac{s^2_{e2}}{\sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2}$$

with n_1 and n_2 the total number of data points in each regression line.

If the residual variances, σ_1^2 and σ_2^2, estimated by s_{e1}^2 and s_{e2}^2, are equal (comparison can be performed by means of an F-test), the pooled estimated variance is calculated as:

$$s^2_{ep} = \frac{(n_1 - 2)\, s^2_{e1} + (n_2 - 2)\, s^2_{e2}}{n_1 + n_2 - 4} \tag{8.36}$$

The test is then performed by calculating

$$t = \frac{b_{11} - b_{12}}{\sqrt{s^2_{ep}\left(\frac{1}{\sum (x_{i1} - \bar{x}_1)^2} + \frac{1}{\sum (x_{i2} - \bar{x}_2)^2}\right)}} \tag{8.37}$$

which should be compared with the tabulated t-value with $n_1 + n_2 - 4$ degrees of freedom at the chosen significance level.

If the residual variances are not equal, an approach similar to the Cochran test for the comparison of two means with unequal variances as described in Section 5.1.1.2 can be used. If $\sigma^2_{b_{11}} \neq \sigma^2_{b_{12}}$, the theoretical t values, t_1 and t_2, at the chosen level of significance and $n_1 - 2$ and $n_2 - 2$ degrees of freedom, respectively, are obtained from a t-table. The following t' is then calculated:

$$t' = \frac{t_1\, s^2_{b_{11}} + t_2\, s^2_{b_{12}}}{s^2_{b_{11}} + s^2_{b_{12}}} \tag{8.38}$$

and the calculated t as obtained from eq. (8.35) is then compared with t' in the usual way. It is not necessary to calculate t' if both regression lines are based on the same number of data points ($n_1 = n_2$). Then $t' = t_1 = t_2$.

The comparison of the slopes of two regression lines is a useful tool in the validation of some analytical methods (see Chapter 13).

Example 9:

As an example, consider the analysis of Al in serum by means of graphite furnace atomic absorption spectrometry. To validate a new method an aqueous calibration line and a standard addition line from a serum sample are compared. Signal evaluation was by means of the integrated absorbances (A.s). The following results are obtained:

Calibration line (1)		Standard addition (2)	
x_{i1} (µg/l)	y_{i1} (A.s)	x_{i2} (µg/l added)	y_{i2} (A.s)
0	0	0	0.050
41	0.039	41	0.083
81	0.073	81	0.122
162	0.149	122	0.161
244	0.215	162	0.179
325	0.280	203	0.215
		325	0.313
$n_1 = 6$		$n_2 = 7$	

$\hat{y}_1 = 8.629\ 10^{-4}x + 0.0033$ $\hat{y}_2 = 8.026\ 10^{-4}x + 0.0533$

$\sum(x_{i1} - \bar{x}_1)^2 = 78379$ $\sum(x_{i2} - \bar{x}_2)^2 = 71582$

$s_{e1}^2 = 1.532\ 10^{-5}$ $s_{e2}^2 = 3.039\ 10^{-5}$

As judged from an F-test, the residual variances can be considered to be similar since $F = 3.039\ 10^{-5}/1.532\ 10^{-5} = 1.98$ which is smaller than $F_{0.05;6,5} = 4.95$.

Consequently the pooled estimated variance is calculated:

$$s_{ep}^2 = \frac{4\times1.532\ 10^{-5} + 5\times3.039\ 10^{-5}}{9}$$

$$= 2.369\ 10^{-5}$$

Therefore:

$$t = \frac{6.03\ 10^{-5}}{\sqrt{2.369\ 10^{-5}\left(\frac{1}{78379} + \frac{1}{71582}\right)}}$$

$$= 2.40$$

As the calculated t (= 2.40) is larger than the tabulated $t_{0.025;9}$ (= 2.26), it should be concluded that the slopes of the aqueous calibration line and the standard addition line are significantly different and that this indicates the presence of matrix effects.

8.2.10 The intersection of two regression lines

In some titrations (e.g., conductometric and photometric) the end point is obtained as the intersection of two straight lines. If

$y_1 = b_0 + b_1x_1$ with n_1 data points

and

$y_2 = b'_0 + b'_1\ x_2$ with n_2 data points

are the two lines, their estimated point of intersection is

$$\hat{x}_I = \frac{(b_0 - b'_0)}{(b'_1 - b_1)} = \frac{\Delta b_0}{\Delta b_1} \tag{8.39}$$

The limits of the 95% confidence interval for the true value of this estimate, $\hat{x}_I$, can be obtained as the roots of the following equation [8,10]:

$$\hat{x}_I^2 \left((\Delta b_1)^2 - t^2\, s^2_{\Delta b_1}\right) - 2\hat{x}_I\,(\Delta b_0\, \Delta b_1 - t^2\, s_{\Delta b_0 \Delta b_1}) + \left((\Delta b_0)^2 - t^2\, s^2_{\Delta b_0}\right) = 0 \tag{8.40}$$

with $t = t_{0.025;n_1+n_2-4}$, the tabulated t value at $n_1 + n_2 - 4$ degrees of freedom

$$s^2_{\Delta b_1} = s^2_{ep}\left(1/\sum (x_{i1} - \bar{x}_1)^2 + 1/\sum (x_{i2} - \bar{x}_2)^2\right) \tag{8.41}$$

$$s^2_{\Delta b_0} = s^2_{ep}\left(1/n_1 + 1/n_2 + \bar{x}_1^2/\sum (x_{i1} - \bar{x}_1)^2 + \bar{x}_2^2/\sum (x_{i2} - \bar{x}_2)^2\right) \tag{8.42}$$

$$s_{\Delta b_0 \Delta b_1} = s^2_{ep}\left(\bar{x}_1/\sum (x_{i1} - \bar{x}_1)^2 + \bar{x}_2/\sum (x_{i2} - \bar{x}_2)^2\right) \tag{8.43}$$

Notice that it is assumed here that the error variances s^2_{e1} and s^2_{e2} are comparable since they are pooled into s^2_{ep} (eq. (8.36)).

Example 10:

The following results are obtained for the conductometric titration of 0.1 M HCl with 0.1 M NaOH. They are also represented in Fig. 8.14. The end point of the titration is the point of intersection of the two lines.

Line 1		Line 2	
x	y	x	y
ml NaOH	arbitrary units	ml NaOH	arbitrary units
3.0	430	25.5	129
6.0	388	27.0	147
9.0	343	30.0	181
12.0	302	33.0	215
15.0	259	36.0	251
18.0	214		
21.0	170		

$\hat{y}_1 = 474.00 - 14.43\,x$
$s^2_{e1} = 1.314$
$s_{b_1} = 0.07222$
$s_{b_0} = 0.96890$
$n_1 = 7$
$\bar{x}_1^2 = 12^2 = 144$
$\sum (x_{i1} - \bar{x}_1)^2 = 252.0$

$\hat{y}_2 = -165.45 + 11.55\,x$
$s^2_{e_2} = 0.4146$
$s_{b'_1} = 0.07496$
$s_{b'_0} = 2.28934$
$n_2 = 5$
$\bar{x}_2^2 = 30.3^2 = 918.09$
$\sum (x_{i2} - \bar{x}_2)^2 = 73.8$

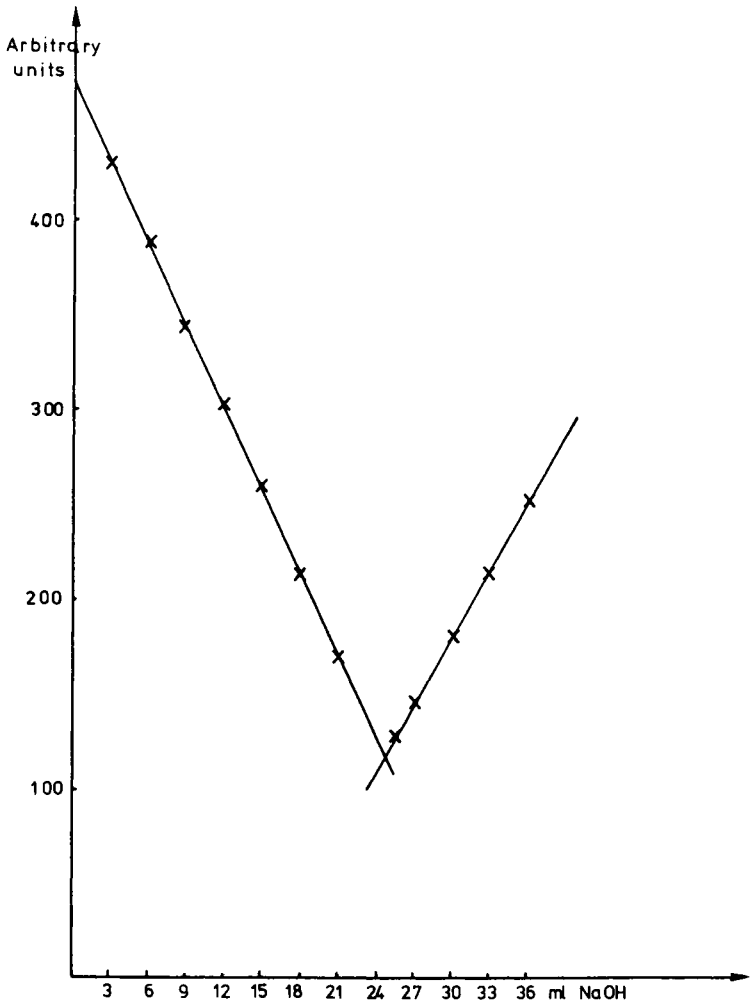


Fig. 8.14. Conductometric titration of 0.1 M HCl with 0.1 M NaOH. Data from Example 10.

This information is necessary to calculate $\hat{x}_I = 24.61$ ml from eq. (8.39), $s^2_{\Delta b_1} = 0.0171$ from eq. (8.41), $s^2_{\Delta b_0} = 13.043$ from eq. (8.42), $s_{\Delta b_0 \Delta b_1} = 0.4475$ from eq. (8.43) and the pooled variance $s^2_{ep} = 0.9767$ from eq. (8.36).

With $t_{0.025;8} = 2.306$ the 95% confidence interval for the true end point, x_I, is obtained from eq. (8.40) which becomes

$$674.8695\, x_I^2 - 2x_I(16610.5313) + 408826.9445 = 0$$

Therefore the 95% confidence limits for the true end point, estimated by $\hat{x}_I = 24.61$ ml, are [24.51 ml, 24.72 ml].

8.2.11 Regression when both the predictor and the response variable are subject to error

Up to now it has been assumed that only the response variable, y, is subject to error and that the predictor variable, x, is known without error (Model I regression). However, there are situations for which the assumption that x is error free is not justified. An example of the Model II regression case is the comparison of results obtained under different experimental conditions (e.g. different measurement or pretreatment techniques) where both the x and y variables are measured quantities which are subject to experimental error. In general, the study of the relationship between two variables that are measured quantities or that show natural variability (examples from the biological science are weights and lengths) require regression methods that take the error in both variables into account. They are called *errors-in-variables regression methods.*

If η_i represents the true value of y_i and ξ_i the true value of x_i then:

$$y_i = \eta_i + \varepsilon_i \tag{8.44}$$

$$x_i = \xi_i + \delta_i \tag{8.45}$$

with ε_i and δ_i the experimental errors.

The model which describes the straight line relationship between η_i and ξ_i is

$$\eta_i = \beta_0 + \beta_1 \xi_i \tag{8.46}$$

Consequently the combination of eq. (8.46) with eq. (8.44) and eq. (8.45) yields:

$$y_i = \beta_0 + \beta_1(x_i - \delta_i) + \varepsilon_i$$

or

$$y_i = \beta_0 + \beta_1 x_i + (\varepsilon_i - \beta_1 \delta_1) \tag{8.47}$$

where $(\varepsilon_i - \beta_1 \delta_i)$ represents the error term.

If the error in x is neglected and the regression coefficients are estimated as described in Section 8.2.1, by minimizing

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

it can be shown [6] that the least-squares slope b_1 is a biased estimator of β_1. The error term and x_i in eq. (8.47) are correlated since both depend on δ_i. This invalidates the use of ordinary least squares, in which the error term is assumed to be independent of x_i.

Since both variables are affected by random measurement errors (here we assume $\sigma_\varepsilon^2 = \sigma_\delta^2$) an unbiased estimation of the regression coefficients can be obtained by minimizing $\sum d_i^2$, i.e. the sum of the squares of the perpendicular

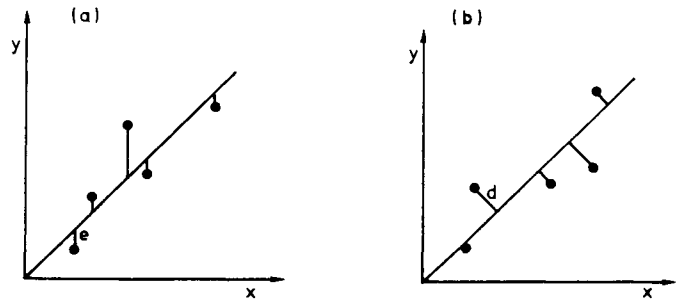


Fig. 8.15. (a) In the least squares method (LS) the residual, e_i, is obtained parallel to the y-axis. (b) In the orthogonal distance regression method (ODR) the residual, d_i, is obtained perpendicular to the estimated line.

distances from the data points to the regression line. The meaning of both e_i and d_i is compared in Fig. 8.15: while e_i in the classical least-squares method is obtained parallel to the y-axis, d_i is determined perpendicular to the estimated line.

The fitted line is then the one for which the least sum of squared d_is is obtained and the method has been called *orthogonal distance regression* [11]. This is equivalent to finding the first principal component of a data set consisting of 2 variables ($p = 2$) and n samples (see Chapters 17 and 31).

The expressions for b_1 and b_0 are [12,13]:

$$b_1 = \frac{s_y^2 - s_x^2 + \sqrt{(s_x^2 - s_y^2)^2 + 4(\mathrm{cov}(y,x))^2}}{2\mathrm{cov}(y,x)} \tag{8.48}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

with s_y^2 and s_x^2 the variance of the y variable and the x variable, respectively; $\mathrm{cov}(y,x) = (\sum(y_i - \bar{y})(x_i - \bar{x}))/(n-1)$ the covariance of y and x (see Section 8.3.1).

Mandel [13] gives expressions for the standard deviation of the slope and the intercept. An approximate relationship between the least squares slope, b_1(LS), and the orthogonal distance slope, b_1(ODR), has been formulated [13]:

$$b_1(\mathrm{ODR}) = b_1(\mathrm{LS}) \Big/ \left(1 - \frac{s_{ex}^2}{s_x^2}\right)$$

with s_{ex}^2 the variance of a single x value (involves replicate observations of the same x); and s_x^2 the variance of the x variable $= \sum(x_i - \bar{x})^2/(n-1)$. The latter depends on the range of the x_i and their distribution.

The ratio s_{ex}/s_x has been proposed by Cornbleet [14] as an estimate of the effect of errors in the x variable. Significant errors in the least squares estimate of b_1 can be expected if this ratio is large (for example >0.2). If this ratio is small, which means that the spread of the x values is large compared to the measurement error

of these x values, the latter can be ignored and the classical least squares analysis used.

This is also the case if the x_is are subject to error but are set at fixed values by the experimenter (= Berkson Model). For example, consider regressions in which different doses of a drug or different concentrations of an analyte preassigned by the experimenter are involved. Although the actual doses or concentrations added may differ from their target values, the ordinary least squares method may be applied. It can be shown [13] that in this situation the error term is independent of x_i and consequently the assumptions of ordinary least squares are fulfilled.

Example 11:

Let us consider the results of Example 5 (Table 8.6) in their original context, namely the comparison of a new potentiometric method (= y variable) with a reference flameless AAS method (= x variable) for the determination of Pb in fruit juices.

$n = 10$ $\quad \bar{x} = 160.5 \quad \bar{y} = 158.5$

$s_x = 98.4731 \quad s_y = 95.3954$

$\text{cov}(x,y) = 9342.5$

$s_x^2 = 9697.0 \quad s_y^2 = 9100.3$

Therefore, from eq. (8.48),

$$b_1 = \frac{9100.3 - 9697.0 + \sqrt{(9697.0 - 9100.3)^2 + 4(9342.5)^2}}{2 \times 9342.5}$$

$$= 0.9686$$

and

$$b_0 = 158.5 - 160.5 \times 0.9686 = 3.04$$

It can be verified that the ODR line is very similar to the LS line calculated in Example 5:

ODR line: $\hat{y} = 3.04 + 0.969\,x$

LS line: $\hat{y} = 3.87 + 0.963\,x$

This is due to the fact that for the comparison a large range of x values has been considered. Therefore s_x^2 is large compared with the spread of errors likely to occur in the x's (s_{ex}^2)

Different comparisons of the least-squares regression and the orthogonal distance regression for method comparison [14,15] have shown that, depending on the experimental design, least squares can lead to wrong estimates of the regression coefficients and consequently invalidates the conclusion concerning the usefulness

of the method tested. Other methods that take errors in both the x and y variables into account have also been described. Some of these are critically examined by MacTaggart [16].

8.2.12 Straight line regression through a fixed point

In some situations the fitted line may be constrained to pass through a fixed point (x_0,y_0). Since this point must lie on the straight line we have:

$$y_0 = b_0 + b_1 x_0$$

and consequently $\hat{y} = b_0 + b_1 x$ can be rewritten as:

$$\hat{y} = y_0 + b_1(x - x_0) \tag{8.49}$$

A model involving only one parameter b_1 is obtained. Minimization of

$$\Sigma e_i^2 = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - y_0 - b_1(x_i - x_0))^2$$

with respect to b_1 now leads to the following expression for b_1:

$$b_1 = \frac{\Sigma (x_i - x_0)\,(y_i - y_0)}{\Sigma (x_i - x_0)^2} \tag{8.50}$$

The residual variance s_e^2 which, as stated earlier, is an estimate of the pure experimental error σ^2 if the model is correct, is now given by:

$$s_e^2 = \frac{\Sigma (e_i)^2}{n-1} = \frac{\Sigma (y_i - \hat{y}_i)^2}{n-1} \tag{8.51}$$

Notice that we divide here by $n - 1$ since the residuals are obtained from a fitted line for which only one parameter, b_1, has to be estimated. If $x_0 = 0$ and $y_0 = 0$, which means that the regression line must pass through the origin, eq. (8.49) and eq. (8.50), respectively, simplify to:

$$\hat{y} = b_1 x$$

and

$$b_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

The standard deviations to be used in the calculation of confidence intervals are then as follows [8]:

- the standard deviation of the estimated slope:

$$s_{b_1} = s_e \sqrt{\frac{1}{\Sigma x_i^2}} \tag{8.52}$$

- the standard deviation of an estimated point of the true regression line at given value of x,x_0:

$$s_{\hat{y}_0} = s_e \, x_0 / \sqrt{\sum x_i^2} \tag{8.53}$$

- the standard deviation of a new mean response predicted from the regression line at a given value of x,x_0:

$$s_{\hat{y}_0} = s_e / \sqrt{1/m + x_0^2 / \sum x_i^2} \tag{8.54}$$

- the standard deviation of x_s predicted from y_s, the mean of m values of y.

$$s_{\hat{x}_0} = (s_e / b_1) \sqrt{1/m + y_s^2 / b_1^2 \sum x_i^2} \tag{8.55}$$

To calculate the confidence limits the appropriate t-value at $n-1$ (and not $n-2$) degrees of freedom should of course be used. Moreover as opposed to the unconstrained model (see Section 8.2.4.3) these confidence limits will be valid over the whole range of x values since only one parameter b_1 is estimated here.

It is necessary to use this model only if there are good *a priori* reasons to do so. For example it is not because the intercept is found not to be significantly different from zero in the unconstrained model that the model $\eta = \beta_1 x$ should be used.

8.2.13 Linearization of a curved line

When the relationship between two variables cannot be represented by a straight line, polynomial (see Chapter 10) or non-linear (see Chapter 11) regression methods should be applied. However, by transformation of one or both variables, some of these models can be converted into a simpler straight line relationship.

Well known linearizations are, for example:

- the transformation of the exponential relation between radioactivity and time

$$A_t = A_0 \, e^{-0.693t/t_{1/2}}$$

into

$$\log A_t = \log A_0 - \frac{0.301}{t_{1/2}} t$$

- the transformation of the Michaelis–Menten equation which defines the quantitative relationship between the initial rate of an enzyme reaction, v, and the substrate concentration [S]

$$v = \frac{V_{max}[S]}{K_m + [S]}$$

into

$$\frac{1}{v} = \frac{1}{V_{max}} + \frac{K_m}{V_{max}}\left[\frac{1}{[S]}\right]$$

where $1/V_{max}$ represents the intercept and K_m/V_{max} the slope of the straight line that gives the relationship between $1/v$ and $1/[S]$.

When the functional relationship between the variables is not known linearization becomes much more difficult. Mosteller and Tukey [17] have proposed a general rule to find an appropriate re-expression to straighten hollow upward (concave) or hollow downward (convex) curved lines by using a transformation of the form:

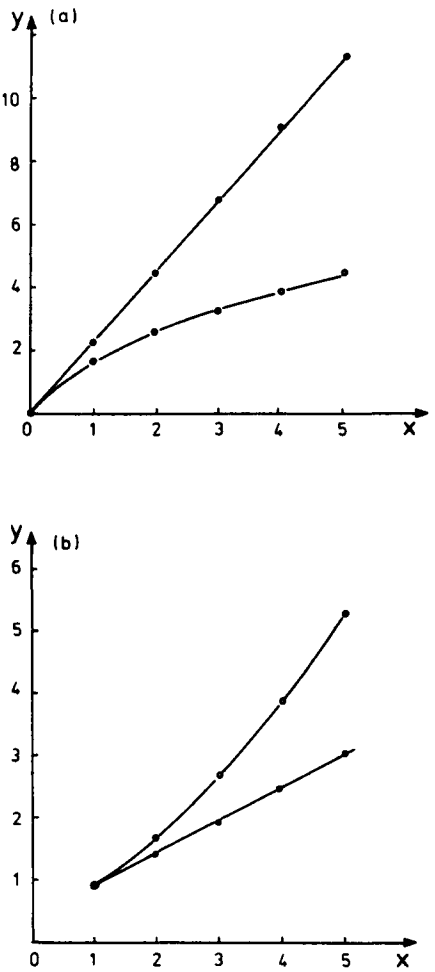


Fig. 8.16. (a) An example of a hollow downward curve linearized by the transformation $y^* = y^p$ with $p = 1.61$. (b) An example of a hollow upward curve linearized by the transformation $y^* = y^p$ with $p = 0.67$.

$y^* = y^p$

where y represents the original variable and y^* the transformed variable.

The value of p depends on the direction of the hollowness: for hollow downward curves that are either monotonically increasing or decreasing, p should be larger than 1 while with values <1 hollow upward curves can be linearized. Examples are shown in Fig. 8.16. The linearization procedure has been used by Wang [18] to linearize curved atomic absorption calibration lines. The p value is determined iteratively and the quality of each transformation is measured in terms of the residuals.

It is important to realize that the transformation is carried out to obtain a straight line relationship but that the condition of homoscedasticity might not be fulfilled for the transformed data. Kalantar [19] showed, with simulated data containing different error structures, the extent to which weighting can improve the precision of the estimated parameters of log linearized data.

8.3 Correlation

As pointed out in the introduction, correlation analysis is applied to study how strong the association between two random variables is. One variable is not expressed as a function of the other since both are equivalent. There is neither a dependent nor an independent variable.

Consider, for example, the data from biomedical analysis in Table 8.8. They represent Cu, Mn and Zn concentrations determined in 12 different structures of

TABLE 8.8

Concentration of Cu, Zn and Mn in different brain structures

Brain structure	$\mu g\ g^{-1}$ dry weight		
	Cu	Mn	Zn
1	25.8	1.03	78.0
2	24.2	0.96	81.8
3	27.3	1.05	69.4
4	32.8	1.49	76.1
5	27.3	1.84	62.5
6	17.9	1.23	60.1
7	14.0	1.09	34.2
8	13.3	0.96	35.5
9	10.0	0.80	33.3
10	10.9	0.77	38.9
11	10.7	0.80	40.8
12	16.0	1.10	46.4

$\bar{y}_{Cu} = 19.18$ $s_{Cu} = 7.89$
$\bar{y}_{Mn} = 1.09$ $s_{Mn} = 0.31$
$\bar{y}_{Zn} = 54.75$ $s_{Zn} = 18.59$

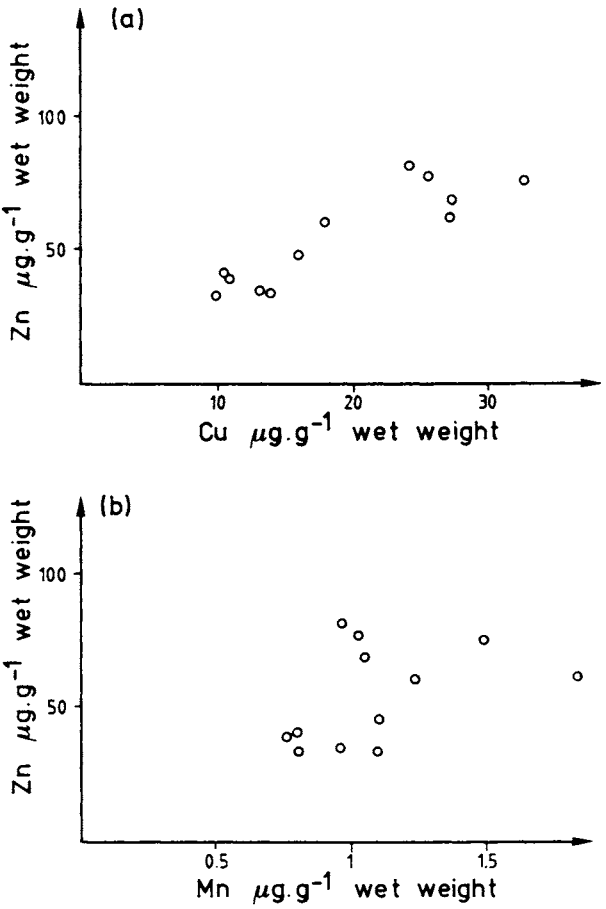


Fig. 8.17. Scatter diagrams of random variables with different degrees of association. Data from Table 8.8.

the human brain. The data for Cu and Zn, displayed graphically in a correlation or scatter diagram in Fig. 8.17a, indicate that high values of Cu are associated with high Zn concentrations, while low values of Cu are associated with low Zn concentrations. Knowledge of the concentration of Cu or Zn gives *a priori* information concerning the concentration of the other element. Both variables are related in a positive sense. They are positively correlated. Relationships can of course also be negative: the higher the fluoride intake (within certain limits) the lower the incidence of tooth caries in children. Much less association is found between the Zn and Mn concentrations in the brain structures of our example. As a result of the large scatter shown in Fig. 8.17b, knowledge of the concentration of one of these elements does not give *a priori* information concerning the concentration of the other element.

The strength of the linear relationship between a pair of variables is quantified by the *covariance* and the *correlation coefficient*. They are both measures of the joint variation between two random variables. Correlation also plays an important role in clustering (see Chapter 30) where it can be used as a measure of similarity.

8.3.1 The correlation coefficient

Consider y_{11}, y_{12}, y_{13}, ..., y_{1n} and y_{21}, y_{22}, y_{23}, ..., y_{2n} which are two sets of n corresponding measurements with respective means $\bar{y}_1$ and $\bar{y}_2$. The *covariance* of the variables y_1 and y_2 is given by:

$$\text{cov}(y_1,y_2) = \frac{1}{n-1}\Sigma(y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2) \tag{8.56}$$

It is a measure of the degree to which the two variables vary together. $\text{Cov}(y_1,y_2)$ is an estimate of the population covariance $\gamma(y_1,y_2)$:

$$\gamma(y_1,y_2) = \frac{1}{N}\Sigma(y_{1i} - \mu_1)\,(y_{2i} - \mu_2) \tag{8.57}$$

obtained with all possible observation pairs and the true population means μ_1 and μ_2 of the two sets.

For our examples in Fig. 8.17 the covariance of the Cu and Zn concentrations is:

$$\text{cov(Zn,Cu)} = \frac{\Sigma(\text{Cu}_i - 19.18)\,(\text{Zn}_i - 54.75)}{11} = \frac{1448.17}{11}$$

$$= 131.65$$

and the covariance of the Mn and Zn concentrations is

$$\text{cov(Mn,Zn)} = \frac{\Sigma(\text{Mn}_i - 1.09)\,(\text{Zn}_i - 54.75)}{11} = \frac{27.361}{11}$$

$$= 2.49$$

It should be noted that a covariance can take any value between $-\infty$ and $+\infty$. The covariance will be negative if the variables are negatively associated. In that case high values of y_1 are accompanied by low values of y_2 and *vice versa*. Consequently, when one of the deviations $(y_{1i} - \bar{y}_1)$ or $(y_{2i} - \bar{y}_2)$ in eq. (8.56) is positive, the other is negative and the sum of their products is negative. The main disadvantage of the use of the covariance as a measure of association between pairs of measurements is that it depends on the scale chosen. In our example the covariances are increased by a factor of 10^6 if concentrations are given in ng g^{-1} instead of μg g^{-1}!

A parameter which is independent of the measurement units used is obtained if the covariance is divided by the standard deviation of both sets of measurements. This quantity is the *product–moment correlation coefficient* or the *Pearson correlation coefficient, r*:

$$r(y_1 y_2) = \frac{\text{cov}\,(y_1 y_2)}{s_{y_1}\, s_{y_2}} = \frac{(\sum (y_{1i} - \bar{y}_1)\,(y_{2i} - \bar{y}_2))\,/n-1}{\sqrt{\dfrac{\sum (y_{1i} - \bar{y}_1)^2}{n-1} \dfrac{\sum (y_{2i} - \bar{y}_2)^2}{n-1}}}$$

$$= \frac{\sum (y_{1i} - \bar{y}_1)\,(y_{2i} - \bar{y}_2)}{\sqrt{\sum (y_{1i} - \bar{y}_1)^2 \sum (y_{2i} - \bar{y}_2)^2}} \tag{8.58}$$

This is an estimate of the population correlation coefficient $\rho(y_1,y_2)$. The correlation coefficient is a dimensionless number between –1 and +1. Values of –1 or +1 indicate a perfect linear relationship between the two variables. A correlation coefficient which is not significantly different from zero indicates that the variables are uncorrelated. This does not imply that there is no relationship between the variables but only indicates that there is no *linear* relationship.

For our example the correlation coefficients between Cu and Zn and between Mn and Zn are respectively:

r(Cu,Zn) = 131.65/(7.89 × 18.59) = 0.898
r(Zn,Mn) = 2.49/(0.31 × 18.59) = 0.432

It follows from a comparison of the correlation coefficients that Cu and Zn are indeed more correlated than Zn and Mn.

The scatter plots in Fig. 8.18 illustrate how r behaves for data with a different degree of association. Compare Figs. 8.18d and 8.18e both with $r = 0$. In Fig. 8.18e there is an obvious relationship between the two variables but in this case, because the relationship is not linear, the correlation coefficient is zero.

The linear relationship between two random variables can be described by two regression lines. The regression of y_1 on y_2 is given by:

$$y_1 = b_0 + b_1\, y_2$$

and the regression of y_2 on y_1 by:

$$y_2 = b'_0 + b'_1\, y_1$$

Both lines go through the point $(\bar{y}_1,\bar{y}_2)$ which is their point of intersection. If there is a perfect correlation between the two variables ($r = +1$ or $r = -1$) the two regression lines coincide. In the case of no correlation, which means $r = 0$, the two regression coefficients, β_1 and β'_1 are also zero (see Section 8.3.3) and the two lines are perpendicular.

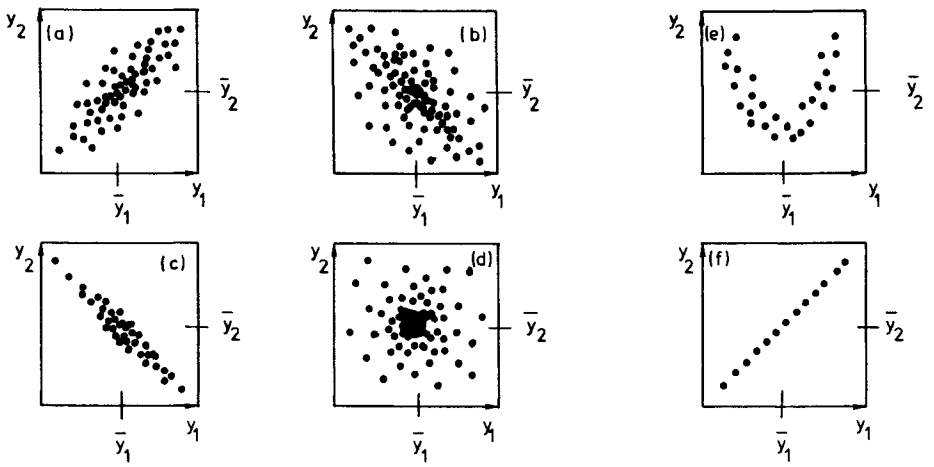


Fig. 8.18. Scatter plots of random variables with various degrees of correlation: (a) $r = 0.75$; (b) $r = -0.32$; (c) $r = -0.95$; (d) $r = 0$; (e) $r = 0$; (f) $r = 1$.

8.3.2 Hypothesis tests and confidence limits

Before discussing significance tests and confidence limits for the population regression coefficient, ρ, it is useful to have a closer look at the bivariate population from which the sample of observation pairs is drawn. A bivariate population provides the probability that the two variables jointly take particular values.

We assume this population to have a *bivariate normal distribution* for which the probability function is given by:

$$f(y_1 y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp$$

$$\left[-\frac{1}{2(1-\rho^2)} \left\{ \left(\frac{y_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{y_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{y_1 - \mu_1}{\sigma_1} \right) \left(\frac{y_2 - \mu_2}{\sigma_2} \right) \right\} \right] \qquad (8.59)$$

where μ_i and σ_i^2 are the mean and variance of y_i ($i = 1,2$) and ρ is the correlation coefficient. This function can be represented as a bell-shaped probability surface (Fig. 8.19) with the following properties:

(i) both variables y_1 and y_2 taken separately are normally distributed (one says that the *marginal distributions* are normal);

(ii) for a given y_1 the distribution of y_2 is normal and similarly for given y_2 the distribution of y_1 is normal (one says that the *conditional distributions* of y_1 and y_2 are normal). The latter is shown in Fig. 8.19: cross sections of the bell-shaped surface at any value of y_1 or y_2 yield Gauss curves;

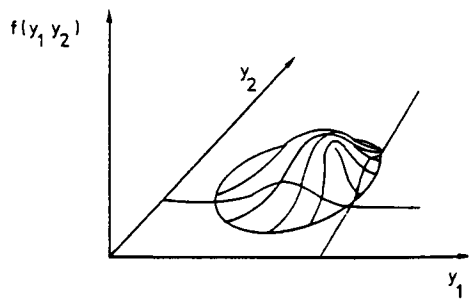


Fig. 8.19. Bivariate normal distribution surface.

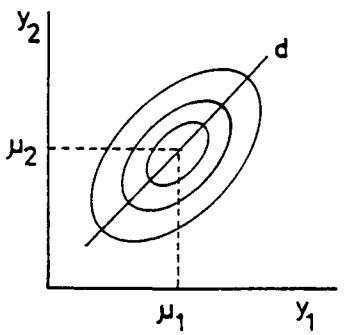


Fig. 8.20. Isoprobability ellipses for a bivariate normal distribution with $\sigma_1 = \sigma_2$, $\rho = 0.6$.

(iii) cross-sections with planes parallel to the (y_1,y_2) plane yield ellipses representing all y_1,y_2 combinations with the same probability density. They are called *isoprobability ellipses*. The bivariate normal distribution can be represented as a set of isoprobability ellipses (Fig. 8.20) with equations:

$$a = \frac{1}{1-\rho^2}\left\{\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right)\right\} \tag{8.60}$$

a is a positive constant; the smaller the constant the higher up the hill the cross section is performed. For $a = 5.99$ an ellipse is obtained containing 95% of the data points. The centre of the ellipses is the point with coordinates (μ_1,μ_2); the major axis of the ellipse (d) and the minor axis perpendicular to it, are common to all isoprobability ellipses. These axes correspond with principal components to be discussed in Chapters 17 and 31. The shape of the ellipses and their position in the (y_1,y_2) plane are determined by the values of σ_1,σ_2 and ρ. If the variables are not associated, ρ is zero and the axes of the ellipses are parallel to the co-ordinate axes. If, in addition, $\sigma_1 = \sigma_2$ the isoprobability contour is a circle. For $\rho = 1$, the ellipse as computed with eq. (8.60) is undefined and in fact it is found that all data points lie on a straight line.

Together with μ_1, μ_2 and σ_1 and σ_2, the correlation coefficient ρ is an important parameter of the bivariate normal distribution. The scatter plot of the data (as in Fig. 8.18) is a graphical representation of this distribution. The sample correlation coefficient r being an estimate of the population correlation coefficient ρ, inferences about ρ can be made from r. The most common null hypothesis to be tested is whether $\rho = 0$ (H_0: $\rho = 0$; H_1: $\rho \neq 0$). Accepting H_0 means that the two variables are uncorrelated or more precisely that a non-zero correlation has not been detected. One calculates t:

$$t = r\sqrt{n-2} \,/\, \sqrt{1-r^2} \tag{8.61}$$

and compares it with the tabulated value of t with $n - 2$ degrees of freedom.

For the correlation coefficients calculated in Section 8.3.1, this test yields the following results:

r(Cu,Zn) = 0.898 $\qquad t = 0.898\sqrt{10} \,/\sqrt{1-(0.898)^2}$

$$= 6.45 > t_{0.025,10} = 2.23$$

Therefore at the 5% significance level the correlation between Cu and Zn is significant. Also:

r(Mn,Zn) = 0.432 $\qquad t = 0.432\sqrt{10} \,/\sqrt{1-(0.432)^2}$

$$= 1.51 < t_{0.025,10} = 2.23$$

indicating that Mn and Zn are not significantly correlated.

Table 8.9 tabulates significance levels of r that allow direct inspection of the correlation coefficient. The 5% level of r for 10 degrees of freedom, which inserted in eq. (8.61) would yield significance, is 0.576. This means that to be significant at the 5% confidence level the correlation coefficient between twelve pairs of measurements should at least be 0.576.

The t-test of eq. (8.61) can be applied only to test H_0: $\rho = 0$, since for $\rho \neq 0$ the frequency distribution of r is not normal but is asymmetrical. Therefore to calculate the $100(1 - \alpha)\%$ confidence interval of ρ the correlation coefficient is transformed to a new variable.

$$z = 0.5 \ln[(1 + r)/(1 - r)] \tag{8.62}$$

This new variable is distributed almost normally with an expected standard deviation of approximately $\sigma_z = \sqrt{1/(n-3)}$.

The calculation of the 95% confidence interval of the correlation coefficient between Cu and Zn is performed as follows:

$$z = 0.5 \ln[(1 + 0.898)/(1 - 0.898)] = 1.462$$

$$\sigma_z = 1/\sqrt{9} = 0.333$$

TABLE 8.9
Critical levels of r (for $p = 0$). p is the two-sided significance level.

df	$p = 0.1$	$p = 0.05$	$p = 0.01$
1	0.988	0.997	1.000
2	0.900	0.950	0.990
3	0.805	0.878	0.959
4	0.729	0.811	0.917
5	0.669	0.754	0.875
6	0.621	0.707	0.834
7	0.582	0.666	0.798
8	0.549	0.632	0.765
9	0.521	0.602	0.735
10	0.497	0.576	0.708
11	0.476	0.553	0.684
12	0.457	0.532	0.661
13	0.441	0.514	0.641
14	0.426	0.497	0.623
15	0.412	0.482	0.606
16	0.400	0.468	0.590
17	0.389	0.456	0.575
18	0.378	0.444	0.561
19	0.369	0.433	0.549
20	0.360	0.423	0.537
21	0.352	0.413	0.526
22	0.344	0.404	0.515
23	0.337	0.396	0.505
24	0.330	0.388	0.496
25	0.323	0.381	0.487
26	0.317	0.374	0.478
27	0.311	0.367	0.470
28	0.306	0.361	0.463
29	0.301	0.355	0.456
30	0.296	0.349	0.449
40	0.257	0.304	0.393
50	0.231	0.273	0.354
60	0.211	0.250	0.325
80	0.183	0.217	0.283
100	0.164	0.195	0.254

The 95% confidence interval for z is therefore:

$$z \pm 1.96 \times 1 / \sqrt{9} = 1.462 \pm 1.96 \times 0.333 = 1.462 \pm 0.653$$

$$0.809 \leq z \leq 2.115$$

Retransforming these z-values to the corresponding r-values gives the 95% confidence limits for ρ:

$$0.669 \leq \rho(\text{Cu, Zn}) \leq 0.971$$

Notice that due to the skewed distribution of r this confidence interval is not symmetrical around $r = 0.898$. Tables that allow the transformation of r to z and the back transformation of z to r are available in most books on statistics.

In a similar way, the 95% confidence interval of the correlation coefficient between Mn and Zn is found to be:

$$-0.189 \leq \rho(\text{Mn, Zn}) \leq 0.806$$

This also leads to the conclusion that the correlation coefficient is not significantly different from zero since zero is included in this very large interval. Notice that the confidence interval is much more informative.

A two-sided test based on the z statistic defined above is used for the comparison of two correlation coefficients (H_0; $\rho_1 = \rho_2$, H_1: $\rho_1 \neq \rho_2$). Both are converted to z (eq. (8.62)) and the significance of the difference between the two z's is tested as follows:

$$t = \frac{z_1 - z_2}{\sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}} = \frac{z_1 - z_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}} \tag{8.63}$$

with n_1 and n_2 the sample size on which r_1 and r_2, respectively, are based. At the 5% significance level this quantity can be compared to 1.96 (z-value from the standardized normal distribution) since it has been obtained from a population standard deviation.

As an example, suppose that in our biomedical application another sample of 15 brain structures was analyzed for Cu and Zn. The correlation coefficient was found to be 0.703. We want to know whether both correlation coefficients $r_1 = 0.898$ ($n_1 = 12$) and $r_2 = 0.703$ ($n_2 = 15$) estimate the same parametric value of ρ. Since

$z_1 = 1.462$ and $z_2 = 0.873$ it follows that

$$t = \frac{1.462 - 0.873}{\sqrt{\dfrac{1}{9} + \dfrac{1}{12}}} = 1.34 < 1.96$$

Therefore, there is not enough evidence to conclude that both samples come from differently correlated populations.

8.3.3 Correlation and regression

Although correlation and regression analysis are used for different purposes there are obvious mathematical relationships between the correlation coefficient and the regression coefficients.

In Section 8.2.1 (eq. (8.4)) the expression for the slope of the least-squares line through n data points, b_1 was derived. From eq. (8.58) the correlation coefficient between x and y, $r(x,y)$, can be obtained. If b_1 is divided by $r(x,y)$ we obtain:

$$\frac{b_1}{r(x,y)} = \frac{\sqrt{\sum (x_i - \bar{x})^2}\ \sqrt{\sum (y_i - \bar{y})^2}}{\sum (x_i - \bar{x})^2} = \frac{\sqrt{\sum (y_i - \bar{y})^2}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Dividing both the numerator and the denominator by $\sqrt{n-1}$ yields the following relationship between the estimated slope and the estimated correlation coefficient:

$$\frac{b_1}{r(x,y)} = \frac{\sqrt{\sum (y_i - \bar{y})^2 / n - 1}}{\sqrt{\sum (x_i - \bar{x})^2 / n - 1}} = \frac{s_y}{s_x}$$

This expression which can be rewritten as

$$b_1 = r(x,y) \frac{s_y}{s_x} \tag{8.64}$$

which indicates that if either b_1 or r is zero the other is also zero (since neither s_y nor s_x are zero): if there is no correlation between x and y, a significant linear regression between these variables cannot exist. Therefore, the test for the significance of r (H_0: $\rho = 0$), described in Section 8.3.2, could also be answered by the test for the significance of b_1 (H_0: $\beta_1 = 0$), described in Section 8.2.4.1. Both are mathematically equivalent and the acceptance of $\beta_1 = 0$ implies acceptance of $\rho = 0$.

Since $b_0 = \bar{y} - b_1 \bar{x}$ (eq. (8.5)), substituting b_1 by eq. (8.64) yields:

$$b_0 = \bar{y} - r(x,y) \frac{s_y}{s_x} \bar{x}$$

Combining this with $y = b_0 + b_1 x$ gives:

$$y = \bar{y} - r(x,y) \frac{s_y}{s_x} \bar{x} + r(x,y) \frac{s_y}{s_x} x$$

and

$$(y - \bar{y}) = (x - \bar{x})\, r(x,y) \frac{s_y}{s_x} \tag{8.65}$$

This relationship will be applied in the discussion of autocorrelation and autoregression in Chapter 20.

Finally, let us consider the square of the correlation coefficient between x and y, r^2. In eq. (8.7) $(\sum (x_i - \bar{x})\ (y_i - \bar{y}))^2$ can be substituted by the equivalent expression $r^2 \sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2$ (see eq. (8.58)) which yields:

$$\Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - \bar{y})^2 - r^2\Sigma(y_i - \bar{y})^2$$
$$= (1 - r^2)\,\Sigma(y_i - \bar{y})^2$$

From this the following expression for r^2 is obtained:

$$r^2 = \frac{\Sigma\,(y_i - \bar{y})^2 - \Sigma\,(y_i - \hat{y}_i)^2}{\Sigma\,(y_i - \bar{y})^2} = \frac{\Sigma\,(\hat{y}_i - \bar{y})^2}{\Sigma\,(y_i - \bar{y})^2} = \frac{SS_{REG}}{SS_{TOT}} \tag{8.66}$$

The latter equality in eq. (8.66) follows from eq. (8.10). Therefore in regression analysis the square of the correlation coefficient between x and y, r^2 which is called the *coefficient of determination*, expresses the proportion of the total variation that is explained by the regression. If $r = 1$ or $r = -1$ all observations perfectly fit a straight line and consequently the total variation in y can be explained in terms of the regression line ($r^2 = 1$). If, on the other hand, $r = 0$ ($r^2 = 0$) there is no regression at all between x and y. The regression line which parallels the x axis ($b_1 = 0$) cannot explain any variation of y.

References

1. J.C. Miller and J.N. Miller, Statistics for Analytical Chemistry. Ellis Horwood, Chichester, 3rd ed., 1993, pp. 140 and 211.
2. G.S. Land, W.J. Leavens and B.C. Weatherley, Comparison of two Methods of Calibrating Linear HPLC Assays. E. Reid and I.D. Wilson (Editors), Bioanalytical Approaches for Drugs, including Anti-asthmatics and Metabolites — Methodological Surveys in Biochemistry and Analysis, Volume 22. Royal Society of Chemistry, Cambridge, 1992, pp. 103–110.
3. G. Kornblum and L. de Galan, Personal communication.
4. H.M. Hill, A.G. Causery, D. Lessard, K. Selinger and J. Herman, Choice and Optimization of Calibration Functions. E. Reid and I.D. Wilson (Editors), Bioanalytical Approaches for Drugs, including Anti-asthmatics and Metabolites — Methodological Surveys in Biochemistry and Analysis, Volume 22. Royal Society of Chemistry, Cambridge, 1992, pp. 111–118.
5. M. Davidian and P.D. Haaland, Regression and calibration with nonconstant error variance. Chemometr. Intell. Lab. Systems, 9 (1990) 231–248.
6. N.R. Draper and H. Smith, Applied Regression Analysis. Wiley, New York, 1981.
7. S. Mannino, Determination of lead in fruit juices and soft drinks by potentiometric stripping analysis. Analyst, 107 (1982) 1466–1470.
8. P.D. Lark, B.R. Crowen and R.L.L. Bosworth, The Handling of Chemical Data. Pergamon Press, Oxford, 1968.
9. P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection. Wiley, New York, 1987.
10. C. Liteanu and L. Rica, Statistical Theory and Methodology of Trace Analysis. Ellis Horwood, Chichester, 1980.
11. P.T. Boggs, C.H. Spiegelman, J.R. Donaldson and R.B. Schnabel, A computational examination of orthogonal distance regression. J. Econometrics, 38 (1988) 169–201.
12. W.E. Deming, Statistical Adjustment of Data. Wiley, New York, 1943.

13. J. Mandel, The Statistical Analysis of Experimental Data. Dover Publications, New York, 1964.
14. P.J. Cornbleet and N. Gochman, Incorrect least-squares regression in method-comparison analysis. Clin. Chem., 25 (1979) 432–438.
15. C. Hartmann, J. Smeyers-Verbeke and D.L. Massart, Problems in method-comparison studies. Analusis, 21 (1993) 125–132.
16. D.L. MacTaggart and S.O. Farwell, Analytical use of linear regression. Part II: Statistical error in both variables. J. AOAC Int., 75 (1992) 608–614.
17. F. Mosteller and J.W. Tukey, Data Analysis and Regression. Addison-Wesley, 1977, pp. 84–87.
18. X. Wang, J. Smeyers-Verbeke and D.L. Massart, Linearization of atomic absorption calibration curves. Analusis, 20 (1992) 209–215.
19. A.H. Kalantar, Large inefficiencies of unweighted least-squares treatment of logarithmically transformed A exp(–kt) data. Int. J. Chem. Kinetics, 19 (1987) 923–927.

Additional recommended reading

Book

R.R. Sokal and J. Rohlf, Biometry — The Principles and Practice of Statistics in Biological Research. W.H. Freeman, New York, 1981.

Articles

Analytical Methods Committee, Uses (proper and improper) of correlation coefficients. Analyst, 113 (1988) 1469–1471.

F.J. Anscombe, Graphs in statistical analysis. Am. Statistician, 27 (1973) 17–22.

R.J. Carroll and C.H. Spiegelman, The effect of ignoring small measurement errors in precision instrument calibration. J. Qual. Technol., 18 (1986) 170–173.

D.L. MacTaggart and S.O. Farwell, Analytical use of linear regression. Part I: Regression procedures for calibration and quantification. J. AOAC Int., 75 (1992) 594–608.

J. Riu and F.X. Rius, Univariate regression models with errors in both axes. J. Chemometr., 9 (1995) 343–362.

Chapter 9

Vectors And Matrices

9.1 The data table as data matrix

Table 9.1 gives the results of the determination of Al, Si, Mn and Fe in five minerals. This collection of data can be considered as a *matrix* (see further Section 9.3). It consists of sub-sets of data for the different metals (column-wise) and for the objects (row-wise). These sub-sets are called *vectors* (see further Section 9.2). The *data matrix* **X** is then given by:

$$\mathbf{X} = \begin{bmatrix} 200 & 300 & 100 & 360 \\ 380 & 580 & 420 & 840 \\ 200 & 320 & 400 & 380 \\ 500 & 760 & 250 & 1060 \\ 50 & 70 & 25 & 100 \end{bmatrix} \tag{9.1}$$

The use of matrices is particularly useful in multivariate analysis to simplify the notations that can otherwise become quite complex. Here we introduce only some elementary concepts which are essential to understand Chapters 10 and 11 on multivariate and non-linear regression and Chapter 17 on principal components. Some definitions and the rules relative to some simple vector and matrix operations will be given. A more systematic introduction is given in Part B (Chapter 29), because a deeper understanding of linear algebra is needed for many chapters of that volume.

TABLE 9.1

Concentration of Al, Si, Mn and Fe in 5 minerals (arbitrary measurement units)

Object	Al	Si	Mn	Fe
1	200	300	100	360
2	380	580	420	840
3	200	320	400	380
4	500	760	250	1060
5	50	70	25	100

9.2 Vectors

9.2.1 Definitions

The matrix shown in eq. (9.1) can be considered as a collection of 5 rows, representing the results for the 5 objects. These rows are called *row vectors*. For instance $\mathbf{r}_1$ (the vector for object 1) is given by:

$$\mathbf{r}_1 = [200 \quad 300 \quad 100 \quad 360]$$

Similarly, we can view the data matrix as a collection of 4 columns, each representing the results for one of the variables. This is called a *column vector*. The column vector for the Al results, $\mathbf{c}_1$, is given by:

$$\mathbf{c}_1 = \begin{bmatrix} 200 \\ 380 \\ 200 \\ 500 \\ 50 \end{bmatrix}$$

When we use the word *vector* without specifying that it is a column or row vector, then, by convention, it is a column vector. This is written as:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

The results x_1 to x_n are called the *elements* of the vector. If there are p vectors, the jth one is written as:

$$\mathbf{x}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ . \\ . \\ . \\ x_{nj} \end{bmatrix}$$

In this convention the row vector is considered to be the transpose of a column vector. Consider the following vector:

$$\mathbf{x} = \begin{bmatrix} 10 \\ 20 \\ 30 \end{bmatrix}$$

then its *transpose*, $\mathbf{x}^T$, is written as:

$\mathbf{x}^T = [10 \quad 20 \quad 30]$

In Section 9.2 there is no need to follow this convention yet and therefore we will not.

A vector also has a geometrical meaning. It can be defined as a directed line segment. Consider first the results for Mn and Fe only: there are 5 row vectors each consisting of two elements and each representing one of the 5 objects. We can plot the results of Mn against those of Fe. In Fig. 9.1a this is done for object 1 and in

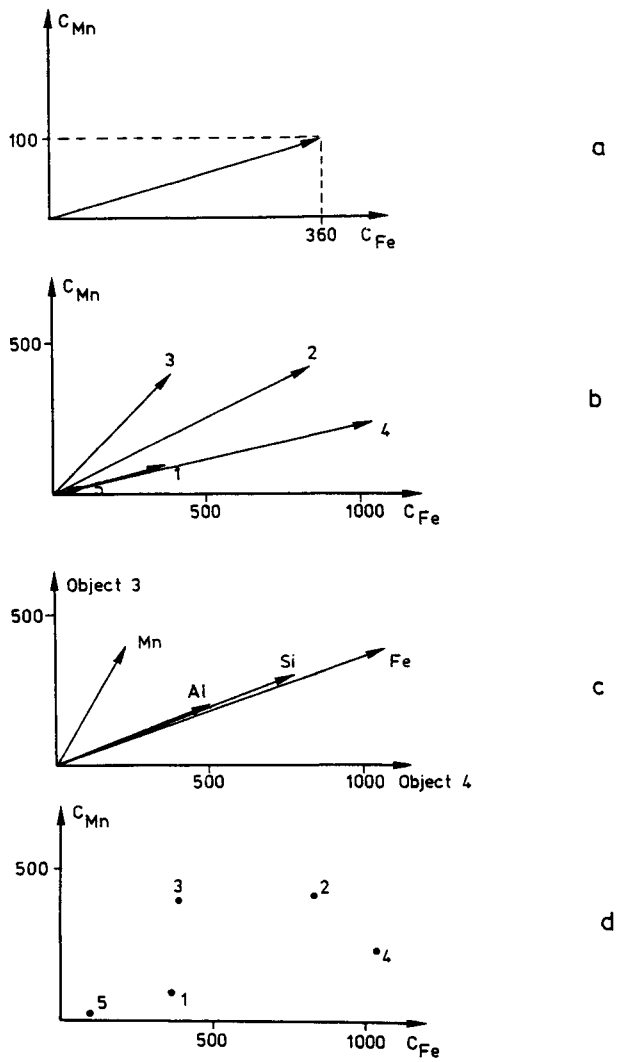


Fig. 9.1. Geometrical meaning of vectors. (a) Vector describing object 1 of Table 9.1 by its concentrations for Mn and Fe; (b) all five objects in the column space defined by variables Mn and Fe; (c) the variables in the row space defined by objects 3 and 4; (d) an alternative representation of (b).

Fig. 9.1b for the 5 objects. The arrow representation illustrates the fact that a vector should be viewed as a directed line segment. It is however more usual not to draw the arrow, but only the point in which the arrow ends. This then leads to the more usual scatterplot of Fig. 9.1d which contains the same information as Fig. 9.1b. Clearly, we can view the 5 objects of Fig. 9.1b as 5 row vectors in two-dimensional space. The axes are defined by the variables (here the concentrations of Fe and Mn). Since there is a column for each variable, we could say that the axes are associated with the columns and therefore the objects or rows are said to be represented in *column space* or *variable space*. In the same way, it would have been possible to represent the variables or columns in *row space* or *object space*. For two rows, this can be represented visually as in Fig. 9.1c. where the four variables are plotted in function of their values for objects 3 and 4.

The column and row spaces represented in Fig. 9.1 are two-dimensional. The vectors consist of two elements. We can therefore state that the *dimension* of a vector is equal to the number of elements it contains. The row space for the complete set of vectors of Section 9.1 is therefore five-dimensional. This is symbolised as an S^5 space.

Vectors and matrices, since they are collections of vectors, allow us to represent data sets or tables in multivariate space (objects in p-variate column space and variables in n-variate row space). This duality of representing the same data in two different spaces is less important in Part A, but will become very important when studying subjects such as multivariate modelling and calibration.

9.2.2 Operations on vectors

9.2.2.1 Addition of vectors

Vector addition is possible for vectors of the same dimension. The elements of the resulting vector are the sums of the corresponding elements of the summed vectors.

$$\mathbf{x}+\mathbf{y}=\begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}+\begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}=\begin{bmatrix} x_1+y_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n+y_n \end{bmatrix}$$

Applying this for a two-dimensional example

$$\mathbf{x}=\begin{bmatrix} 2 \\ 1 \end{bmatrix};\quad \mathbf{y}=\begin{bmatrix} 2 \\ 3 \end{bmatrix};\quad \mathbf{x}+\mathbf{y}=\begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

When representing this graphically, we observe that vector addition is equivalent to placing the added vectors head to tail (Fig. 9.2a).

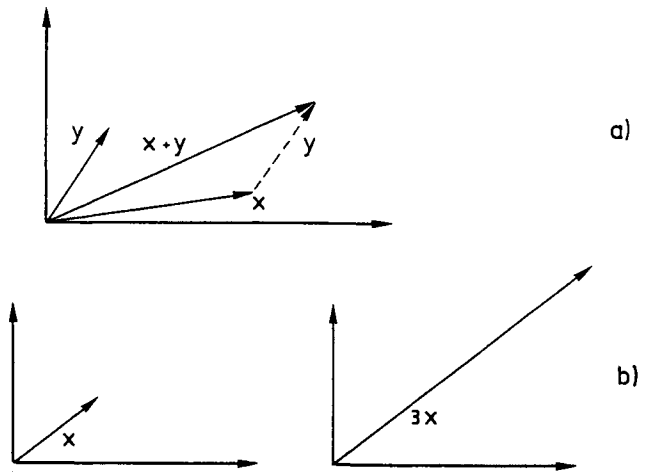


Fig. 9.2. (a) Vector addition. (b) Scalar multiplication.

Vector addition is *commutative*:

$$\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$$

and *associative*:

$$\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$$

9.2.2.2 Multiplication by a scalar

A *scalar* is a single number and a *vector* can therefore also be defined as an ordered array of scalars. Multiplication of a vector **x** by a scalar c (*scalar multiplication*) yields a new vector, the elements of which are obtained by multiplication of the elements of **x** by c

$$c\mathbf{x} = \begin{bmatrix} cx_1 \\ \cdot \\ \cdot \\ \cdot \\ cx_n \end{bmatrix}$$

With a two-dimensional example

$$\mathbf{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}; \quad c = 3; \quad c\mathbf{x} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}$$

we observe (Fig. 9.2b) that scalar multiplication consists of stretching **x** by a factor c.

Scalar multiplication is distributive with respect to addition:

$c(\mathbf{x} + \mathbf{y}) = c\mathbf{x} + c\mathbf{y}$

9.2.2.3 Vector multiplication

There are two types of vector multiplication. The one which is of interest to us yields what is called the *inner product*, *dot product*, or *scalar product*.

The inner product is obtained from vectors that have the same dimensions and consists of the sum of products of the corresponding elements. Thus, if one multiplies two row vectors consisting of the results for objects 1 and 2, one writes

$\mathbf{x} = [200 \quad 300 \quad 100 \quad 360]$

$\mathbf{y} = [380 \quad 580 \quad 420 \quad 840]$

$\mathbf{x} \cdot \mathbf{y} = 200.380 + 300.580 + 100.420 + 360.840 = 594400$

This way of writing the product explains why it is called the dot product. There is a second way which is consistent with the view that a vector is a 1-column or 1-row matrix. Matrix multiplication will be explained in Section 9.3. It will be seen there that elements of the matrix are obtained by summing the products of the ith element of a row with the ith element of a column. The product is then written as

$$\mathbf{x}\mathbf{y}^{\mathrm{T}} = [200 \quad 300 \quad 100 \quad 360] \begin{bmatrix} 380 \\ 580 \\ 420 \\ 840 \end{bmatrix}$$

The name scalar product is due to the fact that the result is a single value, and therefore a scalar. Later, when we have learned how to norm vectors, we will see that the scalar product is related to the angle between the vectors.

9.2.3 Length and distance

Consider a point $\mathbf{x}_1$ in two dimensions (see Fig. 9.3a). We can represent it as a vector.

$\mathbf{x}_1 = [x_{11} \; x_{12}]$

where x_{11} and x_{12} are the values for object 1 on variables x_1 and x_2.

The square distance from the origin to the point $\mathbf{x}_1$, $\|\mathbf{x}_1\|^2$, can be derived using the properties of the triangle to be:

$\|\mathbf{x}_1\|^2 = x_{11}^2 + x_{12}^2$

This result would also be obtained as the scalar product of $\mathbf{x}_1$ with itself.

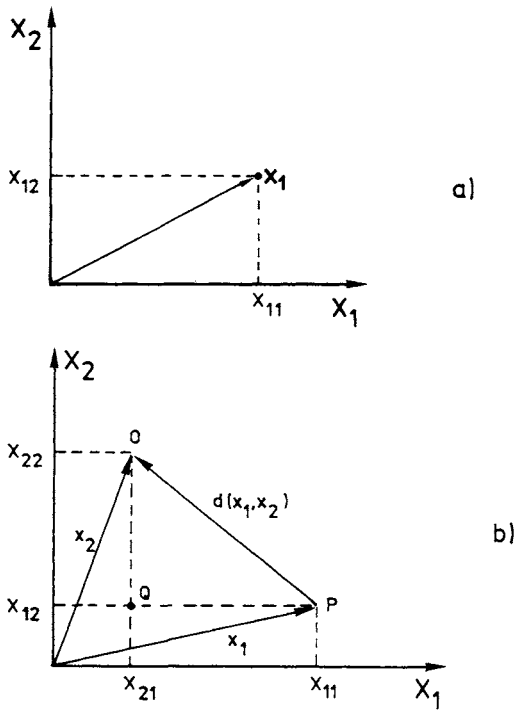


Fig. 9.3. (a) Distance of $\mathbf{x}_1$ from the origin. (b) The distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by $d(\mathbf{x}_1,\mathbf{x}_2)$. The value of $d(\mathbf{x}_1,\mathbf{x}_2)$ is obtained by considering that it is the hypothenusa of triangle OQP.

$$\mathbf{x}_1 \cdot \mathbf{x}_1 = \mathbf{x}_1 \mathbf{x}_1^T = x_{11}^2 + x_{12}^2$$

The non-negative square root, symbolized as $\|\mathbf{x}_1\|$ is called the *length* or the *norm* of the vector and is equal to:

$$\|\mathbf{x}_1\| = \sqrt{x_{11}^2 + x_{12}^2} \tag{9.2}$$

Note that we introduced the length of a vector as a distance and, in fact, the distance between the origin and the point of the arrow is called the *Euclidean distance* from the origin. Generalizing to n dimensions the length of a vector $\mathbf{x}$ is:

$$\|\mathbf{x}_1\| = \sqrt{x_1^2 + x_2^2 + \ldots x_n^2} \tag{9.3}$$

and

$$\|\mathbf{x}_1\| = \sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1} = \sqrt{\mathbf{x}_1 \mathbf{x}_1^T} \tag{9.4}$$

Let us calculate as an example the length of the vectors for objects 1 and 2 of Table 9.1. They are equal to

$$\|\mathbf{x}_1\| = \sqrt{200^2 + 300^2 + 100^2 + 360^2} = 519$$

$$\|\mathbf{x}_2\| = \sqrt{380^2 + 580^2 + 420^2 + 840^2} = 1167$$

Let us now go back to two dimensions and ask the question what the distance is between (points or vectors) $\mathbf{x}_1$ and $\mathbf{x}_2$ (Fig. 9.3b). From the properties of the triangle OPQ, it is easy to derive the distance $d(\mathbf{x}_1,\mathbf{x}_2)$ as being given by

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2} \tag{9.5}$$

This can be interpreted as the length of the vector starting in $\mathbf{x}_1$ and going to $\mathbf{x}_2$, or, in other words, the length of the new vector obtained by shifting the origin to $\mathbf{x}_1$. This distance is also called the Euclidean distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ and eq. (9.5) can also be generalized to n dimensions.

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \ldots + (x_{1n} - x_{2n})^2}$$

$$= \sqrt{\sum_{i}^{n} (x_{1i} - x_{2i})^2} \tag{9.6}$$

The distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ in the example is equal to

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(380 - 200)^2 + (580 - 300)^2 + (420 - 100)^2 + (840 - 360)^2} = 666$$

There is an interesting parallel between the length of a vector and the standard deviation. To understand this let us consider a 3-dimensional vector:

$$\mathbf{y} = \begin{bmatrix} 5 \\ 1 \\ 3 \end{bmatrix}$$

We *centre* or *mean-centre* this column vector by subtracting the mean of the three numbers.

$$\text{Calling } \bar{\mathbf{y}} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

$$\mathbf{y}^* = \mathbf{y} - \bar{\mathbf{y}} = \begin{bmatrix} 5 \\ 1 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \\ 0 \end{bmatrix}$$

The length of the new vector $\mathbf{y}^*$ is given by:

$$\|\mathbf{y}^*\| = \sqrt{(5 - 3)^2 + (1 - 3)^2 + (3 - 3)^2} = 2.83$$

If we had divided the sum of squares under the root by $n - 1 = 2$, $\mathbf{y}^*$ would have been the standard deviation of $\mathbf{y}$. Therefore, we can conclude that the length of a centred vector is proportional to the standard deviation of its elements.

9.2.4 Normed vectors

A *normed* or *normalised vector* is equal to the vector divided by its norm. Since the norm is a scalar, the elements of a normed vector are the original values each divided by the norm. Because the norm is related to the standard deviation, we can view the normed variables as related to standardized variables. Let us apply this again to the first two row vectors for the objects of Table 9.1.

$\mathbf{u} = \mathbf{x} / \|\mathbf{x}\|$

$\mathbf{v} = \mathbf{y} / \|\mathbf{y}\|$

$\mathbf{u} = [200 \quad 300 \quad 100 \quad 360] / 519$

$= [0.3854 \quad 0.5780 \quad 0.1927 \quad 0.6936]$

$\mathbf{v} = [0.3256 \quad 0.4970 \quad 0.3599 \quad 0.7198]$

An inherent property is that a normed vector has length 1. One can verify that

$0.3854^2 + 0.5780^2 + 0.1927^2 + 0.6936^2 = 1$

The sum of squared elements for such a vector is equal to 1. This is not the case for a vector consisting of standardised data where the sum of elements equals $n - 1$.

All the absolute values of the elements of the normed vector are $< |1|$. In fact, we can show that the elements are the cosines of the angle with the coordinate axes. This can be demonstrated with a simple example

$\mathbf{x} = [1 \quad 1]$

$\|\mathbf{x}\| = \sqrt{1^2 + 1^2} = \sqrt{2}$

$\mathbf{u} = \mathbf{x}/\|\mathbf{x}\| = [1/\sqrt{2} \quad 1/\sqrt{2}] = [\cos 45° \quad \cos 45°]$

For this reason the elements of a normed vector are also called *direction cosines.*

9.2.5 Angle between vectors

The angle θ between column vectors $\mathbf{x}$ and $\mathbf{y}$ is given by

$$\cos\theta = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \tag{9.7}$$

If $\mathbf{u}$ and $\mathbf{v}$ are the normed vectors of $\mathbf{x}$ and $\mathbf{y}$, this is equivalent to writing

$$\cos\theta = \mathbf{u}^T\mathbf{v} = \mathbf{u}\cdot\mathbf{v} \tag{9.8}$$

In other words, the cosine of the angle of two vectors is equal to the scalar product between the normed vectors. Since $\cos 90° = 0$, it follows that vectors are *orthogonal* if

$$\mathbf{u} \cdot \mathbf{v} = 0 \tag{9.9}$$

When the angle between two vectors is zero, then $\cos \theta = 1$. The angle between two vectors is zero when the corresponding elements are equal, except for a proportionality constant. We would say that two such series of numbers show a correlation coefficient = 1. This leads us to suspect that there must be a relationship between the angle between two vectors and the correlation coefficient of the two arrays of numbers, treated as vectors.

Rewriting eq. (8.58) for the correlation coefficient in terms of $r(\mathbf{y}, \mathbf{x})$, we obtain:

$$r(\mathbf{y},\mathbf{x}) = \frac{(\mathbf{y} - \bar{\mathbf{y}}) \cdot (\mathbf{x} - \bar{\mathbf{x}})}{\|\mathbf{y} - \bar{\mathbf{y}}\| \, \|\mathbf{x} - \bar{\mathbf{x}}\|} \tag{9.10}$$

If $\mathbf{y}' = \mathbf{y} - \bar{\mathbf{y}}$ and $\mathbf{x}' = \mathbf{x} - \bar{\mathbf{x}}$ are the mean centred vectors of $\mathbf{y}$ and $\mathbf{x}$, then

$$r(\mathbf{y},\mathbf{x}) = \frac{\mathbf{y}' \cdot \mathbf{x}'}{\|\mathbf{y}'\| \, \|\mathbf{x}'\|} = \cos \theta'$$

where θ' is the angle between the mean centred vectors. It follows that the correlation coefficient of two sets of numbers is equal to the scalar product of the normed mean-centred vectors or to the cosine of the angle between the mean-centred vectors. One can also show that the covariance of two variables is equal to the scalar product of the two mean-centred vectors divided by $n - 1$. In Section 9.2.4 we have already seen that standard deviation and therefore variance are related to the length of a vector. Since this relationship between statistical and vector concepts is very important for our further understanding of the methods of multivariate data analysis, we will describe it in more detail later (Chapter 29).

9.2.6 Orthogonal projection

The orthogonal projection $\mathbf{u}$ of a vector $\mathbf{x}$ on another vector $\mathbf{y}$ is shown in Fig. 9.4. It is a vector with the same direction as $\mathbf{y}$ multiplied by $\cos \theta$ and a scalar. It can be shown that this scalar is $\|\mathbf{x}\| / \|\mathbf{y}\|$, so that:

$$\mathbf{u} = \mathbf{proj}\ \mathbf{x} = \|\mathbf{x}\| \, \frac{\mathbf{y}}{\|\mathbf{y}\|} \cos \theta \tag{9.11}$$

or by replacing $\cos \theta$ by its value in eq. (9.7)

$$\mathbf{proj}\ \mathbf{x} = (\mathbf{x} \cdot \mathbf{y})\ \mathbf{y} / \|\mathbf{y}\|^2 \tag{9.12}$$

The length of the projection is given by:

$$\|\mathbf{proj}\ \mathbf{x}\| = (\mathbf{x} \cdot \mathbf{y}) / \|\mathbf{y}\| \tag{9.13}$$

Consider the row vectors $\mathbf{x} = [2, -1, 3]$ and $\mathbf{y} = [4, -1, 2]$. Then:

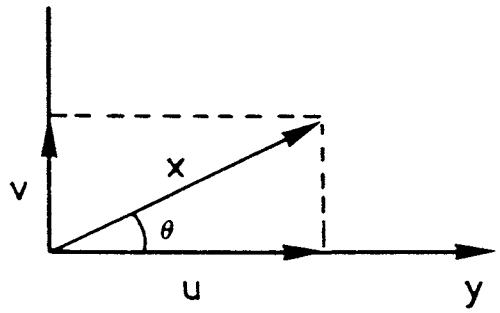


Fig. 9.4. Orthogonalization. Vector **x** is decomposed in two orthogonal vectors **u** and **v**; **u** is the orthogonal projection of **x** on **y** and **v** is the vector orthogonal to **y**.

$$\mathbf{x}\cdot\mathbf{y} = 2.4 + (-1)\,(-1) + 3.2 = 15$$

$$\|\mathbf{y}\|^2 = 4^2 + (-1)^2 + 2^2 = 21$$

$$\mathbf{proj}\ \mathbf{x} = \frac{15}{21}[4\ \ -1\ \ 2] = \left[\frac{60}{21}\ \ -\frac{15}{21}\ \ \frac{30}{21}\right]$$

9.2.7 Orthogonalization

In the preceding section, we saw how to project **x** on **y**. In Fig. 9.4 we have called this projection **u**. We would now also like to obtain **v**, the projection of **x** orthogonal to **y**. Since

$$\mathbf{u} + \mathbf{v} = \mathbf{x}$$

it follows from eq. (9.12) that:

$$\mathbf{v} = \mathbf{x} - (\mathbf{x}\cdot\mathbf{y})\mathbf{y} / \|\mathbf{y}\|^2 \tag{9.14}$$

u and **v** are orthogonal, which means that $\mathbf{u}\cdot\mathbf{v} = 0$.

For the vectors **x** and **y** given in the preceding section

$$\mathbf{u} = \left[\frac{60}{21}\ \ -\frac{15}{21}\ \ \frac{30}{21}\right]$$

and

$$\mathbf{v} = [2\ \ -1\ \ 3] - \left[\frac{60}{21}\ \ -\frac{15}{21}\ \ \frac{30}{21}\right]$$

$$= \left[-\frac{18}{21}\ \ -\frac{6}{21}\ \ \frac{33}{21}\right]$$

We can verify that

$$\mathbf{u}\cdot\mathbf{v} = \frac{60}{21}\cdot\frac{-18}{21} + \frac{-15}{21}\cdot\frac{-6}{21} + \frac{30}{21}\cdot\frac{33}{21} = 0$$

The procedure described can be generalized to more than two vectors. Suppose that one has a set of vectors ($\mathbf{x}_1,\mathbf{x}_2, ..., \mathbf{x}_k$). How can we obtain from this set a set of k orthogonal vectors ($\mathbf{v}_1,\mathbf{v}_2, ..., \mathbf{v}_k$) describing the same space? It can be shown that this can be done by the so-called Gram–Schmidt orthogonalization.

$$\mathbf{v}_1 = \mathbf{x}_1$$

$$\mathbf{v}_2 = \mathbf{x}_2 - (\mathbf{x}_2\cdot\mathbf{v}_1)\mathbf{v}_1 / \|\mathbf{v}_1\|^2$$

$$\mathbf{v}_3 = \mathbf{x}_3 - (\mathbf{x}_3\cdot\mathbf{v}_1)\mathbf{v}_1 / \|\mathbf{v}_1\|^2 - (\mathbf{x}_3\cdot\mathbf{v}_2)\mathbf{v}_2 / \|\mathbf{v}_2\|^2$$

$$\mathbf{v}_k = \mathbf{x}_k - (\mathbf{x}_k\cdot\mathbf{v}_1)\mathbf{v}_1 / \|\mathbf{v}_1\|^2 - (\mathbf{x}_k\cdot\mathbf{v}_2)\mathbf{v}_2 / \|\mathbf{v}_2\|^2 - ... - (\mathbf{x}_k\cdot\mathbf{v}_{k-1})\mathbf{v}_{k-1} / \|\mathbf{v}_{k-1}\|^2 \qquad (9.15)$$

In other words, $\mathbf{v}_i$ is the difference between $\mathbf{x}_i$ and the sum of the projections of $\mathbf{x}_i$ on the vectors $\mathbf{v}_1,\mathbf{v}_2...\mathbf{v}_{i-1}$. Decomposing a set of vectors in a set of orthogonal vectors is an important operation in data handling. The *Gram–Schmidt orthogonalization* is the simplest. Further orthogonalizations under additional constraints such as principal components are described in Chapter 17 and are used to a large extent in many of the chapters in Part B.

The computation of $\mathbf{v}$ is useful, for instance in the detection of minor substances below a chromatographic peak in HPLC with a diode array detector. This instrument is used to measure UV spectra in the eluate. At each time t_i we obtain a set of absorbances at different wavelengths. This set is the spectrum and can be viewed as a row vector (see Fig. 9.5). To detect the impurity we compare each spectrum or vector to the same row vector, called base vector. Let us consider as base vector the spectrum obtained at the top of the chromatographic peak. We call this row vector $\mathbf{x}_t = [x_{t1},...x_{tm}]$, where m is the number of wavelengths at which measurements are made. At each other time t_i a vector $\mathbf{x}_i = [x_{i1},...x_{im}]$ is obtained and we

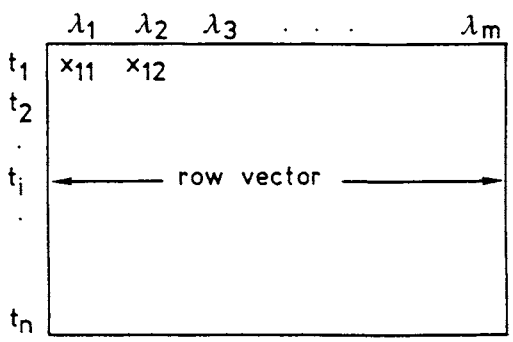


Fig. 9.5. Data table obtained in HPLC-DAD. The spectra can be represented by row vectors.

want to compare vector $\mathbf{x}_i$ to $\mathbf{x}_t$. There are many ways in which we could do that. We could, for instance, centre and normalize both $\mathbf{x}_t$ and $\mathbf{x}_i$ and obtain their product. This would yield the correlation coefficient r_i between $\mathbf{x}_t$ and each $\mathbf{x}_i$. We could then observe r_i in function of t_i. In regions where the pure product is present this would yield a higher r_i than in regions where the impurity influences the spectrum. It is also possible to determine the length of the orthogonal projection of $\mathbf{x}_i$ on $\mathbf{x}_t$, as was shown by Cuesta et al. [1]. These authors proposed several variants of methods where this length is measured. We will consider one of them (not necessarily the best one, but the easiest to explain).

In Fig. 9.6 the different steps are illustrated for the simplest possible example, namely two wavelengths (e.g. $\lambda_1 = 244$ nm and $\lambda_2 = 280$ nm). The concentration of the impurity is about half that of the main compound (Fig. 9.6a) and the substances are rather well separated. It should be understood that for such a good separation, a method like this one makes no practical sense: the intention is didactical. In practice, this method is used to detect poorly resolved minor peaks, often with spectra that do not differ much from that of the major compound.

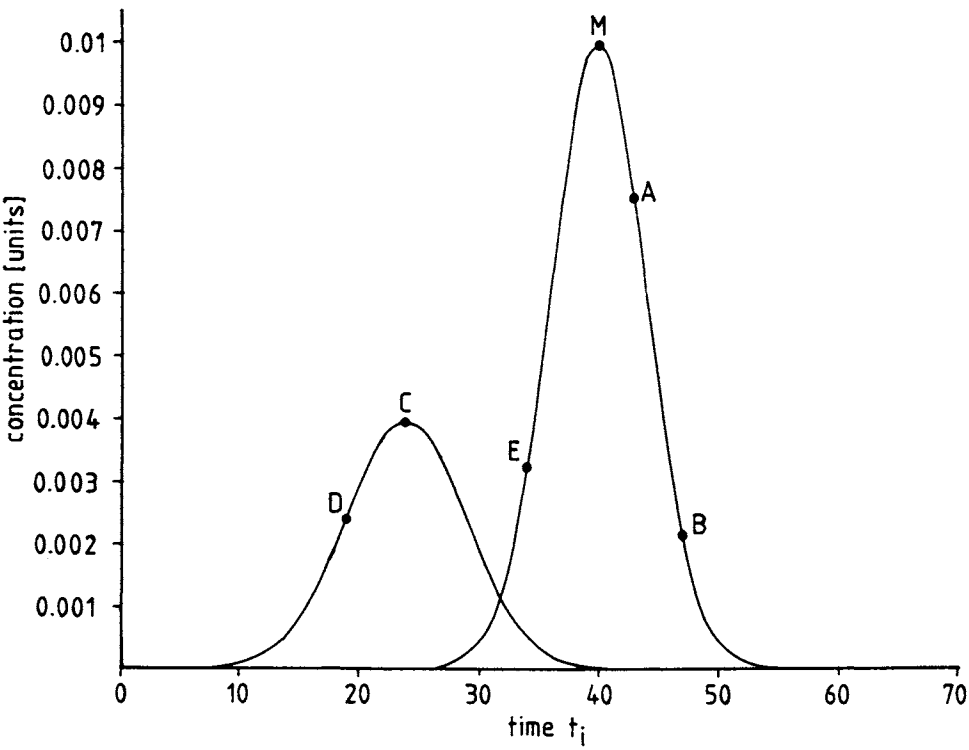


Fig. 9.6. (a) Chromatogram obtained by HPLC-DAD.

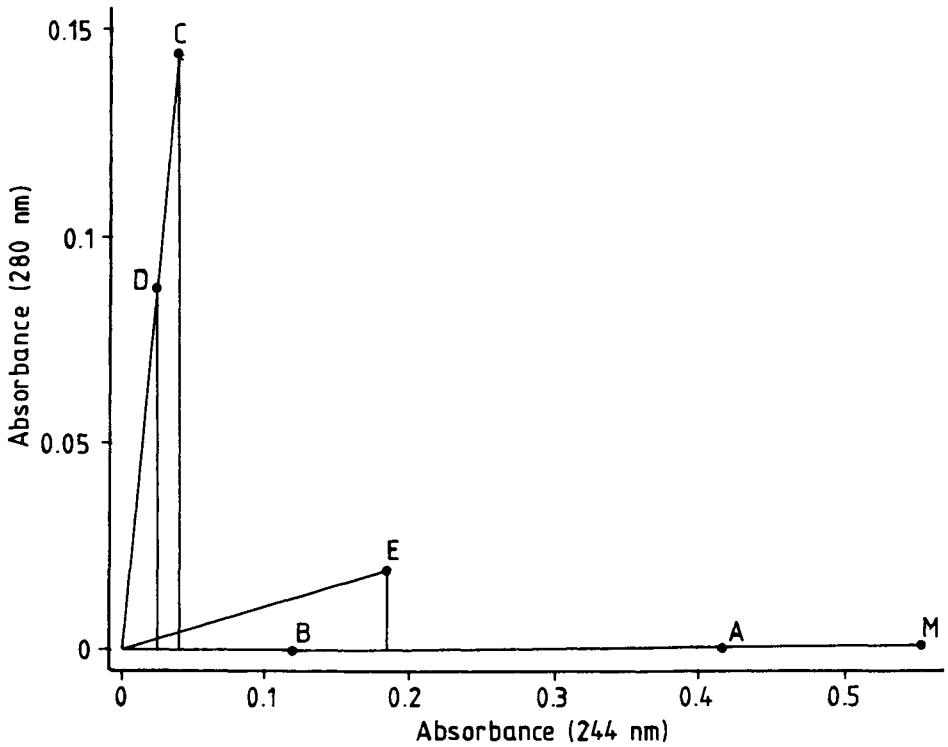


Fig. 9.6. (b) Plot of absorbance at $\lambda = 244$ nm against absorbance at $\lambda = 280$ nm. The meaning of A, B, C, D, E and M is given in the text.

In Fig. 9.6b we observe points such as M, the point representing the base vector $\mathbf{x}_t$ and A and B situated on a line with other points representing measurements of the pure main compound. Indeed the ratio of $A_{\lambda 244}/A_{\lambda 280}$ is constant so that all these points must fall on a line. Points C and D are on another line, and due to the pure "impurity". Point E is a mixture point. We then apply eq. (9.14) for each time t_i.

$$\mathbf{v}_i = \mathbf{x}_i - (\mathbf{x}_i \cdot \mathbf{x}_t)\mathbf{x}_t / \|\mathbf{x}_t\|^2$$

and measure the length of $\mathbf{v}_i$, $\|\mathbf{v}_i\|$.

In Fig. 9.6c the plot of $\|\mathbf{v}_i\|$ in function of time is given. Points such as A and B will yield orthogonal vectors with length close to zero, since the angle with respect to the base vector is close to zero. Points such as C and D have the same angle with respect to the base vector, but the length of $\mathbf{v}_i$ is larger for C than for D, because $\mathbf{x}_i$ is longer for C than for D. Therefore, the length of the $\mathbf{v}_i$ obtained from the projection of the points due to the impurity, follows the same evolution as the chromatographic profile of the impurity.

Gram–Schmidt orthogonalization is applied for instance in high resolution gas chromatography–Fourier transform IR to extract information from interferograms [2].

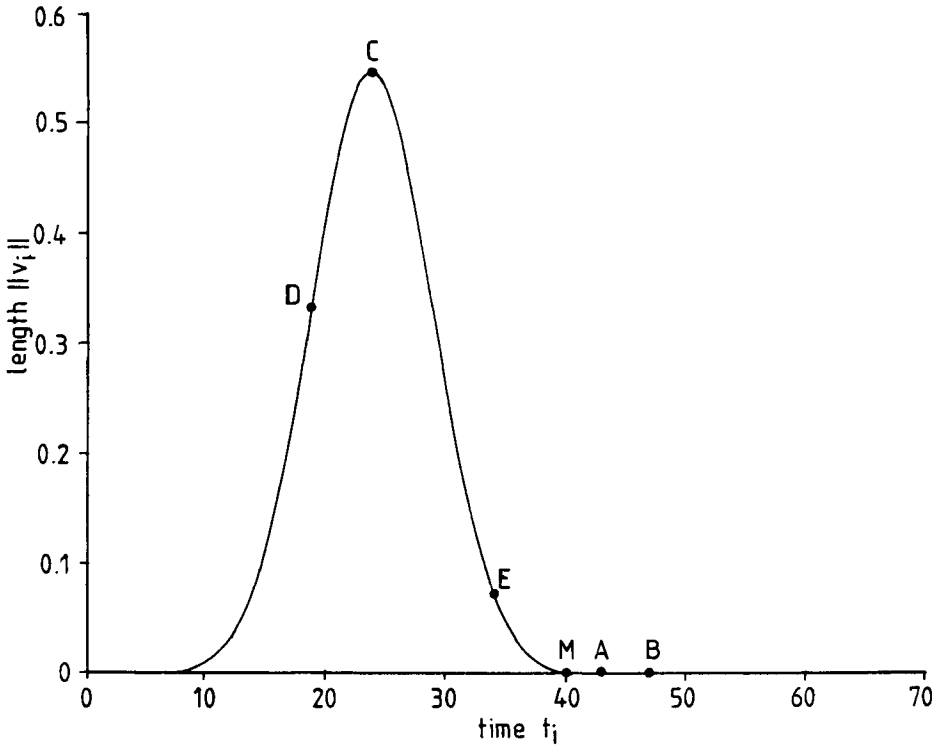


Fig. 9.6. (c) Plot of $\|\mathbf{v}_i\|$ in function of t_i.

9.2.8 Linear combinations, linear dependence and collinearity

When two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same direction, they are called *collinear*. This means that there is a scalar c, such that

$$\mathbf{x}_1 = c\,\mathbf{x}_2 \tag{9.16}$$

From Section 9.2.2.2, we know that eq. (9.16) describes an operation whereby $\mathbf{x}_1$ is shrunk or expanded by a factor c. The scalar c may also be negative, in which case $\mathbf{x}_1$ and $\mathbf{x}_2$ point in exactly opposite directions. The term collinear is sometimes used in an approximate sense. When two variables, are strongly correlated this means that there is a very small angle between the vectors that represent them (see Section 9.2.5). It follows that they have nearly, although not exactly, the same direction in space. Nevertheless, in regression this situation is described as highly collinear. A corollary to definition eq. (9.16) is that two vectors are *non-collinear* when no c satisfying eq. (9.16) is found.

Collinearity is a special case of linear dependence. If one vector of a set of k vectors can be written as a linear combination of the other vectors in the set, then the vectors are called *linearly dependent*. If not, they are linearly independent.

One calls *linear combination* of a set of vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_k}$ a new vector of the form

$$c_1 \mathbf{x_1} + c_2 \mathbf{x_2} + \ldots c_k \mathbf{x_k}$$

where $c_1, c_2, \ldots, c_k$ are scalars called the *coefficients* or *weights* of the linear combination.

Another definition of linear dependence is that there exists at least one $\mathbf{c} \neq \mathbf{0}$ for which their linear combination yields a $\mathbf{0}$ vector, i.e. a vector containing only 0 values

$$c_1 \mathbf{x_1} + c_2 \mathbf{x_2} + \ldots c_k \mathbf{x_k} = \mathbf{0} \tag{9.17}$$

Definition (9.17) follows from the earlier definition. Let us take an example

$$\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

are linearly dependent because

$$\begin{bmatrix} 2 \\ 3 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix} \qquad \text{(first definition)}$$

$$\text{or} \begin{bmatrix} 2 \\ 3 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + (-1) \begin{bmatrix} 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{(definition from (eq. 9.17))}$$

The example can be generalized. Any vector from S^2 space can be written as a combination of two non-collinear vectors from that same space. For instance, in Fig. 9.2a $\mathbf{x}$ and $\mathbf{y}$ are linearly independent. Any other vector in that space can be represented as a linear combination of $\mathbf{x}$ and $\mathbf{y}$. A set of only two non-collinear vectors in S^2 is necessarily linearly independent. Such a set of two linearly independent vectors for S^2 is said to constitute a *basis* for the vector space. From that set all other vectors in that space, here in the plane, can be obtained by linear combination. It is usual to use as basis the set of orthogonal vectors

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

However, this is not necessary. It is not even necessary that the vectors should be orthogonal. This can be generalized to S^n space.

9.2.9 Dimensionality

In Section 9.2.1 we defined the dimension of a vector (row or column) space as the number of elements in the vector. There are situations where the full number of dimensions is not needed.

The simplest such situation, where this occurs, is when there are fewer points than dimensions. Suppose we measure a spectrum of 19 NIR wavelengths for 10 objects. We can represent the data as 10 row vectors each with 19 elements and therefore we would define a 19-dimensional space. However, we really need only 10 dimensions. To understand this, consider the simpler situation of two row vectors (objects) of three elements (wavelengths). This situation is shown in Fig. 9.7a $\mathbf{x}_1$ and $\mathbf{x}_2$ together define a plane in the three-dimensional space. Thus only two dimensions are needed. The reduction of three to two dimensions means that we do not need the whole space, but only a subspace.

A vector *subspace* is defined as the set of vectors containing all linear combinations of $\mathbf{x}_1,\ldots,\mathbf{x}_k$. This is also called the *span* of a set of vectors. Geometrically it is the smallest space (line, plane or hyperplane) that contains all these vectors. The vectors are represented in that subspace without error, i.e. lengths and angles are preserved.

A special case occurs with *closed* data, such as data that describe mixtures. The components of a mixture always add up to 100% or, expressed as fractions, to 1. For instance, in two dimensions all mixtures must be situated on the line connecting pure components (fraction = 1) and in three dimensions on a plane (see Fig. 9.7b and c).

In general, reduction of dimensionality occurs when vectors are linearly dependent. Consider the following example:

$$\mathbf{x}_1 = \begin{bmatrix} 3 \\ 4 \\ 2 \\ -1 \\ 5 \end{bmatrix} \qquad \mathbf{x}_2 = \begin{bmatrix} 2 \\ 5 \\ 3 \\ 0 \\ -2 \end{bmatrix} \qquad \mathbf{x}_3 = \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \\ 7 \end{bmatrix}$$

Then $1\mathbf{x}_1 + (-1)\mathbf{x}_2 + (-1)\mathbf{x}_3 = \mathbf{0}$. We can plot the 5 objects described by $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, but observe that in fact all objects fall into a two-dimensional plane or subspace. It sometimes happens that all vectors fall near but not quite in a subspace. They are approximately linearly dependent. In that case the subspace can be used to represent with a small error the original vectors. This is the basis of dimensionality reduction in methods based on principal components (see Chapter 17).

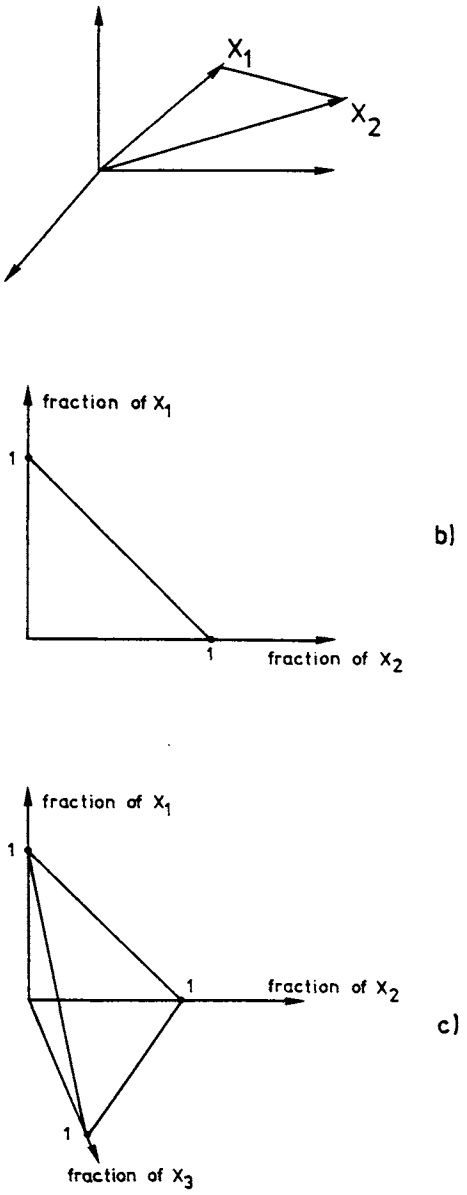


Fig. 9.7. Dimensionality reduction (a) due to fewer points than dimensions, (b) closure in two dimensions, and (c) closure in 3 dimensions.

9.3 Matrices

9.3.1 Definitions

A matrix is a rectangular arrangement of numbers and is represented by a capital letter in bold face:

$$\underset{n\times p}{\mathbf{X}} = \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & x_{ij} & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{9.18}$$

The matrix $\mathbf{X}$ has n rows and p columns and is called a $n\times p$ matrix. It is often represented as $\underset{n\times p}{\mathbf{X}}$. This matrix (see Section 9.1) can be viewed as a collection of p column vectors or n row vectors. The individual values of the vectors and the matrix are the elements of the matrix. They are denoted by the corresponding lower case letter: x_{ij} represents the element in the ith row and the jth column of the matrix $\mathbf{X}$. Thus the matrix $\mathbf{X}$ of eq. (9.1) can be represented as $\underset{5\times 4}{\mathbf{X}}$ and, e.g., $x_{53} = 25$.

A matrix for which the number of rows is equal to the number of columns ($n = p$) is a *square matrix*:

$$\underset{2\times 2}{\mathbf{X}} = \begin{bmatrix} 2 & -1 \\ 5 & 0 \end{bmatrix}$$

A square matrix can also be:

symmetric ($x_{ij} = x_{ji}$ for all i and j)

$$\underset{3\times 3}{\mathbf{X}} = \begin{bmatrix} 2 & -5 & 8 \\ -5 & 6 & 0 \\ 8 & 0 & 3 \end{bmatrix}$$

diagonal ($x_{ij} = 0$ for all $i \neq j$)

$$\underset{3\times 3}{\mathbf{X}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

triangular ($x_{ij} = 0$ for all $i > j$ or for all $i < j$)

$$\underset{3\times3}{\mathbf{X}} = \begin{bmatrix} 4 & 0 & 0 \\ 8 & -1 & 0 \\ -5 & 2 & 7 \end{bmatrix}$$

A square matrix for which the elements of the principal diagonal (from top-left to bottom-right) are all equal to 1 with all other elements being zero is an *identity matrix* ($x_{ij} = 1$ for all $i = j$ and $x_{ij} = 0$ for all $i \neq j$). An identity matrix is represented by the symbol **I**:

$$\underset{3\times3}{\mathbf{I}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

A matrix with all elements equal to zero is a *null matrix* represented by the symbol **0**:

$$\underset{2\times3}{\mathbf{0}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

By the *trace* of a square matrix is meant the sum of the diagonal elements of the matrix. For example the trace of the triangular matrix **X** given higher is:

$$\mathrm{tr}\left(\underset{3\times3}{\mathbf{X}}\right) = 4 - 1 + 7 = 10$$

The *transpose* $\mathbf{X}^T$ of a matrix **X** is obtained by interchanging the rows and columns of **X**. The elements x_{ij} of matrix **X** therefore become the elements x_{ji} of matrix $\mathbf{X}^T$. If for example:

$$\underset{3\times4}{\mathbf{X}} = \begin{bmatrix} 2 & 0 & 5 & 3 \\ 1 & 3 & 8 & 2 \\ 7 & 6 & 0 & 4 \end{bmatrix} \quad \text{then} \quad \underset{4\times3}{\mathbf{X}^T} = \begin{bmatrix} 2 & 1 & 7 \\ 0 & 3 & 6 \\ 5 & 8 & 0 \\ 3 & 2 & 4 \end{bmatrix}$$

It should be evident that a symmetric matrix and its transpose are identical.

9.3.2 Matrix operations

9.3.2.1 Addition and subtraction

The addition and subtraction of matrices is only possible if they have the same number of rows and columns. To add or subtract matrices the corresponding elements in the matrices are added or subtracted. The new matrix obtained has the same dimension as the original matrices. For example if:

$$\underset{2\times3}{\mathbf{X}} = \begin{bmatrix} 7 & 3 & 9 \\ 4 & 8 & 6 \end{bmatrix} \qquad \underset{2\times3}{\mathbf{B}} = \begin{bmatrix} -1 & 2 & 6 \\ 3 & 5 & 0 \end{bmatrix} \qquad \underset{2\times3}{\mathbf{C}} = \begin{bmatrix} 3 & 9 & -7 \\ 5 & 3 & 1 \end{bmatrix}$$

then:

$$\mathbf{D} = \mathbf{X} - \mathbf{B} + \mathbf{C} = \begin{bmatrix} 11 & 10 & -4 \\ 6 & 6 & 7 \end{bmatrix}$$

As for vector addition matrix addition is commutative and associative (see Section 9.2.2.1)

9.3.2.2 Multiplication by a scalar

To multiply a matrix by a number k, each of the elements of the matrix is multiplied by that number:

$$\mathbf{B} = k\mathbf{A} = \mathbf{A}k$$

Therefore if

$$\mathbf{A} = \begin{bmatrix} 2 & 5 & 8 \\ 3 & -4 & 7 \end{bmatrix}, \quad \text{then } \mathbf{B} = 3\mathbf{A} = \begin{bmatrix} 6 & 15 & 24 \\ 9 & -12 & 21 \end{bmatrix}$$

9.3.2.3 Matrix multiplication

The product of two matrices only exists if the number of columns of the first matrix is equal to the number of rows of the second matrix. A new matrix is obtained with a number of rows equal to the number of rows of the first matrix and with a number of columns equal to the number of columns of the second matrix.

For example the two matrices **X** and **B** can be multiplied to obtain:

$$\underset{n\times m}{\mathbf{C}} = \underset{n\times p}{\mathbf{X}} \; \underset{p\times m}{\mathbf{B}} \tag{9.19}$$

The product **B X** will therefore only be possible if $n = m$. The order in which the multiplication is performed is clearly important. One sometimes uses terms such as postmultiplication or premultiplication. For instance, when one states that **X** is postmultiplied by **B**, this means that one performs the operation **X B** and not **B X**. The elements of the matrix **C** are obtained as:

$$c_{ij} = \sum_{k=1}^{p} x_{ik}\, b_{kj} \qquad (i = 1, ..., n; j = 1, ..., m)$$

If, for example,

$$\underset{3\times3}{\mathbf{X}} = \begin{bmatrix} 2 & 3 & 4 \\ 0 & 1 & 7 \\ 1 & 2 & 5 \end{bmatrix} \qquad \text{and} \qquad \underset{3\times2}{\mathbf{B}} = \begin{bmatrix} 1 & 3 \\ -2 & 1 \\ 4 & 5 \end{bmatrix}$$

$$\underset{3\times2}{\mathbf{C}} = \underset{3\times3}{\mathbf{X}}\ \underset{3\times2}{\mathbf{B}} = \begin{bmatrix} 12 & 29 \\ 26 & 36 \\ 17 & 30 \end{bmatrix}$$

since

$$c_{11} = (x_{11} \times b_{11}) + (x_{12} \times b_{21}) + (x_{13} \times b_{31})$$
$$= (2 \times 1) + \Big(3 \times (-2)\Big) + (4 \times 4) = 12$$

$$c_{12} = (x_{11} \times b_{12}) + (x_{12} \times b_{22}) + (x_{13} \times b_{32})$$
$$= (2 \times 3) + (3 \times 1) + (4 \times 5) = 29$$

$$c_{21} = (x_{21} \times b_{11}) + (x_{22} \times b_{21}) + (x_{23} \times b_{31})$$
$$= (0 \times 1) + \Big(1 \times (-2)\Big) + (7 \times 4) = 26$$

.
.
.

$$c_{32} = (x_{31} \times b_{12}) + (x_{32} \times b_{22}) + (x_{33} \times b_{32})$$
$$= (1 \times 3) + (2 \times 1) + (5 \times 5) = 30$$

Matrix multiplication is distributive and associative. Therefore:

$$\underset{n\times p}{\mathbf{X}} \left(\underset{p\times r}{\mathbf{B}} + \underset{p\times r}{\mathbf{C}} \right) = \underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\mathbf{B}} + \underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\mathbf{C}}$$

$$\underset{n\times p}{\mathbf{X}} \left(\underset{p\times r}{\mathbf{D}} + \underset{p\times r}{\mathbf{E}} \right) = \underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\mathbf{D}} + \underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\mathbf{E}}$$

$$\underset{n\times p}{\mathbf{X}} \left(\underset{p\times m}{\mathbf{F}}\ \underset{m\times r}{\mathbf{G}} \right) = \left(\underset{n\times p}{\mathbf{X}}\ \underset{p\times m}{\mathbf{F}} \right) \underset{m\times r}{\mathbf{G}}$$

However in general it is not commutative. It is easily verified that the result of the multiplication of two square matrices depends, in general, on the order in which the multiplication is carried out. Therefore generally:

$$\underset{p\times p}{\mathbf{X}}\ \underset{p\times p}{\mathbf{B}} \neq \underset{p\times p}{\mathbf{B}}\ \underset{p\times p}{\mathbf{X}}$$

Useful properties are:

$$\underset{n\times n}{\mathbf{I}}\ \underset{n\times p}{\mathbf{X}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times p}{\mathbf{I}} = \underset{n\times p}{\mathbf{X}} \tag{9.20}$$

$$\underset{n\times p}{\mathbf{X}}\ \underset{p\times n}{\mathbf{X}^{\mathrm{T}}} \text{ is a symmetric } n\times n \text{ matrix} \tag{9.21}$$

$$\underset{p\times n}{\mathbf{X}^{\mathrm{T}}}\ \underset{n\times p}{\mathbf{X}} \text{ is a symmetric } p\times p \text{ matrix} \tag{9.22}$$

$$\underset{n\times n}{\mathbf{X}}\ \underset{n\times n}{\mathbf{X}} = \underset{n\times n}{\mathbf{X}^2} \tag{9.23}$$

$$\underset{n\times p}{\mathbf{0}}\ \underset{p\times q}{\mathbf{X}} = \underset{n\times q}{\mathbf{0}} \text{ and } \underset{n\times p}{\mathbf{X}}\ \underset{p\times q}{\mathbf{0}} = \underset{n\times q}{\mathbf{0}} \tag{9.24}$$

9.3.2.4 Examples of matrix multiplication

9.3.2.4.1 Linear mixture models

Let us consider some examples of matrix multiplication which are of special interest and will be required in later chapters.

In many cases, the response of a mixture can be modelled as a weighted sum of responses of the individual components, the weights being proportional to the concentrations. For example, consider the absorbance at three wavelengths of a two-component mixture. This can be written as

$$\begin{aligned} A_1 &= \varepsilon_{11}x_1 + \varepsilon_{12}x_2 \\ A_2 &= \varepsilon_{21}x_1 + \varepsilon_{22}x_2 \\ A_3 &= \varepsilon_{31}x_1 + \varepsilon_{32}x_2 \end{aligned} \tag{9.25}$$

where A_1 is the absorbance at λ_1, ε_{11} is the absorptivity at λ_1 for x_1 and x_1 is the concentration of component 1. Calling **a** the column vector of absorbances

$$\mathbf{a} = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix}$$

x the column vector of concentrations

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and **E** the matrix of ε-coefficients

$$\mathbf{E} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \\ \varepsilon_{31} & \varepsilon_{32} \end{bmatrix}$$

we can write:

$$\mathbf{a} = \mathbf{E}\,\mathbf{x}$$

This yields the equations (9.25) (see Chapter 10).

9.3.2.4.2 Weighting by multiplication with a diagonal matrix

In some cases, it will be necessary to weight certain variables or objects, for instance for standardization or in weighted regression. This can be done with a weight matrix. This is a diagonal matrix. Suppose we have a matrix **X** and want to weight the first column by multiplying it by w_1, the second column by w_2, etc. We can then write

$$\mathbf{X}_w = \mathbf{X}\,\mathbf{W}$$

where $\mathbf{X}_w$ is the weighted matrix and **W** the weight matrix

$$\begin{bmatrix} x_{w11} & x_{w12} \\ x_{w21} & x_{w22} \\ x_{w31} & x_{w32} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} \cdot \begin{bmatrix} w_1 & 0 \\ 0 & w_2 \end{bmatrix}$$

We can verify that, for instance

$x_{w11} = x_{11}\,w_1 + x_{12}\,0 = x_{11}\,w_1$ and $x_{w32} = x_{32}\,w_2$

To weight rows, one pre-multiplies **X** with a diagonal weight matrix **W**, i.e. $\mathbf{X}_w = \mathbf{W}\,\mathbf{X}$

$$\begin{bmatrix} x_{w11} & x_{w12} \\ x_{w21} & x_{w22} \\ x_{w31} & x_{w32} \end{bmatrix} = \begin{bmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ 0 & 0 & w_3 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}$$

For instance, now:

$x_{w11} = w_1\,x_{11} + 0\,x_{21} + 0\,x_{31} = w_1\,x_{11}$ and $x_{w32} = w_3\,x_{32}$

9.3.2.4.3 Regression models

In multiple regression (Chapter 10), one needs equations of the type

$$\mathbf{y} = \mathbf{X}\,\mathbf{b}$$

where **y** is the column vector of responses, **b** the column vector of b-parameters and **X** the matrix of x_1 and x_2 values.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

We can verify that

$$y_1 = b_1 x_{11} + b_2 x_{12}$$

If we want to include a constant term, b_0

$$y = b_0 + b_1 x_1 + b_2 x_2$$

then this can be achieved by writing

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

9.3.2.4.4 Variance–covariance matrix

Looking back at Table 9.1, we would like to determine the standard deviations (or variances) of the four variables and the correlations (or covariances) between them. This information can be summarized in a 4×4 table which takes the following form:

$$\begin{bmatrix} var\,(1) & cov\,(1,2) & cov\,(1,3) & cov\,(1,4) \\ cov\,(2,1) & var\,(2) & cov\,(2,3) & cov\,(2,4) \\ cov\,(3,1) & cov\,(3,2) & var\,(3) & cov\,(3,4) \\ cov\,(4,1) & cov\,(4,2) & cov\,(4,3) & var\,(4) \end{bmatrix}$$

If we consider this table to be a matrix, then this is a *variance–covariance matrix*. Let us consider this more generally for **X**.

Matrix **X** has column means $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_p$. We subtract the column means from the elements of the corresponding column (column-centring, see Section 9.2.3). This yields:

$$\begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \ldots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \ldots & x_{2p} - \bar{x}_p \\ & & \ldots & \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \ldots & x_{np} - \bar{x}_p \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \ldots & u_{1p} \\ u_{21} & u_{22} & \ldots & u_{2p} \\ & \ldots & & \\ u_{n1} & u_{n2} & & u_{np} \end{bmatrix} = \mathbf{U}$$

We now premultiply $\mathbf{U}$ with $\mathbf{U}^T$ and divide by $(n-1)$

$$\frac{1}{n-1}(\mathbf{U}^T\mathbf{U}) = \begin{bmatrix} u_{11} & u_{21} & \dots & u_{n1} \\ u_{12} & u_{22} & \dots & u_{n2} \\ & \dots & & \\ u_{1p} & u_{2p} & \dots & u_{np} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1p} \\ u_{21} & u_{22} & \dots & u_{2p} \\ & \dots & & \\ u_{n1} & u_{n2} & \dots & u_{np} \end{bmatrix} / (n-1)$$

$$= \begin{bmatrix} \Sigma u_{i1}^2 & \Sigma u_{i1}u_{i2} & & \Sigma u_{i1}u_{ip} \\ \Sigma u_{i1}u_{i2} & \Sigma u_{i2}^2 & & \Sigma u_{i2}u_{ip} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \Sigma u_{i1}u_{ip} & \Sigma u_{i2}u_{ip} & \dots & \Sigma u_{ip}^2 \end{bmatrix} / (n-1)$$

where Σ stands for $\sum_{i=1,n}$. The $(\mathbf{U}^T\mathbf{U})$ matrix is called the *dispersion matrix* for $\mathbf{U}$.

By dividing each element by $n-1$, we obtain the variance–covariance matrix of $\mathbf{X}$:

$$\text{cov}(\mathbf{X}) = \begin{bmatrix} s_1^2 & \text{cov}(1,2) & \dots & \text{cov}(1,p) \\ \text{cov}(2,1) & s_2^2 & \dots & \text{cov}(2,p) \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \text{cov}(p,1) & \text{cov}(p,2) & \cdots & s_p^2 \end{bmatrix} \tag{9.26}$$

where s_1^2 is the variance of the x-values in column 1 and cov(1,2) = cov(2,1) is the covariance between the x-values in columns 1 and 2. Often the shorter term *covariance matrix* is used. This matrix and the derived correlation matrix will be used in several later chapters in this book, starting with Chapter 10.

9.3.2.5 Inverse of a square matrix

In analogy with the inverse of a non-zero number which multiplied by the initial number equals unity, the inverse $\mathbf{X}^{-1}$ of a *non-singular* or *regular* square matrix $\mathbf{X}$ is such that

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^{-1}\mathbf{X} = \mathbf{I} \tag{9.27}$$

where $\mathbf{I}$ is an identity matrix. If $\mathbf{X}$ is singular, $\mathbf{X}^{-1}$ does not exist (see also Section 9.3.5). Since multiplying with the inverse of a number is equivalent to dividing by that number, matrix inversion can be seen as the equivalent of division. The

computation of an inverse is tedious especially with large matrices and will not be discussed further here. Moreover, it is best to use available computer subroutines which ensure accurate calculations. This is important to avoid round-off errors. The following characteristics of the inverse matrix are useful for some of the following chapters:

$$(\mathbf{X}^{-1})^{-1} = \mathbf{X} \tag{9.28}$$

$$(\mathbf{X}^{-1})^{T} = (\mathbf{X}^{T})^{-1} = \mathbf{X}^{-T} \tag{9.29}$$

$$(\mathbf{X}\,\mathbf{B})^{-1} = \mathbf{B}^{-1}\,\mathbf{X}^{-1} \tag{9.30}$$

$$\text{If } \mathbf{X} = \begin{bmatrix} x_{11} & 0 & 0 \\ 0 & x_{22} & 0 \\ 0 & 0 & x_{33} \end{bmatrix} \quad \text{then } \mathbf{X}^{-1} = \begin{bmatrix} 1/x_{11} & 0 & 0 \\ 0 & 1/x_{22} & 0 \\ 0 & 0 & 1/x_{33} \end{bmatrix} \tag{9.31}$$

9.3.3 Regression modelling and projection

Suppose we have measured the UV absorbance, y, of a substance with concentration x at one wavelength. The following results are obtained:

$y = 0.11 \quad x = 10$; $y = 0.19 \quad x = 20$; $y = 0.30 \quad x = 30$.

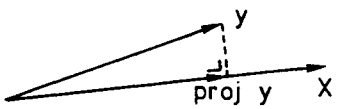
We can also consider that the y results are elements of a column vector **y** and that the corresponding x values constitute a column vector **x**.

$$\mathbf{y} = \begin{bmatrix} 0.11 \\ 0.19 \\ 0.30 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} 10 \\ 20 \\ 30 \end{bmatrix}$$

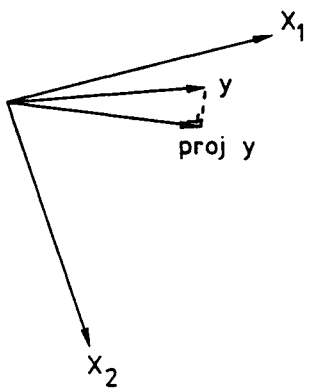
As can be seen, the two vectors diverge slightly. To simplify, we suppose that the relation between y and x is $y = ax$, i.e. there is no intercept. In vector notation,

$$\mathbf{y} = a\mathbf{x}$$

If there were no random error, **y** and **x** would therefore have exactly the same direction in space (Fig. 9.8a). Because there is random error this is not true. To fit the model, one selects a vector that has the same direction as **x** and is as close as possible to **y**. This vector is given by the projection of **y** on **x**, **proj y**. The difference between **y** and **proj y** is **e**, the vector of model errors or residuals. One wants this as small as possible, i.e. one wants it to have the smallest length. The length is given by $\|\mathbf{e}\| = \sqrt{\Sigma\, e_i^2}$. Since this needs to be as small as possible, this means one must minimize the sum of squared residuals, which is exactly what was done in Chapter 8.



a)



b)

Fig. 9.8 Regression of **y** on **x** (a) and on the plane defined by $\mathbf{x}_1$ and $\mathbf{x}_2$ (b).

Let us now take a first look at multiple regression, the subject of the next chapter (Chapter 10). Instead of measuring at only one wavelength, we measure now at two wavelengths. We still measure only one substance. The results now can be written as a matrix with two columns

$$\mathbf{y} = \begin{bmatrix} 10 \\ 20 \\ 30 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 0.11 & 0.21 \\ 0.19 & 0.40 \\ 0.30 & 0.55 \end{bmatrix}$$

In row space the two columns of **X** are column vectors $\mathbf{x}_1$ (the measurement results at λ_1) and $\mathbf{x}_2$ (at λ_2). These two column vectors together define a plane (see Fig. 9.8b). If there were no random errors, we could write

$$\mathbf{y} = a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2$$

In other words, **y** is a linear combination of $\mathbf{x}_1$ and $\mathbf{x}_2$ and **y** is therefore situated in the plane $\mathbf{x}_1$,$\mathbf{x}_2$. When there is random error, **y** will not fit exactly into the plane. To estimate **y** we would then select a vector in the plane that is closest to **y**, in other

words we would project **y** on the plane. Again the difference is **e**, the length of which is given by the root of the sum of the squared residuals.

It can be shown that the projection of any vector on the subspace spanned by linearly independent vectors ($\mathbf{x_1}$, ..., $\mathbf{x_m}$), forming together matrix **X**, is obtained with the *orthogonal projection operator* $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. This result will be applied in Chapter 10 and 29 to obtain the coefficients in multiple regression.

9.3.4 Determinant of a square matrix

A square matrix **X** can be characterized by a number, called the *determinant* $|\mathbf{X}|$. For a 2×2 matrix

$$\underset{2\times2}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

the determinant is:

$$|\mathbf{X}| = x_{11}\,x_{22} - x_{12}\,x_{21} \tag{9.32}$$

e.g. for

$$|\mathbf{X}| = \begin{vmatrix} 39 & 24 \\ 24 & 21 \end{vmatrix} = 39 \times 21 - 24 \times 24 = 243$$

Equation (9.32) is called the expanded form of the determinant. For a 3×3 matrix

$$\underset{3\times3}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}$$

the determinant can be obtained via the so-called *minors*, M_{ij}, and cofactors, A_{ij}.

The minor M_{ij} is the determinant of **X** after deletion of the ith row and jth column. Suppose one deletes row 1 and column 1, then:

$$M_{11} = \begin{vmatrix} x_{22} & x_{23} \\ x_{32} & x_{33} \end{vmatrix}$$

The cofactor is given by

$$A_{ij} = (-1)^{i+j} M_{ij}$$

so that:

$$A_{11} = (-1)^{1+1} M_{11} = M_{11}$$

The determinant of **X** is then obtained by selecting any column- or row-vector. The scalar products of the elements of this vector and the corresponding cofactors are then formed and summed

$$|\mathbf{X}| = \sum_{i \text{ or } j} a_{ij} A_{ij} \qquad \text{for any } i \text{ and } j$$

In our example, we can for instance decide to delete the first row vector. The scalar product for the first element of the deleted vector and its cofactor is $x_{11} M_{11}$ and $|\mathbf{X}|$ is obtained as:

$$|\mathbf{X}| = x_{11} \begin{vmatrix} x_{22} & x_{23} \\ x_{32} & x_{33} \end{vmatrix} - x_{12} \begin{vmatrix} x_{21} & x_{23} \\ x_{31} & x_{33} \end{vmatrix} + x_{13} \begin{vmatrix} x_{21} & x_{22} \\ x_{31} & x_{32} \end{vmatrix}$$

which is equal to:

$$x_{11} x_{22} x_{33} - x_{11} x_{23} x_{32} + x_{12} x_{23} x_{31} - x_{12} x_{21} x_{33} + x_{13} x_{21} x_{32} - x_{13} x_{31} x_{22}$$

Instead of passing through the minors, we can apply the following equation:

$$\underset{n \times n}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$$

$$|\mathbf{X}| = \sum (-1)^r x_{1k_1} x_{2k_2} \dots x_{nk_n}$$

The symbol $\sum$ indicates the sum of all terms for the $n!$ (n-factorial) possible permutations (k_1, k_2, ... k_n) of the numbers 1, 2, ... n. The integer r represents the number of inversions in the permutation (k_1, k_2, ... k_n). In the permutation (k_1, k_2, ... k_n) the numbers k_j and k_k ($j < k$) form an inversion if $k_j > k_k$. For example in the permutation (3,1,5,4,2) of the numbers 1,2,3,4,5 each of the pairs (3,1), (3,2), (5,4), (5,2) and (4,2) is an inversion. Consequently this permutation possesses 5 inversions. The number n is the *order* of the determinant.

For a 3×3 matrix the determinant is:

$$\begin{aligned} |\mathbf{X}| &= (-1)^0 x_{11} x_{22} x_{33} + (-1)^1 x_{11} x_{23} x_{32} + \\ &\quad (-1)^2 x_{12} x_{23} x_{31} + (-1)^1 x_{12} x_{21} x_{33} + \\ &\quad (-1)^2 x_{13} x_{21} x_{32} + (-1)^3 x_{13} x_{22} x_{31} \\ &= x_{11} x_{22} x_{33} - x_{11} x_{23} x_{32} + x_{12} x_{23} x_{31} - \\ &\quad x_{12} x_{21} x_{33} + x_{13} x_{21} x_{32} - x_{13} x_{22} x_{31} \end{aligned}$$

If for example

$$\mathbf{X} = \begin{bmatrix} 2 & 3 & 4 \\ 0 & -1 & 7 \\ 1 & 2 & 5 \end{bmatrix}$$

then $|\mathbf{X}| = (2\times(-1)\times5) - (2\times7\times2) + (3\times7\times1) -$

$(3\times0\times5) + (4\times0\times2) - (4\times(-1)\times1)$

$= -10 - 28 + 21 - 0 + 0 + 4 = -13$

Determinants are used among others to solve sets of simultaneous equations, for which we refer the reader to introductory books on algebra. The geometrical interpretation of determinants will be discussed in Chapter 29.

9.3.5 Rank of a square matrix

A square matrix is said to be *singular* if there is at least one linear dependency among the rows or columns of the matrix. As a result the determinant will be zero.

In the following matrix:

$$\underset{3\times3}{\mathbf{X}} = \begin{bmatrix} 1 & 0 & -2 \\ 4 & 8 & 0 \\ 3 & 7 & 1 \end{bmatrix}$$

the elements of the second column are equal to twice the elements of the first column added to the elements of the third column. It can be verified that $|\mathbf{X}| = 0$. Matrices that have a very small determinant are close to being singular. They are called *ill conditioned* and are known to be difficult to invert correctly. In spectroscopy, matrices are often ill conditioned because the vectors that constitute it are highly collinear in the sense described in Section 9.2.8.

The *rank*, $r(\mathbf{X})$, of a matrix $\mathbf{X}$ is the maximum number of linearly independent columns or, equivalently, rows. When two rows are linearly dependent or collinear the determinant is zero. It follows that the rank can also be defined as the order of the non-zero determinant of the largest order that it contains.

The last mentioned square matrix $\mathbf{X}$ is only of rank 2 ($r = 2$)

since $|\mathbf{X}| = 0$, but for instance $\begin{vmatrix} 1 & 0 \\ 4 & 8 \end{vmatrix} = 8 \neq 0$.

Therefore a square $\mathbf{X}$ matrix is regular or non-singular if its rank $r(\mathbf{X}) = n$, which means that $|\mathbf{X}| \neq 0$. The concept of rank for a non-square matrix will be discussed in Chapter 29.

References

1. F. Cuesta Sánchez, M.S. Khots, D.L. Massart and J.O. De Beer, Algorithm for the assessment of peak purity in liquid chromatography with photodiode-array detection. Anal. Chim. Acta, 285 (1994) 181–192.
2. R.L. White, G.N. Giss, G.M. Brissey and C.L. Wilkins, Comparison of methods for reconstruction of gas chromatogram for interferometric gas chromatography/infrared spectrometry data. Anal. Chem. 53 (1981) 1778–1782.

Chapter 10

Multiple and Polynomial Regression

10.1 Introduction

In Chapter 8 the simple straight line model

$$\eta = \beta_0 + \beta_1 x$$

that relates the dependent variable η to a single x variable has been described. However if we suspect that η is dependent on different variables $x_1, x_2, \ldots x_m$ multivariate functional relationships should be considered.

In this chapter we only describe multivariate models that are linear or first-order in the regression parameters, which means models that can be written in the following general form

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m = \beta_0 + \sum_{i=1}^{m} \beta_i x_i \qquad (10.1)$$

The following relationship

$$\eta = \beta_0 + \frac{\beta_1}{x_1^2} + \beta_2 \log x_2$$

is also a linear model since by taking $x'_1 = 1/x_1^2$ and $x'_2 = \log x_2$ a relationship as described in eq. (10.1) is obtained.

Non-linear relationships such as the following function

$$\eta = \beta_0 + \log(x - \beta_1)$$

are discussed in Chapter 11. Some non-linear models are intrinsically linear since they can be transformed into a linear relationship. The exponential function

$$\eta = \beta_0 e^{\beta_1 x}$$

for example, can be transformed to a linear function by taking the natural logarithm which results in the following linear form

$$\ln \eta = \beta'_0 + \beta_1 x$$

A special class of linear models consists of *polynomials*. If in eq. (10.1) $x_1 = x$, $x_2 = x^2$, ..., $x_m = x^m$ an mth degree polynomial relationship

$$\eta = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_m x^m = \sum_{i=0}^{m} \beta_i x^i \qquad (10.2)$$

between the dependent variable and a single independent variable is obtained. It is obvious that, except for $m = 1$, the linear models represented by eq. (10.2) do not describe straight lines. Therefore to avoid confusion with "non-linear" used in the sense described earlier, the term *curvilinear* is sometimes preferred to indicate linear models that describe curved lines or surfaces.

10.2 Estimation of the regression parameters

The least squares procedure as described in Section 8.2.1 can be extended to estimate the regression coefficients, $\beta_0,\beta_1,\ldots, \beta_m$, in the multiple linear regression situation. Consider n observations $y_1,y_2,\ldots y_n$, each with variance σ^2, obtained at n different combinations of the independent variables, $x_1,x_2,\ldots,x_m$ ($n > m$). If a multivariate model as given in eq. (10.1) is assumed between the response and the m x-variables, each observation can be represented as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i \qquad (10.3)$$

with y_i: the ith observation ($i = 1,2, \ldots n$)
x_{ik}: the value of the kth independent variable for observation i ($k = 1, \ldots m$)
ε_i: the ith residual. It is again assumed that the ε_i's are independent, normally distributed random variables with mean 0 and constant variance σ^2 (see also Section 8.2.1).

As described in Chapter 8 for the straight line regression, the least squares estimates ($b_0,b_1, \ldots b_m$) of the $p(p = m + 1)$ unknown regression coefficients ($\beta_0,\beta_1 \ldots \beta_m$) are obtained by minimizing the sum of the squared residuals. This requires the solution of a system of p normal equations with p unknowns.

By expressing the regression problem in matrix notation a solution is obtained that is applicable to any linear regression situation, including the simple straight line. We consider the following vectors and matrices

– the vector of observations $\mathbf{y}$

$$\underset{n\times 1}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

– the vector of the parameters to be estimated $\boldsymbol{\beta}$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_m \end{bmatrix}$$

– the independent variable matrix $\mathbf{X}$

$$\underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

The 1's in the first column allow for the estimation of the intercept, β_0. They correspond to the value of the x variable in the first term of eq. (10.1).

– the error vector $\boldsymbol{\varepsilon}$

$$\underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

The model represented by eq. (10.3) then becomes:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{10.4}$$

It can be shown [1] that minimizing the sum of the squared residuals provides the normal equations

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y} \tag{10.5}$$

in matrix notation. When the matrix $\mathbf{X}^T\mathbf{X}$ is non-singular the least squares estimate, $\mathbf{b}$, of $\boldsymbol{\beta}$ is therefore obtained as

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ . \\ b_m \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y} \tag{10.6}$$

This is the least squares solution applicable to all models that are linear in the parameters.

With these regression coefficients the estimated response values $\hat{\mathbf{y}}$ can be calculated

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ . \\ . \\ . \\ \hat{y}_n \end{bmatrix} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y} \tag{10.7}$$

In this expression $\mathbf{X}(\mathbf{X}^T\,\mathbf{X})^{-1}\,\mathbf{X}^T$ is known as the *hat matrix*, $\mathbf{H}$.

From the estimated and the measured y values the residual variance which is an estimate of the experimental error, σ^2, if the model is correct, is obtained as

$$s_e^2 = \frac{\sum e_i^2}{n-p} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p} \tag{10.8}$$

where $e_i = (y_i - \hat{y}_i)$ represents the residual for the ith measurement. In matrix notation this becomes

$$s_e^2 = \frac{\mathbf{e}^T\,\mathbf{e}}{n-p}$$

Equation (10.8) is a generalization of eq. (8.6) for the regression situation with p regression coefficients.

It is easily checked that for the straight line regression, eq. (10.5) indeed yields the normal equations given in Section 8.2.1. Since with $m = 1$

$$\mathbf{X} = \begin{bmatrix} 1\ x_1 \\ 1\ x_2 \\ .\ . \\ .\ . \\ .\ . \\ 1\ x_n \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

and

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

Therefore eq. (10.5) can be written as

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

which yields the normal equations as derived in Section 8.2.1.

$$nb_0 + b_1 \sum x_i = \sum y_i$$

$$b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i$$

Example:

Table 10.1 lists part of the stack loss data set given by Brownlee [2]. The data have been rearranged. They are obtained from a plant for the oxidation of ammonia to nitric acid. The dependent variable, y, is an inverse measure of the overall efficiency of the plant since it is 10 times the percentage of the ingoing ammonia that is lost. It has been studied during 17 days as a function of three predictor variables: x_1 is the rate of operation of the plant, x_2 is the temperature of the cooling water circulated through the coils in the adsorption tower for the nitric acid, and x_3 is the concentration of acid circulating (in arbitrary units).

The model relating y and the three x variables is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

The least squares estimates of β_0, β_1, β_2 and β_3 are obtained from eq. (10.6)

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

where

TABLE 10.1

The adapted stack loss data set [2]

i	x_1 Rate	x_2 Temperature	x_3 Acid concentration	y Stack loss
1	80	27	88	37
2	62	22	87	18
3	62	23	87	18
4	62	24	93	19
5	62	24	93	20
6	58	23	87	15
7	58	18	80	14
8	58	18	89	14
9	58	17	88	13
10	58	18	82	11
11	58	19	93	12
12	50	18	89	8
13	50	18	86	7
14	50	19	72	8
15	50	19	79	8
16	50	20	80	9
17	56	20	82	15

$$\mathbf{X} = \begin{bmatrix} 1 & 80 & 27 & 88 \\ 1 & 62 & 22 & 87 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & 50 & 20 & 80 \\ 1 & 56 & 20 & 82 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & . & . & . & 1 & 1 \\ 80 & 62 & . & . & . & 50 & 56 \\ 27 & 22 & . & . & . & 20 & 20 \\ 88 & 87 & . & . & . & 80 & 82 \end{bmatrix} \begin{bmatrix} 1 & 80 & 27 & 88 \\ 1 & 62 & 22 & 87 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & 50 & 20 & 80 \\ 1 & 56 & 20 & 82 \end{bmatrix}$$

$$= \begin{bmatrix} 17 & 982 & 347 & 1455 \\ 982 & 57596 & 20300 & 84354 \\ 347 & 20300 & 7215 & 29796 \\ 1455 & 84354 & 29796 & 125053 \end{bmatrix}$$

The inverse of this matrix is

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 14.269046605 & 0.000999288 & -0.041431100 & -0.156823712 \\ 0.000999288 & 0.002898139 & -0.005047759 & -0.000763841 \\ -0.041431100 & -0.005047759 & 0.017552937 & -0.000295286 \\ -0.156823712 & -0.000763841 & -0.000295286 & 0.002418253 \end{bmatrix}$$

and

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 80 & 62 & \cdots & 50 & 56 \\ 27 & 22 & \cdots & 20 & 20 \\ 88 & 87 & \cdots & 80 & 82 \end{bmatrix} \begin{bmatrix} 37 \\ 18 \\ \vdots \\ 9 \\ 15 \end{bmatrix} = \begin{bmatrix} 246 \\ 15032 \\ 5295 \\ 21320 \end{bmatrix}$$

Consequently eq. (10.6) becomes

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y}) = \begin{bmatrix} -37.65245229 \\ 0.79767627 \\ 0.57734001 \\ -0.06707647 \end{bmatrix}$$

and the regression equation is

$$\hat{y} = -37.652 + 0.798x_1 + 0.577x_2 - 0.067x_3$$

In this last expression the number of digits has been reduced but for all calculations all reported digits are used. Small differences will nevertheless be noticed when doing the calculations with a computer regression routine since these carry even more digits to reduce the round-off errors.

For the different combinations of the predictor variables of Table 10.1 the estimated response (see eq. (10.7)) and the residuals are summarized in Table 10.2. The residual variance (eq. (10.8)) calculated from these data is

$$s_e^2 = \frac{\Sigma e_i^2}{n-p} = \frac{20.401}{13} = 1.569$$

TABLE 10.2
Estimated stack loss and residuals

i	y	$\hat{y}$	e
1	37	35.8471	1.1529
2	18	18.6693	–0.6693
3	18	19.2466	–1.2466
4	19	19.4215	–0.4215
5	20	19.4215	0.5785
6	15	16.0559	–1.0559
7	14	13.6388	0.3612
8	14	13.0351	0.9649
9	13	12.5248	0.4752
10	11	13.5046	–2.5046
11	12	13.3441	–1.3441
12	8	6.6537	1.3463
13	7	6.8549	0.1451
14	8	8.3713	–0.3713
15	8	7.9018	0.0982
16	9	8.4120	0.5880
17	15	13.0639	1.9361

10.3 Validation of the model

10.3.1 Examination of the overall regression equation

10.3.1.1 Analysis of variance

In Chapter 8 ANOVA was proposed as a useful tool for the validation of the straight line model. ANOVA allows us to verify whether the predictor variables can explain a significant amount of the variance in the response variable. Moreover, if replicate measurements have been performed or if an estimate of the pure experimental error is available, the adequacy of the model chosen can also be checked.

In Table 10.3 the ANOVA table constructed in Section 8.2.2.2 is generalized for multiple regression with p regression coefficients. In this table p is the number of regression coefficients, n the number of observations and k the number of different settings (combinations) of the x variables ($m < k \leq n$).

The total sum of squares (SS_T) can be partitioned into the sum of squares due to regression (SS_{Reg}) and the residual error sum of squares (SS_{Res}). If replicate measurements are available the latter can be further decomposed in the sum of squares due to lack-of-fit (SS_{LOF}) and the sum of squares due to pure experimental error (SS_{PE}).

TABLE 10.3

Analysis of variance table for multiple regression

Source of variation	SS	Degrees of freedom	MS	F
Regression	SS_{Reg}	$p-1$	MS_{Reg}	MS_{Reg}/MS_{Res}
Residual	SS_{Res}	$n-p$	MS_{Res}	
Lack-of-fit	SS_{LOF}	$k-p$	MS_{LOF}	MS_{LOF}/MS_{PE}
Pure error	SS_{PE}	$n-k$	MS_{PE}	
Total	SS_T	$n-1$		

For the computation of these different sums of squares the expressions given in Chapter 8 can be used. They can also be expressed in matrix notation [1]

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \mathbf{y}^T\mathbf{y} - n\bar{y}^2 \qquad (df = n-1)$$

$$SS_{Reg} = \sum_i n_i (\hat{y}_i - \bar{y})^2 = \mathbf{b}^T \mathbf{X}^T \mathbf{y} - n\bar{y}^2 \qquad (df = p-1)$$

$$SS_{Res} = \sum_i \sum_j (y_{ij} - \hat{y}_i)^2 = \mathbf{e}^T \mathbf{e} = \mathbf{y}^T \mathbf{y} - \mathbf{b}^T \mathbf{X}^T \mathbf{y} \qquad (df = n-p)$$

$$SS_{PE} = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \qquad (df = n-k)$$

$$SS_{LOF} = \sum_i n_i (\bar{y}_i - \hat{y}_i)^2 = \mathbf{y}^T \mathbf{y} - \mathbf{b}^T \mathbf{X}^T \mathbf{y} - SS_{PE} \qquad (df = k-p)$$

The symbols used have the same meaning as in Chapter 8:

$\bar{y}$: the grand mean
n_i: the number of replicate measurements performed at a specific combination of the x variables
$\sum_{i=1}^{k} n_i = n$: the total number of observations
y_{ij}: one of the n_i measurements at a specific combination of the x variables
$\bar{y}_i$: the mean of the replicate measurements y_{ij} at a specific setting of the x variables
$\hat{y}_i$: the value of y at a specific combination of the x variables, estimated by the regression parameters.

Example:

For the stack loss data from Table 10.1 the following sums of squares can be calculated

TABLE 10.4

Analysis of variance table for the stack loss data

Source of variation	SS	Degrees of freedom	MS	F
Regression	795.834	3	265.278	169
Residual	20.401	13	1.569	
Total	816.235	16		

$$SS_T = \mathbf{y}^T \mathbf{y} - n\bar{y}^2 = 4376 - 17(14.47059)^2$$
$$= 4376 - 3559.765 = 816.235$$
$$SS_{Res} = \mathbf{e}^T \mathbf{e} = 20.401$$
$$SS_{Reg} = SS_T - SS_{Res} = 816.235 - 20.401 = 795.834$$

Since only 2 replicates (measurements 4 and 5) are available the residual sum of squares (SS_{Res}) has not been further partitioned into SS_{LOF} and SS_{PE}. An estimate of the pure error would only be based on 1 degree of freedom and consequently a possible lack-of-fit would be difficult to detect. Therefore lack-of-fit from the model $\hat{y} = -37.652 + 0.798x_1 + 0.577x_2 - 0.067x_3$ is not verified. The ANOVA results are summarized in Table 10.4. The calculated F being much larger than the tabulated $F_{0.05;3,13} = 3.41$ it can be concluded that the regression model accounts for a significant part of the variance of y.

To illustrate the validation of the model Table 10.1 has been adapted to contain several replicate measurements. These synthetic data are shown in Table 10.5 where the experimental conditions are identical for measurements 2 and 3; 4 and 5; 8 and 9; 12, 13 and 14; 15 and 16. Consequently there are 11 different settings of the x variables ($k = 11$). For these data the regression equation is $\hat{y} = 33.771 + 0.800x_1 + 0.535x_2 - 0.102x_3$ and the following sums of squares are obtained:

$$SS_T = 816.235$$

$$SS_{Res} = 23.400$$

$$SS_{Reg} = 816.235 - 23.400 = 792.835$$

$$SS_{PE} = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 = 2.167$$

$$SS_{LOF} = SS_{Res} - SS_{PE} = 23.400 - 2.167 = 21.233$$

Table 10.6 summarizes these ANOVA results. Since $F = MS_{LOF}/MS_{PE} = 8.40$ is larger than $F_{0.05;7,6} = 4.21$ the lack-of-fit term is significant. This lack-of-fit can be due to a wrong model or to the presence of outlying observations. The latter should be evaluated (see Section 10.9) before one starts adapting the model.

TABLE 10.5

Synthetic data adapted from Table 10.1 to illustrate the validation of the model

i	x_1 Rate	x_2 Temperature	x_3 Acid concentration	y Stack loss
1	80	27	88	37
2	62	23	87	18
3	62	23	87	18
4	62	24	93	19
5	62	24	93	20
6	58	23	87	15
7	58	18	80	14
8	58	17	88	14
9	58	17	88	13
10	58	18	82	11
11	58	19	93	12
12	50	18	86	8
13	50	18	86	7
14	50	18	86	8
15	50	19	80	8
16	50	19	80	9
17	56	20	82	15

TABLE 10.6

ANOVA table for the synthetic data from Table 10.5

Source of variation	SS	df	MS	F
Due to regression	792.835	3	264.278	
Residual	23.400	13		
Lack-of-fit	21.233	7	3.033	$MS_{LOF}/MS_{PE} = 8.40$
Pure error	2.167	6	0.361	
Total	816.235	16		

10.3.1.2 The coefficient of multiple determination

In Chapter 8, it was shown that for straight line regression between x and y the square of the correlation coefficient, $(r_{xy})^2$, represents the proportion of the variation of y that is explained by the x variable

$$r_{xy}^2 = \frac{SS_{Reg}}{SS_T} = \frac{\sum n_i (\hat{y}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2}$$

In multiple regression R^2, the *coefficient of multiple determination* is defined in the same way

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{SS_T - SS_{Res}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \tag{10.9}$$

It is used to estimate the proportion of the variation of y that is explained by the regression. R, which is called the *coefficient of multiple correlation*, is the correlation between $\mathbf{y}$ and $\hat{\mathbf{y}}$. For our example $R^2 = 795.834/816.235$, indicating that 97.5% of the variation in stack loss can be explained by the equation $\hat{y} = -37.652 + 0.798x_1 + 0.577x_2 - 0.067x_3$.

If there is no linear relationship between the dependent and the independent variables $R^2 = 0$; if there is a perfect fit $R^2 = 1$. The value of R^2 can generally be increased by adding additional x variables to the model. It can even reach unity if the number of coefficients in the model equals the number of observations ($p = n$): indeed a straight line ($p = 2$) perfectly fits two data points ($n = 2$). It follows that R^2 should be used with caution.

10.3.1.3 Analysis of the residuals

The analysis of the residuals can be performed as described in Section 8.2.2.1 for simple straight line regression. Statistical and graphical methods can be useful to detect deviations from normality. To detect shortcomings of the model residual plots, in which e_i is plotted against $\hat{y}_i$, should be examined. For the stack loss example the residual plot shown in Fig. 10.1 indicates that the model is adequate since no particular trend in the pattern of residuals is observed.

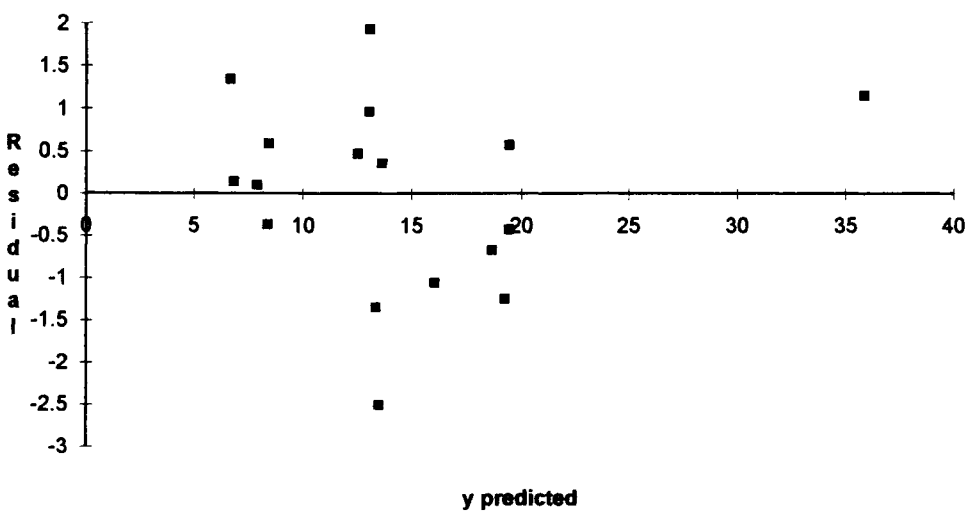


Fig. 10.1. Residual plot (e as a function of $\hat{y}$) for the stack loss data. Model: $\hat{y} = -37.652 + 0.798x_1 + 0.577x_2 - 0.067x_3$.

Additional information can also be obtained from other types of plots [1]: a time effect can be detected if the residuals can be plotted against the order in which the observations are made; the influence of a variable which has been recorded, but has not been included in the model, could be revealed from a plot of the residuals against that variable.

10.3.2 Importance of the predictor variables

From the previous results obtained for the stack loss data we can conclude that the model

$$\hat{y} = -37.652 + 0.798x_1 + 0.577x_2 - 0.067x_3$$

including three x variables (rate of operation, temperature and acid concentration) gives a good description of stack loss. However we might be interested to know whether all these variables are really necessary and what the importance of each variable is.

One way of answering these questions is to include the different terms sequentially in the model and to monitor the changes in the regression sum of squares. If the inclusion of a particular variable results in a significant increase of SS_{Reg}, this indicates that it explains a significant amount of the variation of y which is not accounted for by the other variables that are already in the equation.

Consider, for example, the following models for the same data:

Model 1: $\hat{y} = b_0 + b_1 x_1$ $SS_{Reg}(1)$

Model 2: $\hat{y} = b'_0 + b'_1 x_1 + b'_2 x_2$ $SS_{Reg}(2)$

If $SS_{Reg}(1)$ and $SS_{Reg}(2)$ are the regression sum of squares for these models then $SS_{Reg}(2) - SS_{Reg}(1)$ represents the increase of the regression sum of squares due to the inclusion of x_2 in the model. It is called $SS(x_2|x_1)$, the sum of squares due to x_2 given x_1 is already in the model. Since $SS_{Reg}(1)$ and $SS_{Reg}(2)$ have 1 and 2 degrees of freedom, respectively, there is 1 degree of freedom associated with $SS_{Reg}(x_2|x_1)$. The corresponding mean square, $MS(x_2|x_1) = SS(x_2|x_1)/1$, is compared with $MS_{Res}(2)$, the residual mean square for the more complex model, by means of an F test

$$F = \frac{MS(x_2|x_1)}{MS_{Res}(2)} \tag{10.10}$$

This F test is called a *partial F-test* and is important for the selection of predictor variables in the stepwise regression procedures described in Section 10.3.3. What is tested here is the significance of the regression coefficient β_2, when β_1 is already in the model. Therefore $SS(x_2|x_1)$ is also represented as $SS(b_2|b_1)$. The significance of a regression coefficient (H_0: $\beta_i = 0$; H_1: $\beta_i \neq 0$) can also be obtained from a t-test (see Sections 8.2.4 and 10.4):

$t = b_i / s_{b_i}$

with b_i: the estimate of the ith regression coefficient, β_i

s_{b_i}: the standard deviation of the estimated regression coefficient b_i, which can be obtained from eq. (10.18).

It can be shown that the square of this t-value with $n - p$ degrees of freedom is equal to the partial F value which has 1 and $n - p$ degrees of freedom.

As an example consider Table 10.7 in which the results of the sequential fitting of x_1, x_2 and x_3 are given for the stack loss data. From this it is evident that the addition of x_2 if x_1 is already present is useful since it results in a significant increase of the regression sum of squares. Since the total sum of squares remains constant this also means that a significant reduction of the residual sum of squares is observed. On the other hand, the addition of x_3 if x_1 and x_2 are already present does not significantly improve the model.

Notice that the addition of a new x variable changes the estimates of all the other regression coefficients. This is due to the correlations among the independent

TABLE 10.7

Stack loss data. Results of the sequential fitting of x_1, x_2 and x_3

1. Fitting x_1 $\quad \hat{y} = -40.033 + 0.944\, x_1$

ANOVA

Source	df	SS	MS	F
Regression (x_1)	1	775.482	775.482	285.43
Residual	15	40.753	2.717	
Total		816.235		

2. Addition of x_2 $\quad \hat{y} = -42.001 + 0.777\, x_1 + 0.569\, x_2$

ANOVA

Source	df	SS	MS	F
Regression (x_1,x_2)	2	793.975	396.987	249.68
Residual	14	22.260	1.590	
Total		816.235		

$$\text{partial } F = \frac{SS(x_2|x_1)/1}{MS_{Res}} = \frac{793.975 - 775.482}{1.590} = 11.63$$

$$> F_{0.05;1,14} (= 4.60)$$

3. Addition of x_3 $\quad \hat{y} = -37.652 + 0.798\, x_1 + 0.577\, x_2 - 0.067\, x_3$

ANOVA

Source	df	SS	MS	F
Regression (x_1,x_2,x_3)	3	795.834	265.278	169.04
Residual	13	20.401	1.569	
Total		816.235		

$$\text{partial } F = \frac{SS(x_3|x_1,x_2)/1}{MS_{Res}} = \frac{795.834 - 793.975}{1.569} = 1.18$$

$$< F_{0.05;1,13} (= 4.67)$$

variables (see Table 10.9). Without these correlations the addition of a variable would not influence the coefficients of the variables already in the model.

The information of Table 10.7 can be used to partition the regression sum of squares of the model, including the three variables, into the individual contributions of the different variables: the contribution of x_1, x_2 and x_3 is 775.482, 18.493 (= 793.975 – 775.482) and 1.859 (= 795.834 – 793.975), respectively, when they are entered in that order. This is summarized in Table 10.8 together with the results

TABLE 10.8

Stack loss ANOVA data. Effect of the order in which the x variables are entered into the model.

Source	df	SS	MS	F
Regression x_1,x_2,x_3				
due to x_1	1	775.482	775.482	285.42
residual	15	40.753	2.717	
due to $x_2\|x_1$	1	18.493	18.493	11.63
residual	14	22.260	1.590	
due to $x_3\|x_1,x_2$	1	1.859	1.859	1.18*
residual	13	20.401	1.569	
Total	16	816.235		
Regression x_2,x_1,x_3				
due to x_2	1	567.032	567.032	34.13
residual	15	249.203	16.614	
due to $x_1\|x_2$	1	226.943	226.943	142.73
residual	14	22.260	1.590	
due to $x_3\|x_2,x_1$	1	1.859	1.859	1.18*
residual	13	20.401	1.569	
Total	16	816.235		
Regression x_3,x_2,x_1				
due to x_3	1	134.799	134.799	2.97*
residual	15	681.436	45.429	
due to $x_2\|x_3$	1	441.480	441.480	25.76
residual	14	239.956	17.140	
due to $x_1\|x_3,x_2$	1	219.555	219.555	139.93
residual	13	20.401	1.569	
Total	16	816.235		

*Not significant at 5% significance level.

TABLE 10.9

Correlation matrix for the stack loss data

	x_1	x_2	x_3	y
x_1	1.000	0.754	0.454	0.975
x_2		1.000	0.369	0.833
x_3			1.000	0.406
y				1.000

obtained for the regression in which x_2 is first entered followed by x_1 and then x_3 and also for the regression in which the order of entering is x_3, x_2 and x_1. From this table it follows that the contribution of the different x variables in increasing the regression sum of squares depends on the order in which the variables are introduced into the model. For example the contribution of x_2 is much larger when it is introduced first (567.032) than when it is added after x_1 (18.493). This is due to the relatively high correlation between x_1 and x_2 as follows from the correlation matrix given in Table 10.9. Therefore, if x_1 is already in the regression it explains part of the variation in y that could also be accounted for by x_2. Consequently, the contribution of x_2 in the SS_{Reg} drops when it is added in second place. Nevertheless, x_1 and x_2 are important variables since, whatever the order of introduction, they have a significant contribution in increasing the regression sum of squares. On the other hand, x_3 is not important since it does not significantly contribute to the variation in y.

10.3.3 Selection of predictor variables

The discussion of the previous section brings us to the problem of the selection of the predictor variables: which variables should be used in the regression equation? The most complete approach is to compare *all possible regressions* performed on the m variables. This means that all regression equations with only one variable, with two variables, up to the regression equation including all m variables are fitted.

Several related criteria for the comparison of the $2^m - 1$ different regression equations that are obtained in this way have been proposed such as the value of R^2 (eq. (10.9)) or of the residual mean square (MS_{Res}). In the comparison of models with a different number of x variables R^2 should of course be used with caution (see Section 10.3.1.2). Therefore to compare different regression equations the adjusted R^2 which takes into account the degrees of freedom associated with the sums of squares (SS_{Res} and SS_T) in the expression for R^2 is generally preferred:

$$R_a^2 = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)} = 1 - (1-R^2)\left(\frac{n-1}{n-p}\right) \tag{10.11}$$

A statistic which is related to R_a^2 is the Mallows C_p statistic:

$$C_p = \frac{SS_{Res(p)}}{s^2} - (n-2p) \tag{10.12}$$

with $SS_{Res(p)}$: the residual sum of squares for the model with p parameters ($= MS_{Res(p)}\,(n-p)$)
s^2: an estimate of the experimental error σ^2, e.g. obtained from the residual mean square of the model containing all parameters

The form of the fitted model is adequate if $C_p \approx p$.

Example:

The comparison of all possible regressions for the stack loss data by means of MS_{Res}, R_a^2 and C_p is summarized in Table 10.10. It follows that the best equation (lowest residual variance and highest coefficient of multiple determination) is the one in which all three variables are included:

$$y = -37.652 + 0.798x_1 + 0.577x_2 - 0.067x_3$$

However in that equation the contribution of x_3 is not significant. The elimination of that variable results in a simpler equation with a very similar MS_{Res} (1.59 vs. 1.57) and R_a^2 (96.88 vs. 96.92) and a C_p value, 3.19, close to 3. Therefore the equation

$$y = -42.001 + 0.777x_1 + 0.569x_2$$

is to be preferred.

Of course with a large number of x variables the comparison of all possible regression equations requires a lot of computation and therefore other procedures

TABLE 10.10

Comparison of the quality of all possible regressions for the stack loss data

p	Variables in the equation	MS_{Res}	$100\,R^2$	$100\,R_a^2$	C_p	Variables with a significant contribution to the regression
2	x_1	2.717	95.01	94.67	12.98	x_1
2	x_2	16.614	69.47	67.43	145.83	x_2
2	x_3	45.429	16.51	10.95	421.31	/
3	x_1,x_2	1.590	97.27	96.88	3.19	x_1,x_2
3	x_2,x_3	17.140	70.60	66.40	141.94	x_2
3	x_1,x_3	2.814	95.17	94.48	14.11	x_1
4	x_1,x_2,x_3	1.569	97.50	96.92	4.00	x_1,x_2

are generally preferred over this brute force approach. These are the forward, the backward and the stepwise procedure. It is important to realize that these methods will identify an acceptable model which is not necessarily the best one.

In the *forward selection procedure* the predictor variables are entered one at a time. At each step that variable is added that produces the largest significant increase in the regression sum of squares. The selection starts with the variable that has the largest (positive or negative) correlation with the dependent variable. If this variable results in a significant regression, as judged from the overall F-test, the variable is retained and the selection continues. This means that for all variables, not yet in the equation, the partial F-test of eq. (10.10) is performed. In the forward selection procedure this F-test is called *F-to-enter* and is defined as the partial F-test performed on a variable which is not yet in the regression equation. The variable that results in the largest significant increase of the SS_{Reg} (largest significant F-to-enter value) is then added. This procedure continues until none of the variables left significantly contribute to the regression sum of squares.

The *backward elimination procedure* starts with all the predictor variables in the equation and removes the least important variables one at a time. The criterion for removal is again based on the partial F-test of eq. (10.10). In the backward elimination procedure this F-test is called *F-to-remove* and is defined as the partial F-test performed on a variable already in the equation as though it was added last to the model. In other words, at each step it is checked for each variable of the model whether it significantly contributes to the regression sum of squares, if it were the last variable added to the model. The variable that results in the smallest non-significant increase of the SS_{Reg} (the smallest non-significant F-to-remove value) is dropped. This procedure continues until all the variables not yet dropped significantly contribute to the regression sum of squares as judged from their F-to-remove value being significant.

The forward and the backward procedure do not necessarily lead to the same regression equation when the predictor variables are correlated. This is because a variable that is entered in the forward selection remains in the model, even if after the addition of other correlated variables its contribution may have dropped significantly. Similarly, a variable deleted in the backward elimination is lost even if after the elimination of other variables it might become an important variable.

Therefore the *stepwise regression procedure*, which combines the forward and backward approach, is generally preferred. At each step the F-to-enter values for all variables not yet in the equation are checked and the variable with the highest significant F value is entered. After each step the F-to-remove values for all variables already in the equation are tested. If a variable is detected that does no longer significantly contribute to the regression it is rejected. The procedure is continued until no more variables fulfil the criterion to be entered or to be removed.

TABLE 10.11

Stack loss data. Regression sum of squares, residual sum of squares and % variation explained for different regression equations.

Variables in the equation	Regression		Residual		$100\,R^2$
	SS	df	SS	df	
x_1	775.482	1	40.753	15	95.01
x_2	567.032	1	249.203	15	69.47
x_3	134.799	1	681.436	15	16.51
x_1 and x_2	793.975	2	22.260	14	97.27
x_1 and x_3	776.845	2	39.390	14	95.17
x_2 and x_3	576.279	2	239.956	14	70.60
x_1, x_2 and x_3	795.834	3	20.401	13	97.50

Example:

The results of the stepwise regression performed on the stack loss data are given as an example. The information necessary for the calculations is summarized in Table 10.11.

Step 1:

Since from the correlation matrix (Table 10.9) it follows that the response variable y is most correlated with x_1, that variable is the first to enter the regression equation. For the variables not in the regression the following F-to-enter values are calculated:

x_2: F-to-enter = (793.975 – 775.482) / (22.260/14)
$= 11.63 > F_{0.05;1,14}\ (= 4.60)$

x_3: F-to-enter = (776.845 – 775.482) / (39.390/14)
$= 0.48 < F_{0.05;1,14}\ (= 4.60)$

It follows that x_2 has the highest F-to-enter value. Since it contributes significantly to the regression x_2 is added to the equation.

Step 2:

For the variables in the regression equation (x_1 and x_2) the following F-to-remove values are calculated:

x_1: F-to-remove = (793.975 – 567.032)/(22.260/14)
$= 142.73 > F_{0.05;1,14}\ (= 4.60)$

x_2: F-to-remove = (793.975 – 775.482)/(22.260/14)
$= 11.63 > F_{0.05;1,14}\ (= 4.60)$

The smallest F-to-remove value is observed for x_2. Since it is significant, x_2 is retained in the equation. Of course at this stage of the analysis, with only two variables entered, no other result could be expected. The F-to-remove for x_1 for example cannot be smaller than the F-to-remove for x_2 because that would mean that in the first step, x_2 was the first variable entered. For the variables not in the equation the following F-to-enter values are calculated:

$$x_3\text{: } F\text{-to-enter} = (795.834 - 793.975)/(20.401/13)$$
$$= 1.18 < F_{0.05;1,13} (= 4.67)$$

x_3 which is the only variable left does not significantly improve the regression. Consequently it is not included. Since in a further step no variables can be added or deleted the procedure stops and the final regression equation is:

$$\hat{y} = -42.001 + 0.777x_1 + 0.569x_2$$

In this simple example the stepwise regression procedure happens to yield the same model as the evaluation of all possible regressions. Dagnelie [3] describes an example that shows that this is not always the case.

Variable or *feature selection* can also be performed by means of genetic algorithms described in Chapter 27.

10.3.4 Validation of the prediction performance of the model

It is important to realize that during the modelling as described up to now, the validation has been performed with the data used to construct the model. However if the model has been built for prediction purposes it is of paramount importance to extend the validation to new, independent data. This means that new experiments are performed and that the actual observations are compared with the predictions from the model. If new data can not be obtained an alternative approach known as *cross-validation* can be used. The data set at hand is split into subsets, one subset, the *estimation set* or *training set*, being used to build the model and the other, the *prediction set* or *test set*, to validate the model i.e. to measure the prediction accuracy of the model.

There are different possibilities of splitting the data. In the *leave-one-out,* the first observation is deleted from the data set and is predicted from the model fitted to the remaining $n - 1$ data points. The residual $(y_1 - \hat{y}_{-1})$, which will be called the deleted residual, is calculated (the index -1 refers to the fact that the prediction is from a model built without the first observation). This is repeated for all data points and the *predicted residual error sum of squares* (PRESS) or the *root mean squared prediction error* (RMSPE) is calculated:

$$\text{PRESS} = \sum(y_i - \hat{y}_{-i})^2 \tag{10.13}$$

$$\text{RMSPE} = (\text{PRESS}/n)^{1/2} \tag{10.14}$$

TABLE 10.12

PRESS values for all different models applied to the stack loss data

Variables in the equation	PRESS
x_1	62
x_2	424
x_3	804
x_1,x_2	43
x_2,x_3	458
x_1,x_3	66
x_1,x_2,x_3	43

TABLE 10.13

The deleted residuals from the model $\hat{y} = b_0 + b_1x_1 + b_2x_2$ for the stack loss data

Observation predicted*	Deleted residual
1	4.1
2	–0.7
3	–1.4
4	–1.0
5	0.2
6	–1.4
7	0.9
8	0.9
9	0.4
10	–2.7
11	–2.0
12	1.1
13	–0.1
14	0.4
15	0.4
16	1.0
17	2.3

*The observation is predicted from the model developed with the other 16 data.

PRESS or RMSPE are especially useful in comparing the prediction errors of different regression models. Table 10.12 summarizes the PRESS values for all different models applied to the stack loss data. The model including x_1 and x_2, which was the model selected from the evaluation made in the previous section, is the best model for prediction purposes. An identical PRESS value is obtained with the model including all three variables but the simpler model is to be preferred.

For the different observations the deleted residuals from the best predictive model are listed in Table 10.13. It is interesting to note that the first data point

possesses the largest residual (= 4.1) and consequently contributes most to the prediction error sum of squares. Inspection of the first observation reveals that, with $x_1 = 80$ and $x_2 = 87$, it is a leverage point (see Section 8.2.6 and 10.9) since it is remote from the rest of the data. Being alone outside the domain of the model built with the other observations, it is the worst predicted. This certainly does not mean that it is a bad point, on the contrary it provides useful information concerning the model fitted. For example, it would certainly be useful to include more observations spanning the whole region in which predictions are to be performed. However, with unplanned data such as the stack loss data which are obtained from successive days of operation of a plant, this may be difficult to achieve. When possible, experimental design procedures (see e.g. Chapter 21), which define predetermined settings of the predictor variables, should be used since they allow us among others to obtain balanced data describing the whole domain of interest.

10.4 Confidence intervals

The 95% confidence intervals for the true regression parameters, β_i, are obtained from

$$b_i \pm t_{0.025,n-p}\, s_{b_i} \qquad (i = 0 \ldots m) \tag{10.15}$$

These confidence intervals can also be used to check the significance of the corresponding regression coefficient. If the confidence interval includes the value zero, β_i can be zero and consequently the regression coefficient is not significant at the 5% significance level. This can of course also be checked by means of a t-test in the usual way by calculating

$$t = b_i / s_{b_i} \tag{10.16}$$

A joint $100\,(1 - \alpha)\%$ confidence region for all the regression parameters β_i that takes into account the correlation between these parameters, can be obtained from

$$(\boldsymbol{\beta} - \mathbf{b})^{\mathrm{T}}\, \mathbf{X}^{\mathrm{T}}\, \mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) \leq p s_e^2\, F_{(\alpha;p,n-p)} \tag{10.17}$$

It is a generalization of eq. (8.17) to multiple regression with p regression coefficients and represents the equation of an ellipsoid in p dimensions. Since with increasing p, the interpretation is not straightforward, the joint confidence region is less used in multiple regression.

The variances of the different parameters, $(s_{b_i})^2$, necessary to determine the confidence intervals are obtained from the *variance–covariance matrix* (see Chapter 9) *of the regression coefficients*, $\mathbf{V}(b)$

$$\mathbf{V}(b) = \begin{bmatrix} (s_{b_0})^2 & \text{cov}(b_0,b_1) & \dots & \text{cov}(b_0,b_m) \\ \text{cov}(b_1,b_0) & (s_{b_1})^2 & \dots & \text{cov}(b_1,b_m) \\ . & . & & . \\ . & . & & . \\ \text{cov}(b_m,b_0) & . & \cdots & (s_{b_m})^2 \end{bmatrix}$$

This is a symmetric matrix in which the diagonal elements are the variances of the regression parameters in the same order as they appear in the regression equation. The off-diagonal elements are the covariances between the regression parameters.

It can be shown [1] that $\mathbf{V}(b)$ is given by

$$\mathbf{V}(b) = s_e^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{10.18}$$

with s_e^2 an estimate of the pure experimental error (eq. (10.8)).

This is an important expression, indicating the influence of the $(\mathbf{X}^T\mathbf{X})^{-1}$ matrix on the variance of the regression parameters. It means that the confidence intervals will largely depend on the experimental design used, thus among others on the range considered for the different x variables (i.e. the experimental domain), the distribution of the x values over the experimental domain and the number of measurements. In Chapter 24 different criteria, based on this matrix, are discussed for the evaluation of experimental designs.

If the model is adequate the 95% confidence interval for η, the true mean value of y given a specific combination of the controlled variables, $\mathbf{x}_0$, is obtained from

$$\hat{y}_0 \pm t_{0.025,n-p}\, s_e \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0} \tag{10.19}$$

with $\mathbf{x}_0^T = [1\ x_{01} \dots x_{0m}]$.

If the objective is to predict the mean of g replicate observations at a given combination of the controlled variables $\mathbf{x}_0$, the following expression for the 95% confidence interval should be used

$$\hat{y}_0 \pm t_{0.025,n-p}\, s_e \sqrt{\frac{1}{g} + \mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_0} \tag{10.20}$$

Example:

From Section 10.3.3 the following regression equation $\hat{y} = -42.001 + 0.777\, x_1 + 0.569\, x_2$ was obtained for the stack loss data. The variance–covariance matrix of the parameters is

$$\mathbf{V}(b) = s_e^2(\mathbf{X}^T\mathbf{X})^{-1} = 1.590 \begin{bmatrix} 4.099030468 & -0.048535765 & -0.060580395 \\ -0.048535765 & 0.002656869 & -0.005141029 \\ -0.060580395 & -0.005141029 & 0.017516880 \end{bmatrix}$$

$$= \begin{bmatrix} 6.51746 & -0.07717 & -0.09632 \\ -0.07717 & 0.00422 & -0.00817 \\ -0.09632 & -0.00817 & 0.02785 \end{bmatrix}$$

The 95% confidence interval for β_0 is therefore ($t_{0.025,14} = 2.145$)

$$-42.001 \pm 2.145\sqrt{6.51746} = -42.001 \pm 5.476$$

The 95% confidence interval for β_1 and β_2 are found to be respectively 0.777 ± 0.139 and 0.569 ± 0.358.

To obtain, from eq. (10.19), the 95% confidence interval for η, the true mean of y, at for example

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ 62 \\ 24 \end{bmatrix}$$

the following information is necessary

$$\hat{y}_0 = -42.001 + 0.777\,(62) + 0.569\,(24) = 19.829$$

$$s_e = \sqrt{1.590}$$

$$\sqrt{\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0} =$$

$$\sqrt{[1 \;\; 62 \;\; 24] \begin{bmatrix} 4.099030468 & -0.048535765 & -0.060580395 \\ -0.048535765 & 0.002656869 & -0.005141029 \\ -0.060580395 & -0.005141029 & 0.017516880 \end{bmatrix} \begin{bmatrix} 1 \\ 62 \\ 24 \end{bmatrix}} = 0.419$$

The 95% confidence interval therefore is 19.8 ± 1.1.

10.5 Multicollinearity

To obtain the regression parameters from eq. (10.6) we need to invert the matrix $\mathbf{X}^T\mathbf{X}$. This inverse only exists if the matrix is non-singular, that is if the determinant of $(\mathbf{X}^T\mathbf{X})$ is not zero (see Chapter 9). Singularity ($\det(\mathbf{X}^T\mathbf{X}) = 0$) occurs if any of the independent variables is a perfect linear combination of other independent variables. This means that some of the normal equations given by eq. (10.5) can be exactly expressed as linear combinations of others. Therefore fewer equations are available than there are unknowns and no unique solution can be obtained.

Consider for example the following $\mathbf{X}$ matrix in which x_1 and x_2 are perfectly correlated since $x_2 = 2x_1$:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \\ 1 & 4 & 8 \\ 1 & 5 & 10 \end{bmatrix}$$

The $\mathbf{X}^T\mathbf{X}$ matrix is given by

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 5 & 15 & 30 \\ 15 & 55 & 110 \\ 30 & 110 & 220 \end{bmatrix}$$

Since the determinant of $\mathbf{X}^T\mathbf{X}$ is zero the calculation of the inverse of this matrix is not possible.

The effect of the correlation between x_1 and x_2 can also be evaluated from the normal equations. With the following response vector:

$$\mathbf{y} = \begin{bmatrix} 10 \\ 20 \\ 30 \\ 40 \\ 50 \end{bmatrix}$$

to fit the equation $\hat{y} = b_0 + b_1x_1 + b_2x_2$, which is represented by a plane, eq. (10.5) becomes:

$$5\,b_0 + 15\,b_1 + 30\,b_2 = 150$$

$$15\,b_0 + 55\,b_1 + 110\,b_2 = 550$$

$$30\,b_0 + 110\,b_1 + 220\,b_2 = 1100$$

There are 3 equations with 3 unknowns but since the last equation is simply twice the second, it does not give us independent information. Consequently no unique solution can be generated from these equations. This is also shown in Fig. 10.2 from which it becomes evident that due to the perfect correlation between x_1 and x_2, an infinite number of planes fit these data equally well. This problem of *multicollinearity* can be solved by reducing the number of x variables.

Situations in which the determinant is not zero but is very small (because some variables are almost linear combinations of other independent variables) are more common and result in an *ill-conditioned* $\mathbf{X}^T\mathbf{X}$ matrix. This leads to unstable estimates of the regression coefficients which may be unreasonably large (in absolute value) or have the wrong sign. This is also reflected in their large variances (see eq. (10.18)). Highly correlated x variables therefore easily lead to unreliable predictions. Obviously in regression the $\mathbf{X}^T\mathbf{X}$ matrix is an important matrix (see also Section 10.4) and, as already mentioned, in experimental design (see Chapter 24) it will play an important role in the evaluation of the design of the experiments.

A useful indicator of the interdependency among the x variables is the *tolerance* which for each x_i variable can be calculated as

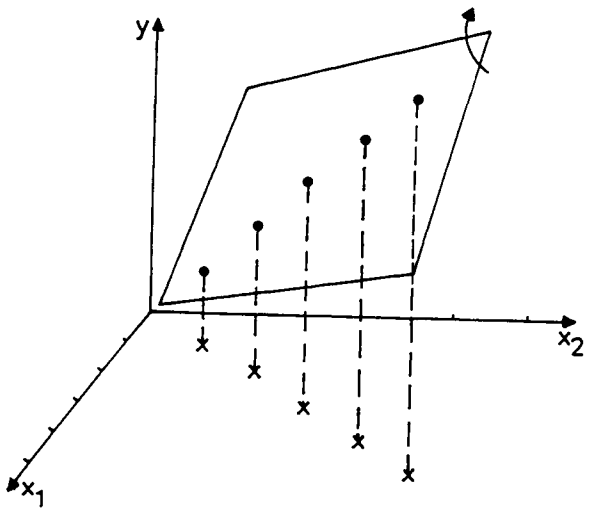


Fig. 10.2. Illustration of the problem of multicollinearity due to the perfect correlation between x_1 and x_2 ($x_2 = 2x_1$).

$$\text{Tolerance } (x_i) = 1 - R^2(x_i) \tag{10.21}$$

where $R^2(x_i)$ is the coefficient of multiple determination (see Section 10.3.1.2) for the regression between x_i (considered here as the dependent variable) and the other independent variables. Since $R^2(x_i)$ then represents the variation in x_i that can be explained by the other x variables, a small tolerance (large $R^2(x_i)$) means that x_i is almost a linear combination of the other x variables.

A related indicator of multicollinearity is the *variance inflation factor* (VIF) which is the reciprocal of the tolerance:

$$\text{VIF}(x_i) = \frac{1}{1 - R^2(x_i)} \tag{10.22}$$

The larger the variance inflation factor, the larger the variance of the regression coefficient. The latter can also be obtained from the following expression [4]:

$$s^2_{b_i} = \frac{s^2_e}{(1 - R^2(x_i))\,(n-1)\,s^2_i}$$

with s^2_e and $(1 - R^2(x_i))$ as defined by eqs. (10.8) and (10.21), respectively, and s^2_i the variance of the ith x variable. A VIF larger than 5 or 10 is generally considered large [5] and is an indication that the corresponding coefficient is poorly estimated.

10.6 Ridge regression

Various alternative regression procedures have been described for the analysis of data in which the predictor variables are highly correlated such as principal component regression, partial least squares regression (see Chapter 36) and ridge regression. The regression coefficients in the *ridge regression* procedure are obtained from:

$$(\mathbf{X}^T\mathbf{X} + k\mathbf{I})\mathbf{b} = \mathbf{X}^T\mathbf{y} \tag{10.23}$$

where $\mathbf{X}$ is the $n \times p$ matrix of the standardized x variables (see Section 31.3), k is a positive number (usually $0 < k < 1$) and $\mathbf{I}$ is the $p \times p$ identity matrix. Comparison of this expression with eq. (10.5) reveals that a constant is added to the diagonal elements of the $\mathbf{X}^T\mathbf{X}$ matrix of the normal equations. With $k = 0$ the least squares solution is obtained since eq. (10.23) then reduces to eq. (10.5). As a result of the addition of the constant k, biased estimates of the regression coefficients are obtained in ridge regression. The estimates of the regression coefficients, $\mathbf{b}$, are not biased if the mean of the sampling distribution of $\mathbf{b}$ (obtained by estimating β repeatedly at the same values of the x variables) is equal to the true regression coefficients, β. (Notice that it can be shown [1] that classical least squares multiple regression also results in biased regression coefficients if, by eliminating x variables, the fitted model differs from the true model). The constant k is therefore known as the *bias parameter* or the *ridge coefficient*. In ridge regression some bias is introduced in order to increase the stability of the regression coefficients. With increasing k values the bias in the estimates increases but their variance largely decreases. The residual sum of squares, SS_{Res}, also increases with increasing k; consequently R^2 decreases. Hoerl and Kennard [6] suggest selecting a value of k by an examination of a *ridge trace*, which is a plot of the regression coefficients for different values of the bias parameter. At the value of k chosen the regression coefficients should have started to stabilize, they should have the proper sign, and the reduction in R^2 should not be too large. The latter can be evaluated from a plot of R^2 against different k values.

Example:

Consider as an example the simulated data in Table 10.14 which have been adapted from Hoerl [7] and for which the true relationship is

$$\eta = 100 + 2x_1 + 3x_2 + 5x_3$$

The least squares regression results and the variance inflation factors for the different x variables are summarized in Table 10.15. The least squares regression equation is

TABLE 10.14
Hoerl data [7]

x_1	x_2	x_3	y
11	11	12	223
14	15	11	223
17	18	20	292
17	17	18	270
18	19	18	285
18	18	19	304
19	18	20	311
20	21	21	314
23	24	25	328
25	25	24	340

TABLE 10.15
Least squares results for the Hoerl data

Variable	b_i	s_{b_i}	VIF
x_1	8.266	5.322	41.92
x_2	−5.516	4.732	34.24
x_3	6.386	2.019	7.53
Constant	121.117	15.388	

$$\hat{y} = 121.12 + 8.27x_1 - 5.52x_2 + 6.39x_3$$

The VIFs, and especially those for x_1 and x_2 are large due to the high correlation between the x variables. Consequently the associated coefficients are poorly estimated, their variance is large and b_2 has the wrong sign. The application of ridge regression with different k values results in the ridge trace as given in Fig. 10.3. A plot of R^2 as a function of k is shown in Fig. 10.4. At a value of $k = 0.15$ the regression coefficients stabilize and the reduction in R^2 is not very large. The ridge regression equation with $k = 0.15$ is

$$\hat{y} = 126.29 + 2.95x_1 + 1.46x_2 + 4.36x_3$$

which agrees much better with the true model than the least squares solution. Moreover, the estimated regression coefficients are more stable ($s_{b_1} = 0.77$; $s_{b_2} = 0.85$; $s_{b_3} = 1.01$).

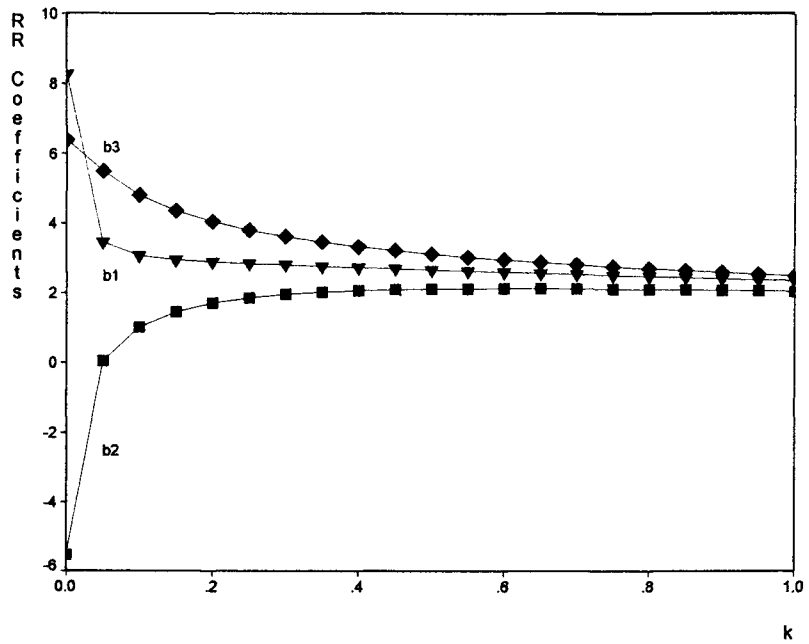


Fig. 10.3. Ridge trace for the Hoerl data.

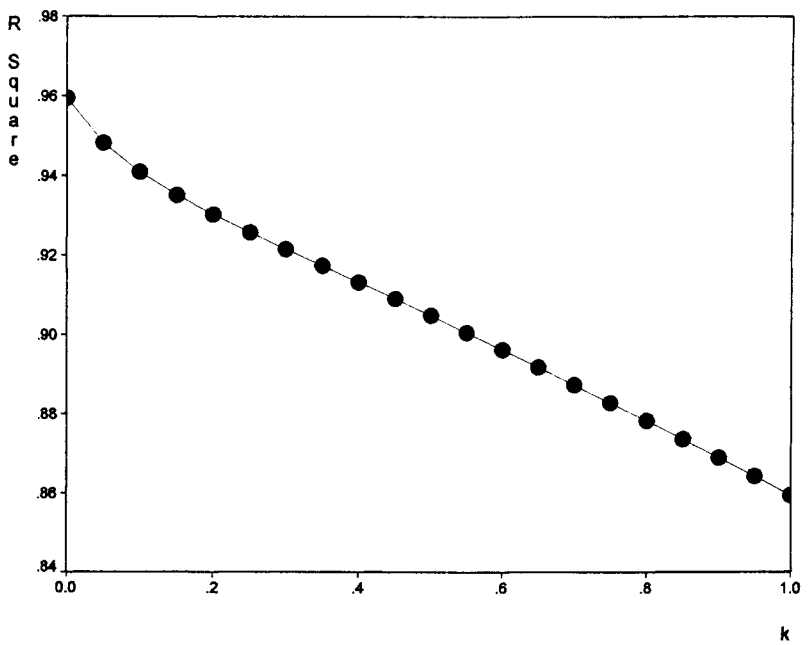


Fig. 10.4. Coefficient of multiple correlation, R^2, as a function of the bias parameter k.

10.7 Multicomponent analysis by multiple linear regression

The term multicomponent analysis is used for procedures in which several components in a sample are determined simultaneously. For the analysis of an m-component mixture at least $n = m$ measurements are required. Linearity in the sense of straight line relationships and additivity of the signals is assumed. When $n = m$, a so-called *exactly determined* system is obtained; when $n > m$ which means that the number of measurements made is larger than the number of components, the system is *over-determined.* In general, it can be expected that the precision of the procedure increases with an increasing number of measurements. To some extent, the effect of using an over-determined system is the same as the effect of repeated measurements on the precision.

The contribution to the signal of each analyte at a given sensor (e.g. a wavelength in UV-visible spectrometry) is weighted by the sensitivity coefficients, k_j ($j = 1, 2, \ldots m$), of each analyte (in spectrometry, k_{ij} is the absorptivity of component j at wavelength i). For the spectrometric analysis of an m-component mixture, for example, for which measurements at n ($n \geq m$) wavelengths are performed the absorbances are:

$$A_1 = k_{11}\, c_1 + k_{12}\, c_2 + \ldots + k_{1m}\, c_m$$

$$A_2 = k_{21}\, c_1 + k_{22}\, c_2 + \ldots + k_{2m}\, c_m$$

. . . .

. . . .

$$A_n = k_{n1}\, c_1 + k_{n2}\, c_2 + \ldots + k_{nm}\, c_m$$

This set of equations can be written in matrix notation as:

$$\mathbf{a} = \mathbf{K}\,\mathbf{c} \tag{10.24}$$

with $\mathbf{a}$ the vector of the absorbances measured at n wavelengths, $\mathbf{c}$ the concentration vector for the m components and $\mathbf{K}$ the $(n \times m)$ absorptivity matrix.

If the $\mathbf{K}$ matrix is known, the concentrations of the components in the mixture can be obtained from:

$$\mathbf{c} = (\mathbf{K}^T\mathbf{K})^{-1}\,\mathbf{K}^T\,\mathbf{a} \tag{10.25}$$

Notice the similarity between this equation and eq. (10.6). It corresponds to the least-squares solution in which the elements of the $\mathbf{K}$ matrix are treated as the independent variables.

The elements of the $\mathbf{K}$ matrix, which are the absorptivities of the m components at the n wavelengths, can be obtained from the spectra of the pure components. Alternatively, as explained in Chapter 36, they can be estimated by multivariate

calibration methods that relate the known concentration of calibration mixtures to the measured calibration spectra. In Section 36.2.1 the limitations of multicomponent analysis by multiple linear regression are discussed.

Example:

The absorptivities of Cl_2 and Br_2 in chloroform at six wavenumbers are given in Table 10.16 [8]. For an optical path length of 1 cm, and concentrations c_1 and c_2 of Cl_2 and Br_2, respectively, the measurements A_1, A_2, A_3, ..., A_6 are obtained at the wavenumbers (22, 24, 26, 28, 30, 32×10^3 cm^{-1})

$A_1 = 4.5\ c_1 + 168\ c_2 = 34.10$

$A_2 = 8.4\ c_1 + 211\ c_2 = 42.95$

$A_3 = 20\ c_1 + 158\ c_2 = 33.55$

$A_4 = 56\ c_1 + 30\ c_2 = 11.70$

$A_5 = 100\ c_1 + 4.7\ c_2 = 11.00$

$A_6 = 71\ c_1 + 5.3\ c_2 = 7.98$

The concentrations c_1, c_2 are given by eq. (10.25), which becomes

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \left(\begin{bmatrix} 4.5 & 8.4 & 20 & 56 & 100 & 71 \\ 168 & 211 & 158 & 30 & 4.7 & 5.3 \end{bmatrix} \begin{bmatrix} 4.5 & 168 \\ 8.4 & 211 \\ 20 & 158 \\ 56 & 30 \\ 100 & 4.7 \\ 71 & 5.3 \end{bmatrix} \right)^{-1}$$

$$\times \begin{bmatrix} 4.5 & 8.4 & 20 & 56 & 100 & 71 \\ 168 & 211 & 158 & 30 & 4.7 & 5.3 \end{bmatrix} \begin{bmatrix} 34.10 \\ 42.95 \\ 33.55 \\ 11.70 \\ 11.00 \\ 7.98 \end{bmatrix}$$

This gives

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 18667.81 & 8214.7 \\ 8214.7 & 98659.18 \end{bmatrix}^{-1} \begin{bmatrix} 4.5 & 8.4 & 20 & 56 & 100 & 71 \\ 168 & 211 & 158 & 30 & 4.7 & 5.3 \end{bmatrix} \begin{bmatrix} 34.10 \\ 42.95 \\ 33.55 \\ 11.70 \\ 11.00 \\ 7.98 \end{bmatrix}$$

TABLE 10.16

Absorptivities of Cl_2 and Br_2 in chloroform (from Ref. [8])

Wavenumber ($cm^{-1} \times 10^3$)	Absorptivities		Absorbance of mixture (simulated)
	Cl_2	Br_2	
22	4.5	168	34.10
24	8.4	211	42.95
26	20	158	33.55
28	56	30	11.70
30	100	4.7	11.00
32	71	5.3	7.98

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 5.56\times10^{-5} & -4.63\times10^{-6} \\ -4.63\times10^{-6} & 1.05\times10^{-5} \end{bmatrix}^{-1} \begin{bmatrix} 4.5 & 8.4 & 20 & 56 & 100 & 71 \\ 168 & 211 & 158 & 30 & 4.7 & 5.3 \end{bmatrix} \begin{bmatrix} 34.10 \\ 42.95 \\ 33.55 \\ 11.70 \\ 11.00 \\ 7.98 \end{bmatrix}$$

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} -0.00053 & -0.00051 & 0.00038 & 0.00298 & 0.00554 & 0.00392 \\ 0.00174 & 0.00218 & 0.00157 & 0.00006 & -0.00041 & -0.00027 \end{bmatrix} \begin{bmatrix} 34.10 \\ 42.95 \\ 33.55 \\ 11.70 \\ 11.00 \\ 7.98 \end{bmatrix}$$

$c_1 = 0.099241$

$c_2 = 0.199843$

By analogy with eq. (10.18) the variance of the concentrations is

$$\mathbf{V}(c) = s_e^2 (\mathbf{K}^T\mathbf{K})^{-1}$$

In fact, the term $(\mathbf{K}^T\mathbf{K})^{-1}$ in this equation gives the error amplification of the measurement error into the analytical result, $\mathbf{V}(c)$. The most important conclusion is that the error propagation depends on the choice of the wavelengths in multicomponent analysis (the $\mathbf{K}$ matrix i.e. the design of the calibration).

The absorbances predicted by the model, $\hat{A}_i$, and the residuals ($e_i = A_i - \hat{A}_i$ are tabulated in Table 10.17. Consequently

TABLE 10.17

Predicted absorbances and residuals for the data of Table 10.16

Wavenumber ($cm^{-1} \times 10^3$)	$\hat{A}$	$A - \hat{A}$	$(A - \hat{A})^2$
22	34.02	0.08	6.4×10^{-3}
24	43.01	–0.06	3.6×10^{-3}
26	33.57	–0.02	0.4×10^{-3}
28	11.59	0.11	12×10^{-3}
30	10.93	0.07	4.9×10^{-3}
32	8.15	–0.17	29×10^{-3}
			sum = 0.0563

$$s_e^2 = \frac{\Sigma(A_i - \hat{A}_i)^2}{n - m} = 0.0563 / (6 - 2) = 1.41 \times 10^{-2}$$

where n is the number of measurements (wavelengths) and m the number of analytes. Moreover

$$(s_{c_1})^2 = 1_{11} s_e^2$$

$$(s_{c_2})^2 = 1_{22}\, s_e^2$$

where 1_{11} and 1_{22} are the corresponding diagonal elements of the $(\mathbf{K}^T\mathbf{K})^{-1}$ matrix. Therefore

$$(s_{c_1})^2 = (5.56 \times 10^{-5})\,(1.41 \times 10^{-2}) = 7.84 \times 10^{-7}$$

$$(s_{c_2})^2 = (1.05 \times 10^{-5})\,(1.41 \times 10^{-2}) = 1.48 \times 10^{-7}$$

The 95% confidence limits of the true concentrations are

$$c_1 \pm t_{0.025,4} \sqrt{(s_{c_1})^2} = 0.099241 \pm 2.776 \sqrt{7.84 \times 10^{-7}}$$
$$= 0.099 \pm 0.0025$$

$$c_2 \pm t_{0.025,4} \sqrt{(s_{c_2})^2} = 0.199843 \pm 2.776 \sqrt{1.48 \times 10^{-7}}$$
$$= 0.200 \pm 0.0011$$

10.8 Polynomial regression

As already mentioned in the introduction, multiple regression can also be used to solve polynomial regression problems. By setting $x_1 = x$, $x_2 = x^2$, ..., $x_m = x^m$ in eq. (10.1) an mth degree polynomial relationship

$$\eta = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_m x^m \tag{10.26}$$

is obtained which can be estimated as described in Section 10.2.

A first degree polynomial is the straight line model. Expanding the model with a quadratic term introduces curvature and a maximum or a minimum in the function values. A second degree (or quadratic) model is the general equation for a parabola and is symmetrical around its extremum. In Fig. 10.5 three second-order polynomials are shown.

$$\hat{y} = 5 + 0.20x - 0.40x^2 \tag{10.27}$$

$$\hat{y} = 5 - 0.25x + 0.10x^2 \tag{10.28}$$

$$\hat{y} = 5 - 1.50x - 0.05x^2 \tag{10.29}$$

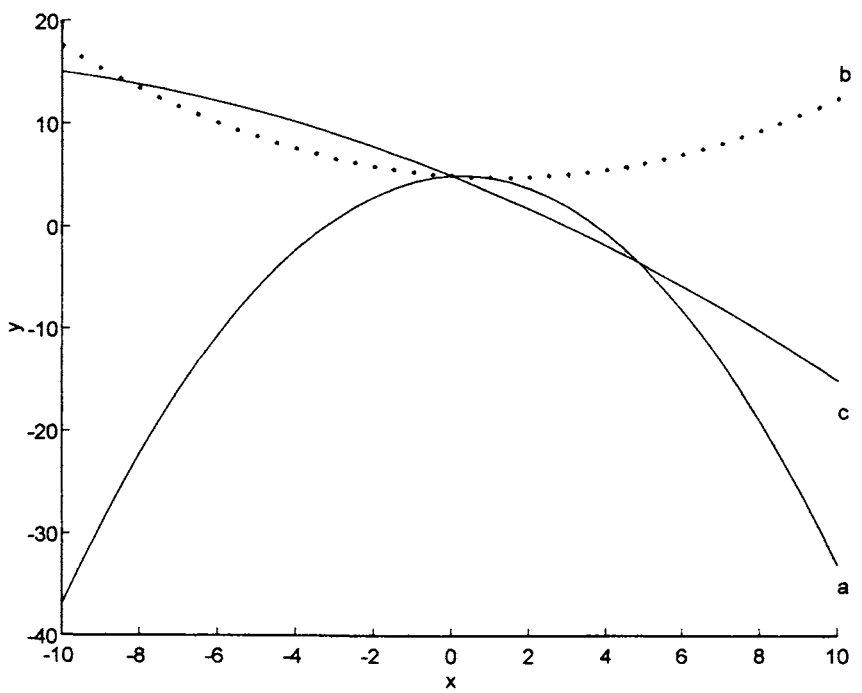


Fig. 10.5. Second order polynomials. (a) $\hat{y} = 5 + 0.20x - 0.40x^2$; (b) $\hat{y} = 5 - 0.25x + 0.10x^2$; (c) $\hat{y}=5-1.50x - 0.05x^2$.

The position of the extremum is determined by the coefficients b_1 and b_2.

$dy/dx = b_1 + 2b_2x = 0$

giving

$x = -b_1/2b_2$

When the quadratic coefficient, b_2, is negative the function has a maximum (eq. (10.27); Fig. 10.5a). When b_2 is positive the function has a minimum (eq. (10.28); Fig. 10.5.b). For the function represented by eq. (10.29) (Fig. 10.5c) the maximum is situated outside the plotted range. The larger the absolute value of the quadratic coefficient, the higher the curvature and the more sharply the extremum is defined. More complicated response relationships can be modelled by means of third or even higher order polynomials. Three examples are shown in Fig. 10.6.

$$\hat{y} = 5 - 1.5x - 0.5x^2 + 0.6x^3 \tag{10.30}$$

$$\hat{y} = 2 + 2.5x + 0.6x^2 - 0.6x^3 \tag{10.31}$$

$$\hat{y} = 5 - 0.5x - 3.5x^2 + 3.6x^3 + 0.8x^4 \tag{10.32}$$

As can be seen from Fig. 10.5 and 10.6 the higher the order of the polynomial, the more complicated relationships can be modelled.

When more descriptor variables are available the number of terms in the polynomial increases rapidly

$$\begin{aligned} \hat{y} = {} & b_0 + b_1x_1 + b_2x_2 + b_3x_3 && \text{(linear terms)} \\ & + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 && \text{(quadratic terms)} \\ & + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 && \text{(cross-product terms)} \end{aligned}$$

This is called a fully quadratic model as it contains all possible terms up to second order. Cross-product terms such as x_1x_2 represent interaction terms (see Section 6.6). It means that the response cannot be described with a purely additive model, i.e. as a sum of independent terms, one for each separate descriptor. Some second order polynomials in two independent variables are shown in Fig. 10.7.

$$\hat{y} = 15 - 7.5x_1 + 1.0x_2 + 0.5x_1^2 \tag{10.33}$$

$$\hat{y} = 15 - 7.5x_1 + 5.0x_2 + 0.5x_1^2 - 0.5x_2^2 \tag{10.34}$$

$$\hat{y} = 15 + 10x_1 + 10x_2 - x_1x_2 \tag{10.35}$$

The graphs in Fig. 10.7 are called *response surfaces*. This term is used in a wider context to denote the form of the response as a function of the predictor variables. When there is one predictor variable the response surface reduces to a curve (e.g. Figs. 10.5 and 10.6). When there are more than two predictor variables the response surface becomes a higher-dimensional hypersurface, which can no longer be visualized.

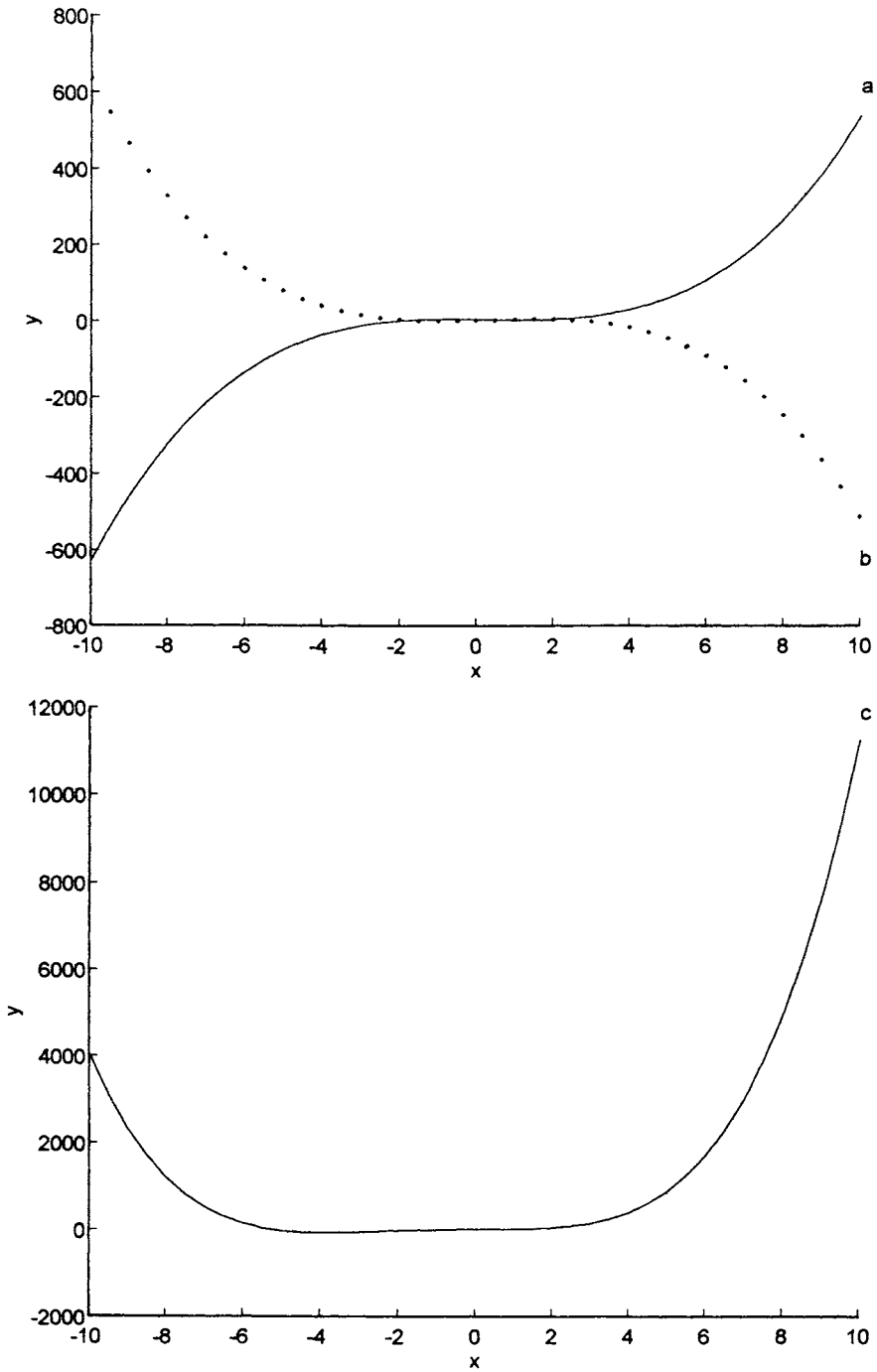


Fig. 10.6. Higher order polynomials. (a) $\hat{y} = 5 - 1.5x - 0.5x^2 + 0.6x^3$; (b) $\hat{y} = 2 + 2.5x + 0.6x^2 - 0.6x^3$; (c) $\hat{y} = 5 - 0.5x - 3.5x^2 + 3.6x^3 + 0.8x^4$.

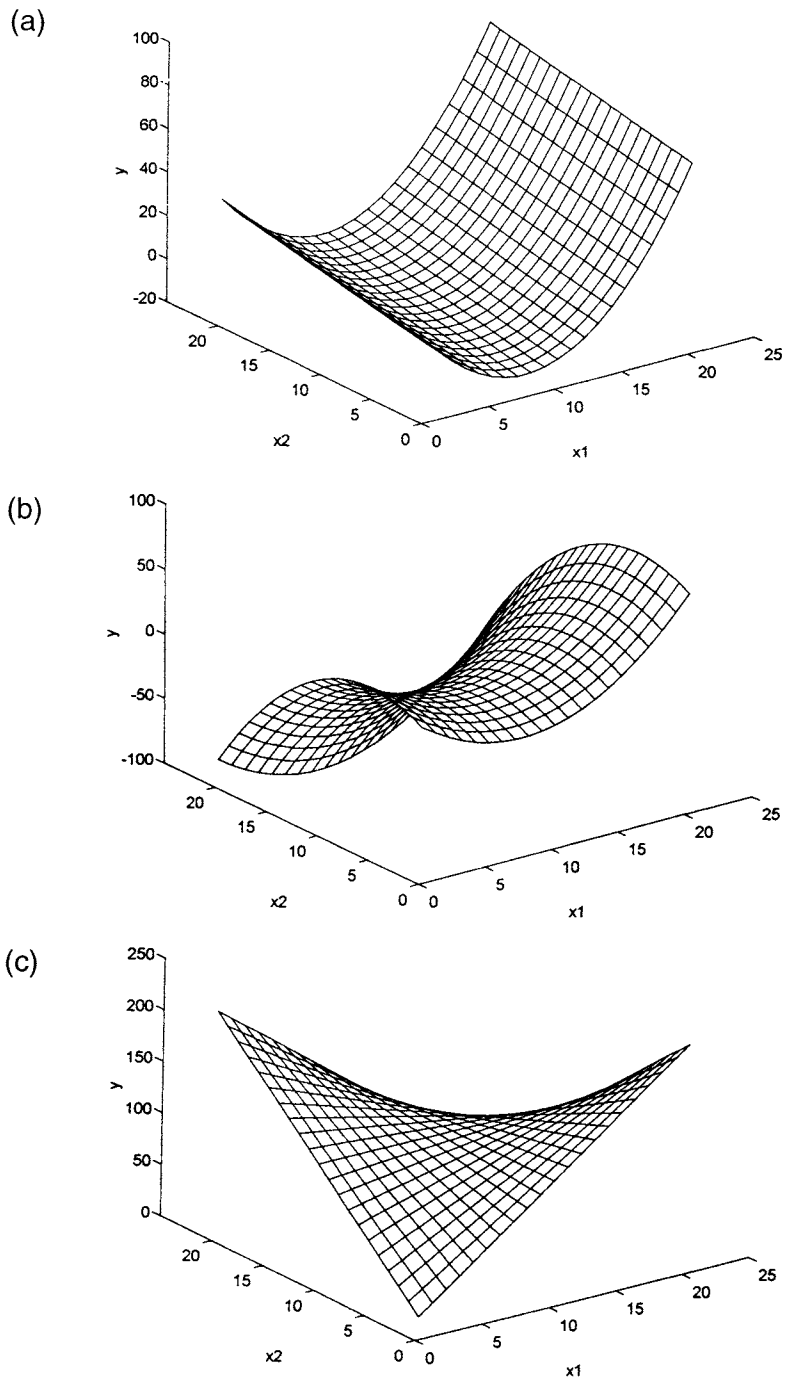


Fig. 10.7. Response surfaces. (a) $\hat{y} = 15 - 7.5x_1 + 1.0x_2 + 0.5x_1^2$; (b) $\hat{y} = 15 - 7.5x_1 + 5.0x_2 + 0.5x_1^2 - 0.5x_2^2$; (c) $\hat{y} = 15 + 10x_1 + 10x_2 - x_1x_2$.

In general, the inclusion of additional terms requires additional experimental data. Moreover, unnecessary terms decrease the quality of the prediction. For the selection of the degree of the polynomial the same methods as described in Section 10.3.3 for the selection of variables in multiple regression can be used. However generally terms are entered from low order to high order. The significance of the term added is then evaluated from the confidence interval (eq. (10.15)), a t-test (eq. (10.16)) or an F-test (eq. (10.10)). Experience tells us that many response relationships can be described by polynomials of degree two or three.

The only possible interpretation of the resulting equation is in terms of the relative contributions of the independent variables to the response. When the term x_1x_2 is significant it can be concluded that there is an important interaction effect between the variables x_1 and x_2. Similarly when the term x_1^2 is significant the variable x_1 contributes in a quadratic way to the response. Terms such as $x_1^3x_2^4$ are, however, not easily interpretable. For this reason too one restricts the polynomial to the second degree in most practical situations.

It must be noted that in polynomial regression the terms are necessarily correlated, at least when the variables are not scaled (see Chapters 22 and 24). This, too, complicates the interpretation of the regression coefficients. The inclusion of higher order terms changes the role of the lower order terms, already in the model.

The combination of modelling response data by low-order polynomial models in conjunction with an appropriate experimental design (e.g. central composite design, see Chapter 24) is known as Response Surface Methodology.

10.9 Outliers

As indicated in Section 8.2.6 the identification of outlying observations is not straightforward. In the multiple regression situation, where visualization of the data is no longer possible, this is even less evident. The diagnostics introduced in Section 8.2.6 for the straight line regression also apply in multiple regression. Cook's squared distance, $CD_{(i)}^2$, can then also be obtained from:

$$CD_{(i)}^2 = (\mathbf{b} - \mathbf{b}_{-i})^T \mathbf{X}^T \mathbf{X}(\mathbf{b} - \mathbf{b}_{-i}) / ps_e^2 \qquad (10.36)$$

where $\mathbf{b}$ is the vector of estimated regression coefficients obtained with all data points included, $\mathbf{b}_{-i}$ is the vector of estimated regression coefficients obtained with observation i excluded from the data set and $\mathbf{X}$, p and s_e^2 are as defined before.

Since $\hat{\mathbf{y}} = \mathbf{Xb}$ (eq. (10.7)), eq. (10.36) can also be written as:

$$CD_{(i)}^2 = (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{-i})^T (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{-i}) / ps_e^2 \qquad (10.37)$$

where $\hat{\mathbf{y}}$ is the vector of estimated response values obtained with all data points included and $\hat{\mathbf{y}}_{-i}$ the vector obtained with observation i excluded from the data set.

The leverage, h_{ii}, for a point i is obtained from the ith diagonal element of the hat matrix, **H**, which is defined by:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T$$

Therefore h_{ii} is equal to:

$$h_{ii} = \mathbf{x}_i\,(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{x}_i^T \qquad (10.38)$$

with $\mathbf{x}_i$ the vector of x variables for observation i ($[1\ x_{i1}\ \ldots\ x_{im}]$). It can be shown that for the simple straight line regression this expression is equivalent to eq. (8.31).

For regression with a constant term b_0, another measure for leverage is the squared Mahalanobis distance which here takes the following form:

$$\mathrm{MD}_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})\,\mathbf{C}^{-1}\,(\mathbf{x}_i - \bar{\mathbf{x}})^T \qquad (10.39)$$

with $\mathbf{x}_i = [x_{i1}\ \ldots\ x_{im}]$

$\bar{\mathbf{x}} = (\sum \mathbf{x}_i)/n$

C = the $m \times m$ variance–covariance matrix (see Chapter 9) of **X**, the $n \times m$ matrix of independent variables.

It is a measure of the distance of $\mathbf{x}_i$ from $\bar{\mathbf{x}}$ that takes correlation into account. For the simple straight line regression expression (10.39) reduces to the square of the standardized value of x_i (see eq. (8.30)).

As already mentioned in Section 8.2.6 the following relationship exists between h_{ii} and MD_i^2:

$$h_{ii} = \frac{1}{n} + \frac{\mathrm{MD}_i^2}{n-1}$$

For all data points of the stack loss example these diagnostics, together with the standardized residuals, $|e_i/s_e|$, (see Section 8.2.6) are listed in Table 10.18. In Section 10.3.4 it was noticed that the first observation is a leverage point since with $x_1 = 80$ and $x_2 = 87$ it is remote from the rest of the data. This is reflected in h_{ii} and MD_i^2 being large. As discussed in Section 8.2.6, the large $\mathrm{CD}_{(i)}^2$ value can be due to the fact that the observation has a considerable influence on the regression estimates but also to the fact that it is a leverage point as indicated by h_{ii}. The former can be evaluated from a comparison of the regression equation obtained without the first data point:

$$\hat{y} = -36.248 + 0.674x_1 + 0.565x_2 \qquad s_e^2 = 1.240$$

$$(3.432)\quad (0.074)\quad (0.147)$$

with the equation for all data points:

$$\hat{y} = -42.001 + 0.777x_1 + 0.569x_2 \qquad s_e^2 = 1.590$$

$$(2.553)\quad (0.065)\quad (0.167)$$

TABLE 10.18

Outlier diagnostics for the stack loss data associated with the LS model $\hat{y} = -42.001 + 0.777x_1 + 0.569x_2$

Observation	$\lvert e_i/s_e \rvert$ (2.00)	$CD^2_{(i)}$ (1.000)	h_{ii} (0.353)*	MD_i^2 (5.991)*
1	1.20	2.158	0.626	9.083
2	0.53	0.009	0.082	0.363
3	0.98	0.045	0.111	0.837
4	0.64	0.035	0.176	1.871
5	0.16	0.002	0.176	1.871
6	0.89	0.066	0.170	1.780
7	0.57	0.026	0.167	1.726
8	0.57	0.026	0.167	1.726
9	0.23	0.009	0.271	3.397
10	1.81	0.262	0.167	1.726
11	1.47	0.086	0.097	0.616
12	0.74	0.031	0.128	1.112
13	0.05	0.000	0.128	1.112
14	0.29	0.005	0.141	1.318
15	0.29	0.005	0.141	1.318
16	0.63	0.038	0.189	2.084
17	1.69	0.068	0.063	0.060

*Cut-off value (see Section 8.2.6).

(the standard deviations of the regression parameters, s_{b_i}, are given between brackets). From this it follows that the observation has some influence on the regression estimates. On the other hand it has a beneficial effect on the standard deviation of most of the parameters, b_i. Therefore the conclusion concerning the first data point is not straightforward. The original stack loss data set [2], which contains several outliers, has been studied by different investigators. Some of them identified the first observation (= observation 2 in the original set) as an outlier while others did not so [9]. A better balanced design, with more observations that cover the whole region, would probably be necessary to come to a decisive answer.

For a more extensive discussion of different outlier diagnostics the reader is referred to the excellent book by Rousseeuw and Leroy [9].

References

1. N.R. Draper and H. Smith, Applied Regression Analysis (2nd edn). John Wiley, New York, 1981.
2. A. Brownlee, Statistical Theory and Methodology in Science and Engineering. John Wiley, New York, 1965.
3. P. Dagnelie, Analyse statistique à plusieurs variables. Les Presses Agronomique de Gembloux, Gembloux, Belgium, 1975.

4. A. Palm, Les critères de validation des équations de regression linéaire. Notes de statistique et d'informatique 88/1 — Faculté des Sciences Agronomique, Gembloux, Belgium.
5. R.D. Snee, Validation of regression models: methods and examples. Technometrics, 19 (1977) 415–428.
6. A.E. Hoerl and R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12 (1970) 55–67.
7. A.E. Hoerl, Application of ridge analysis to regression problems. Chem. Eng. Prog., 58 (1962) 54–59.
8. Landolt-Bornstein, Zallen Werte und Funktionen, Teil 3, Atom und Molekular Physik. Springer Verlag, Berlin, 1951.
9. P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection. John Wiley, New York, 1987.

Additional recommended reading

M. Sergent, D. Mathieu, R. Phan-Tan-Luu and G. Drava, Correct and incorrect use of multilinear regression. Chemom. Intell. Lab. Syst., 27 (1995) 153–162.
J.G. Topliss and R.J. Costello, Chance correlations in structure-activity studies using multiple regression analysis. J. Med. Chem., 15 (1972) 1066–1068.

Chapter 11

Non-linear Regression

11.1 Introduction

In Chapter 8 linear relationships were studied with regression analysis by fitting the straight line model

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{11.1}$$

In Chapter 10 this linear relationship was extended to the case of two or more predictors giving the linear multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon \tag{11.2}$$

Geometrically this model represents a flat (hyper)plane in high (p+1)-dimensional space. In many applications a model such as eq. (11.1) or (11.2) is theoretically correct. In other instances it is at least a valid approximation in the restricted working range of the application. Therefore, the linear approach covers a major part of regression applications in chemistry. In analytical method validation the linear range is even one of the figures of merit of a method.

In many other fields of application, however, the straight line model is not appropriate and non-linear functional relationships should be used. Figure 11.1 represents some of these non-linear relationships, e.g. exponential functions (Fig. 11.1b), trigonometric functions (f), hyperbolas (c), Gaussian functions (e), logistic functions (d), splines (h), rational functions and combinations of these. Notice that some of these curves can be well approximated by a polynomial function. For example, the curve in Fig. 11.1a represents a parabola, which is defined by

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{11.3}$$

It represents y as a quadratic, i.e. a *non-linear* function of x, given the parameters β_0, β_1 and β_2. At the same time eq. (11.3) represents y as a *linear* function of the parameters β_0, β_1 and β_2, given the associated predictor variables $1(=x^0)$, $x(=x^1)$, and x^2. The latter viewpoint is relevant to regression analysis. One generally has available measurements on a set of response and predictor variables and the aim is to fit a model, i.e. to estimate its parameters. Fitting the parabolic model (eq. 11.3) is a linear parameter estimation problem that can be handled by the linear multiple

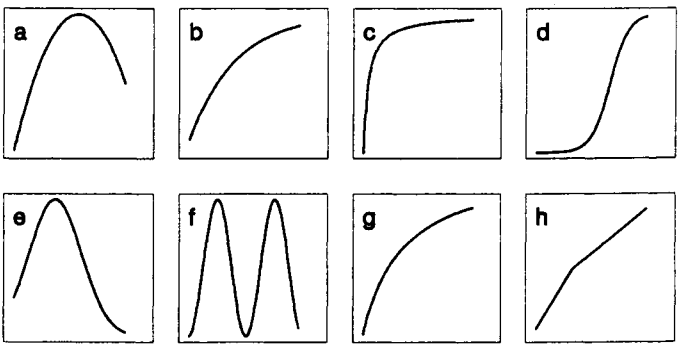


Fig. 11.1. Various non-linear relationships: (a) parabola; (b) exponential; (c) hyperbole; (d) logistic; (e) Gaussian; (f) sine; (g) rational function; (h) linear segments.

regression method discussed in Chapter 10. Non-linear regression analysis then refers to the estimation of a model involving parameters which enter the model in a non-linear way. In practice, such models are also non-linear (curved) when seen as a function of the predictor variable(s).

When studying curved relationships such as those displayed in Fig. 11.1 there are basically two approaches: the *empirical* approach and the *non-empirical* or *mechanistic* approach. In the empirical approach one tries to model as well as possible the form of the response by means of a simple function. The choice of the functional form is suggested by the data and also determined by considerations of computational ease. The resulting fitted model is used mainly for summarizing the relation in the form of a smooth function or for future prediction purposes. Interpretation of the individual model coefficients is only a secondary issue. When the mechanistic approach is used, the process under study must be so well understood that an appropriate functional form can be selected beforehand or can be derived from the underlying physico-chemical phenomena or from theoretical considerations. The experimental data are then modelled with this function. Estimation of the coefficients by fitting the model is in this approach the primary goal. Since the coefficients already have a well-defined meaning, the interpretation becomes straightforward.

11.2 Mechanistic modelling

The term non-empirical or mechanistic modelling is used when the data are modelled with a specific function that is available from theoretical considerations. Examples can be found in (pharmaco)-kinetics (Chapter 39), analytical chemistry (e.g. titration curves), physical chemistry, etc.

Example from kinetics

Suppose the reaction of interest is of the form:

$A + B \rightarrow P$

The simplest possible equation describing the decreasing concentration of A as a function of time (t) complies with first-order reaction kinetics:

$$[A]_t = [A]_0 \exp(-kt) \tag{11.4}$$

Here, $[A]_0$ is the concentration at start ($t = 0$) and k is the reaction rate constant. When the above function and the data do not match, the reaction apparently does not follow a first-order law and a different mechanism must be considered. This brings us immediately an important advantage of mechanistic models over empirical models. Mechanistic models, when applied appropriately, can increase the scientific understanding of the system under study. An additional advantage is that they provide a better basis for extrapolation. A number of questions arise however: how should the experiments be designed to test the proposed model and how can an inappropriate model be detected?

Example from chromatography

In chemometrics the term "curve fitting" is frequently used in the restricted sense of fitting spectroscopic or chromatographic data. Theoretical considerations may indicate the shape of a peak (e.g. Gaussian for chromatographic peaks, Lorentzian for NMR peaks). As an example, the simplest mathematical function for a chromatographic peak, including noise, reads

$$y = \beta_1 \exp[-\{(x - \beta_2)/\beta_3\}^2] + \varepsilon \tag{11.5}$$

Estimations of the parameters of the model yield information on peak characteristics such as the position of the top (β_2), the width (β_3), and the peak height (β_1). The Gaussian function is but one of many alternative mathematical functions suggested for the description of chromatographic peak profiles [1]. When overlapping peaks are studied the simplest model becomes the sum of n Gaussian functions, one for each peak:

$$y = \sum_{i=1,n} \beta_{1i} \exp[-(x - \beta_{2i})^2 / \beta_{3i}^2] + \varepsilon \tag{11.6}$$

These parameters in turn yield information on the system or process under study. In chromatography the peak position is the retention time of the compound under study and yields information on the identity of the compound. The peak height can be related to the concentration. Peak width can be an indicator of the interaction between the compound studied and the medium. Fitting appropriately selected mathematical functions thus yields in an indirect way information on the underlying phenomena.

In many cases idealized functions such as the Gaussian function are not sufficiently precise. In that case empirical modifications are introduced that enhance the fit. For example, a term may be added to make the peak shape asymmetric in order to account for "tailing" peaks. The status of curve fitting is then lying somewhere between empirical modelling and mechanistic modelling.

11.2.1 Linearization

In the previous section some examples of functions to model non-linear relations were given. The parameters in such functions can be estimated from experimental data using least-squares regression. To explain how the least-squares technique can be applied in the non-linear case, three types of curvilinear functions must be distinguished. Examples of each type are:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon \tag{11.7}$$

$$y = \exp(\beta_0 + \beta_1 x) + \varepsilon \tag{11.8}$$

$$y = \beta_1 \exp(-\beta_2 x) + \beta_3 \exp(-\beta_4 x) + \varepsilon \tag{11.9}$$

The term ε denotes the error term, just as in the linear case. The first equation is a polynomial; it is linear in the parameters and can be treated as a (multiple) linear regression model as explained in Chapter 10. The other equations are non-linear in the parameters. There is, however, an important difference between the second and the third equation. Equation (11.8) can be transformed into a linear equation:

$$\ln y = \beta_0 + \beta_1 x + \varepsilon^* \tag{11.10}$$

This form is linear in the parameters β_0 and β_1. Equation (11.8) may therefore be viewed as *intrinsically linear*. Notice that the error term has changed ($\varepsilon \rightarrow \varepsilon^*$) as a result of the transformation. The parameters β_0 and β_1 can be obtained from a simple linear regression of log-transformed y, i.e. $\ln y$, on x. For this reason it is said that the model (11.8) is linearizable. This does not imply that solving eq. (11.8) via a least-squares fit of eq. (11.10) is always adequate. Transforming the response affects the distribution of errors. If the error ε in eq. (11.8) has a homoscedastic normal distribution, then the error ε^* in eq. (11.10) will have a different, heteroscedastic non-normal distribution. In that case (non-linear) least-squares regression of model (11.8) is appropriate, and (unweighted) linear least-squares regression of model (11.10) is not. This topic has already been discussed in Section 8.2.3.

It is impossible to transform the third equation (eq. 11.9) into a form that is linear in the parameters. This equation is therefore said to be *intrinsically non-linear*. Whatever the transformation is, it will always yield a model that is non-linear in the parameters. However, when the non-linear parameters β_2 and β_4 are not too close, one may — to a good approximation — solve the equation in parts, each of which can be linearized (*cf* Chapter 39, 'curve peeling').

11.2.2 Least-squares parameter estimation

The least-squares principle as explained in Chapter 8 can also be applied in the non-linear case. The solution is, however, more complicated as will become clear. As in the linear case, the sum of the squares of the residual differences between the experimental value and the value predicted by the model is minimized:

$$y_i = \hat{y}_i(\mathbf{b}) + e_i \tag{11.11}$$

and

$$e_i = y_i - \hat{y}_i(\mathbf{b}) \tag{11.12}$$

$\hat{y}(\mathbf{b})$ is the estimated value of the response using the non-linear equation with estimated values $\mathbf{b}$ for the parameters $\boldsymbol{\beta}$, the vector containing the values of the unknown model parameters. As for the linear case we assume independence, homoscedasticity and normality of the errors: $\varepsilon \sim N(0,\sigma^2)$.

The sum of squares of the residual errors is

$$SS(\mathbf{b}) = \sum \{y_i - \hat{y}_i(\mathbf{b})\}^2 \tag{11.13}$$

where the sum runs over all n experimental data points. The least squares estimate $\mathbf{b}$ of $\boldsymbol{\beta}$ are those values of the parameters that minimize SS($\mathbf{b}$). To find the least squares solution we need to differentiate the SS($\mathbf{b}$) with respect to the parameters, $\mathbf{b}$. Doing this for all p model parameters yields the normal equations. There are as many normal equations as there are model parameters. These normal equations must be solved for $\mathbf{b}$:

$$\partial SS/\partial b_j = -2 \sum \{y_i - \hat{y}_i(\mathbf{b})\} \{\partial \hat{y}_i(\mathbf{b}) / \partial b_j\} = 0 \qquad \text{for } j = 1,..,p \tag{11.14}$$

Recall that for linear regression the normal equations are also linear in the parameters (*cf* Section 8.2.1). For example, $y_i = b_0 + b_1 x_i + e_i$ (eq. 11.1) yields $\partial \hat{y}_i / \partial b_0 = 1$ and $\partial \hat{y}_i / \partial b_1 = x_1$, leading to the normal equations:

$$\partial SS / \partial b_0 = -2 \sum \{y_i - \hat{y}_i(\mathbf{b})\} \cdot 1 = -2 \sum (y_i - b_0 - b_1 x_i) = 0 \tag{11.15a}$$

$$\partial SS / \partial b_1 = -2 \sum \{y_i - \hat{y}_i(\mathbf{b})\} \cdot x_i = -2 \sum (y_i - b_0 - b_1 x_i) x_i = 0 \tag{11.15b}$$

The normal equations (eqs. 11.15a and b) are linear in the parameters b_0 and b_1 and can be solved as explained in Chapter 8. In the non-linear case the normal equations are no longer linear in the parameters and this makes the solution more difficult. Consider, for example, the following simple non-linear function:

$$\hat{y} = \exp(-bx) \tag{11.16}$$

and suppose there are n observations (y_i, x_i) available. The derivative of the model predictions with respect to the parameter b are:

$$\partial \hat{y}_i / \partial b = -x_i \exp(-bx_i) \tag{11.17}$$

The (single) normal equation then is:

$$\sum \{y_i - \exp(-bx_i)\}\{-x_i \exp(-bx_i)\} = -\sum y_i x_i \exp(-bx_i) + \sum x_i \exp(-2bx_i) = 0 \quad (11.18)$$

Notwithstanding the simplicity of the non-linear model (eq. 11.16), the normal equation (eq. 11.18) is already quite complicated. It has no analytical solution. When the model contains multiple parameters, it is generally not feasible to solve these equations in an analytical way. Therefore, iterative numerical methods are used to estimate the parameters in the non-linear case. The fact that there are often multiple solutions (local minima) complicates the situation even more. The best known methods are *linearization*, *steepest descent* and the *Marquardt* compromise. It is also possible to use sequential optimization methods, such as Simplex (see Chapter 26) to solve non-linear equations.

11.2.3 Gauss–Newton linearization

We will explain the linearization or *Gauss–Newton* method by trying to fit a curve to a chromatographic peak. Figure 11.2 illustrates such a peak. The data points will be modelled by the Gaussian function

$$\hat{y} = b_1 \exp[-\{(x - b_2) / b_3\}^2] \quad (11.19)$$

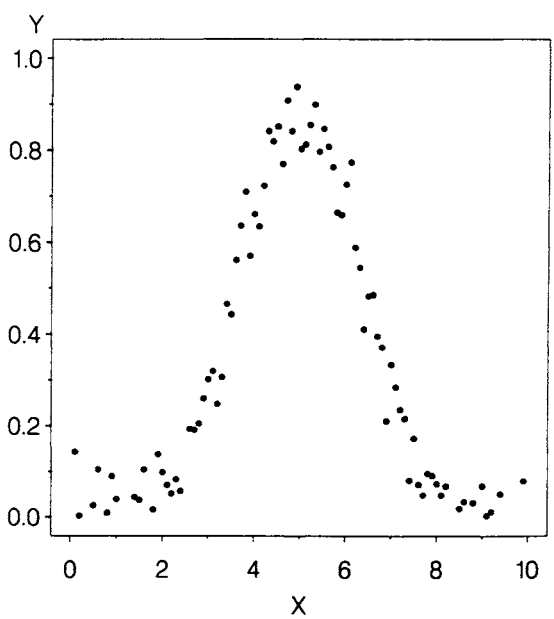


Fig. 11.2. Raw data sampled from a Gaussian peak.

where $\hat{y}$ is the fitted height of the signal at retention time x. The parameter estimate b_1 gives the peak height, b_2 the position (retention time) of the peak maximum, and b_3 is related to the peak width.

The linearization method starts the iteration process with some initial values of the parameters, say $\mathbf{b}^0$, with the superscript indicating the iteration number. In general these initial values may be intelligent guesses or they can be estimated from other procedures. In this example a first estimation of the initial values can be obtained from a visual inspection of the chromatogram. These initial values will be used as a starting point of the iterative process and will, hopefully, be improved during the process.

According to *Taylor's theorem* all continuous functions f(z) can be expanded around some fixed point z^0 as follows:

$$f(z) = f(z^0) + \{\partial f/\partial z\}_{z^0} (z - z^0) + \textit{higher-order terms.} \tag{11.20}$$

When z is close to z^0 we may disregard the terms of higher order. First-order Taylor expansion acts as a local linear approximation of the function f(z) in the neighbourhood of z^0. When there are more variables this may be generalized to:

$$f(\mathbf{z}) \approx f(\mathbf{z}^0) + \sum [\{\partial f / \partial z_j\}_{z^0} (z_j - z_j^0)] \tag{11.21}$$

retaining only the linear terms.

We now apply this to the model estimate $\hat{y}$ which is a non-linear function of the parameter estimate $\mathbf{b}$:

$$\hat{y}_i(\mathbf{b}) \approx \hat{y}_i (\mathbf{b}^0) + \sum[\{\partial \hat{y}_i (\mathbf{b}) / \partial b_j\}_{\mathbf{b}=\mathbf{b}^0} (b_j - b_j^0)] \tag{11.22}$$

where we have again omitted the higher-order terms in the Taylor expansion. For the example this becomes:

$$\hat{y}_i(b_1, b_2, b_3) = \hat{y}_i(b_1^0, b_2^0, b_3^0) + \\ + (\partial \hat{y}_i(\mathbf{b})/\partial b_1)_{\mathbf{b}=\mathbf{b}^0} \Delta b_1^0 + (\partial \hat{y}_i(\mathbf{b}) / \partial b_2)_{\mathbf{b}=\mathbf{b}^0} \Delta b_2^0 + (\partial \hat{y}_i(\mathbf{b}) / \partial b_3)_{\mathbf{b}=\mathbf{b}^0} \Delta b_3^0 \tag{11.23}$$

or, in a simpler notation,

$$\hat{y}_i = \hat{y}_i^0 + J_{i1}^0 \Delta b_1^0 + J_{i2}^0 \Delta b_2^0 + J_{i3}^0 \Delta b_3^0 \tag{11.24}$$

Here, $\hat{y}_i = \hat{y}_i\ (b_1, b_2, b_3)$ is the response predicted for a new set of parameter values different from $\hat{y}_i^0 = \hat{y}_i(b_1^0, b_2^0, b_3^0)$, the prediction using the current parameter estimates. Further, $\Delta b_j^0 = b_j - b_j^0$ is the difference between the new parameter values and the 'old' values and $J_{ij}^0 = \{\partial \hat{y}_i(\mathbf{b}) / \partial b_j\}_{\mathbf{b}=\mathbf{b}^0}$ is the derivative of the predicted response with respect to the jth parameter evaluated at $x = x_i$ and $\mathbf{b} = \mathbf{b}^0$. The terms J_{ij}^0 can be calculated from analytical expressions for the partial derivatives of the model with respect to the parameters. Alternatively it can be computed numerically from finite differences, e.g. $J_{i2}^0 = \{y_i(b_1^0, b_2^0 + \delta, b_3^0) - y_i(b_1^0, b_2^0, b_3^0)\}/\delta$ for some small value δ. Equation (11.24) can be further simplified to

$$\hat{y}_i - \hat{y}_i^0 = \sum_{j=1,p} J_{ij}^0 \, \Delta b_j^0 \tag{11.25a}$$

or, since $y_i = \hat{y}_i + e_i$,

$$\Delta y_j^0 = \sum_{j=1,p} J_{ij}^0 \, \Delta b_j^0 + e_i \tag{11.25b}$$

where $\Delta y_j^0 = y_i - \hat{y}_i^0$ is the deviation of the observed data from the predictions using the parameters $\mathbf{b}^0$. Note that eq. (11.25b) looks like a linear regression equation, the independent variables now being the J_{ij}-terms (cf. the x_{ij} terms in multiple regression). The parameters can now be estimated by applying the classical least squares procedure.

As in multiple linear regression, the terms J_{ij}^0 can be collected in an $n \times p$ matrix $\mathbf{J}^0$ (the so-called *Jacobian*):

$$\mathbf{J}^0 = \begin{bmatrix} J_{11}^0 & J_{12}^0 & \dots & J_{1j}^0 & \dots & J_{1p}^0 \\ J_{21}^0 & J_{22}^0 & \dots & J_{2j}^0 & \dots & J_{2p}^0 \\ . & . & \dots & . & \dots & . \\ J_{i1}^0 & J_{i2}^0 & \dots & J_{ij}^0 & \dots & J_{ip}^0 \\ . & . & \dots & . & \dots & . \\ J_{n1}^0 & J_{n2}^0 & \dots & J_{nj}^0 & \dots & J_{np}^0 \end{bmatrix} \tag{11.26}$$

Likewise, the parameter corrections Δb_j^0 can be placed in a vector $\Delta \mathbf{b}^0$

$$\Delta \mathbf{b}^0 = \begin{bmatrix} \Delta b_1^0 \\ \Delta b_2^0 \\ . \\ \Delta b_j^0 \\ . \\ \Delta b_p^0 \end{bmatrix} \tag{11.27}$$

and so can the current residuals $\Delta y_i^0 = y_i - \hat{y}_i^0$

$$\Delta \mathbf{y}^0 = \begin{bmatrix} y_1 - \hat{y}_1^0 \\ y_2 - \hat{y}_2^0 \\ . \\ y_i - \hat{y}_i^0 \\ . \\ y_n - \hat{y}_n^0 \end{bmatrix} \tag{11.28}$$

Therefore, eq. (11.25b) can be written compactly in matrix notation as:

$$\Delta \mathbf{y}^0 = \mathbf{J}^0 \, \Delta \mathbf{b}^0 + \mathbf{e} \tag{11.29}$$

The least squares solution is given by:

$$\Delta \mathbf{b}^0 = (\mathbf{J}^{0T}\mathbf{J}^0)^{-1}\,\mathbf{J}^{0T}\,\Delta \mathbf{y}^0 \tag{11.30}$$

assuming that the Jacobian $\mathbf{J}^0$ is non-singular (Section 9.3.5). Notice that eq. (11.30) has exactly the same form as the solution for regression coefficient vector in multiple linear regression (eq. 10.6). However, in the present non-linear case the solution is not $\mathbf{b}$ itself, but a correction $\Delta\mathbf{b}^0$ to the current guess $\mathbf{b}^0$, giving $\mathbf{b}^1$, the next better approximation of the true parameter vector $\boldsymbol{\beta}$:

$$\mathbf{b}^1 = \mathbf{b}^0 + \Delta \mathbf{b}^0 \tag{11.31}$$

The revised estimate $\mathbf{b}^1$ of the non-linear model can now be used in exactly the same way as the initial estimate $\mathbf{b}^0$. Since the Jacobian $\mathbf{J}$ depends on the parameters it has to be computed for every update of the parameter vector ($\mathbf{J}^0 \rightarrow \mathbf{J}^1 \rightarrow \mathbf{J}^2 \rightarrow$ etc.). Here lies the essential difference with linear regression where the design matrix $\mathbf{X}$ plays the same role as the Jacobian $\mathbf{J}$ in non-linear regression. The difference is that in linear regression $\mathbf{X}$ is a given fixed matrix that does not depend on the parameters and does not need to be updated.

At each stage during the iterations the error sum of squares is given by:

$$\mathrm{SS}(\mathbf{b}) = \Sigma\{y_i - \hat{y}_i\,(\mathbf{b})\}^2 \tag{11.32}$$

and at each iteration it can be verified whether the sum of squares actually has been reduced. The procedure can be repeated several times until convergence, e.g. until the relative difference between two successive estimates of $\boldsymbol{\beta}$ is smaller than a predefined small value, δ.

$$|b_j^k - b_j^{k+1}| \,/\, |b_j^k| < \delta, \text{ for all } j \tag{11.33}$$

When a parameter value happens to be nearly zero, one should use the absolute difference $|b_j^k - b_j^{k+1}|$ rather than the relative difference as a criterion for convergence. Upon convergence $\mathbf{b} = \mathbf{b}^{\text{final}}$ represents the least-squares estimate of $\boldsymbol{\beta}$.

Upon convergence the error sum of squares $\mathrm{SS}(\mathbf{b}^{\text{final}}) = \sum e_i^2$ can be used to estimate the error variance

$$s_e^2 = \sum e_i^2 \,/(n-p) \tag{11.34}$$

The standard errors of the parameters can be obtained from the appropriate diagonal elements of the matrix $(\mathbf{J}^{\text{final,T}}\mathbf{J}^{\text{final}})^{-1}$, in analogy to the linear regression case (eq. 10.16):

$$s(b_j) = s_e\{[(\mathbf{J}^{\text{final,T}}\,\mathbf{J}^{\text{final}})^{-1}]_{jj}\}^{1/2} \tag{11.35}$$

Standard errors computed in this way are approximate, even with homoscedastic normally distributed errors. For large number of observations they become correct. Given the standard errors one can obtain confidence intervals as $b_j \pm t_{0.025;n-p}\, s(b_j)$, again completely analogous to the case of linear regression (*cf* eq. 10.13).

The linearization procedure has some drawbacks. The convergence is highly dependent on the quality of the initial estimates and can be quite slow. Sometimes the solutions may oscillate and, consequently, no convergence is reached at all.

11.2.4 Steepest descent and Marquardt procedure

In the *steepest descent* or *gradient* approach one determines the sensitivity of the error sum of squares SS(**b**) with respect to each parameter in the neighbourhood of the current estimate $\mathbf{b}^0$. It can be shown that this is given by

$$\mathbf{f}^0 = \{dSS/d\mathbf{b}\}^0 = -2(\mathbf{J}^T\mathbf{e})^0 \tag{11.36}$$

The parameter estimates are then updated in proportion to these sensitivities. This corresponds to the steepest descent direction of the sum of squares as a function of the parameters. Thus the update of $\mathbf{b}^0$ becomes $\Delta\mathbf{b}^0 = \alpha\mathbf{f}^0$, where α is a proportionality constant, and the next best estimate $\mathbf{b}^1$ is

$$\mathbf{b}^1 = \mathbf{b}^0 + \Delta\mathbf{b}^0 \tag{11.37}$$

The choice of the factor α is somewhat arbitrary. The steepest descent method can be particularly effective to improve the parameter estimates when they are far away from their final best-fit values. When the estimates approach their final values convergence can become quite slow.

For this reason the *Marquardt* method provides a useful compromise between the linearization method and the steepest descent procedure. Here the update is written as

$$\Delta\mathbf{b}^0 = (\mathbf{J}^{0\,T}\mathbf{J}^0 + \lambda\mathbf{D})^{-1}\,\mathbf{J}^{0\,T}\,\mathbf{e} \tag{11.38}$$

where the matrix **D** is a diagonal matrix with the same diagonal elements as $(\mathbf{J}^{0\,T}\mathbf{J}^0)$ and λ is a tuning parameter that affects a relative increase of the diagonal elements of $(\mathbf{J}^{0\,T}\mathbf{J}^0)$. When $\lambda \to 0$ we essentially approach the linearization method of Section 11.2.3. As $\lambda \to \infty$, the $\mathbf{J}^{0T}\mathbf{J}^0$-term in eq. (11.38) becomes relatively unimportant and $\Delta\mathbf{b}^0$ becomes proportional to $\mathbf{f}^0$ as in the steepest descent method. A good implementation of the Marquardt method starts with a relatively large value of λ (e.g. $\lambda = 10^{-2}$) and gradually decreases λ as the solution converges and the error sum of squares continues to decrease.

Notice that eq. (11.38) has a similar appearance as the solution of a ridge regression problem (Section 10.6). Indeed, the Marquardt method was originally devised to cope with the situation of highly correlated parameter estimates giving rise to a near-singular Jacobian matrix. There is always a danger in non-linear regression that the solution found does not correspond to the global least squares solution, but rather to a local minimum. One way to decrease the likelihood of such solutions is by redoing the calculations and starting from different initial parameter settings. When the same solution is repeatedly found one can be confident that the global minimum has been found.

11.2.5 An example

Table 11.1 gives the activity of the enzyme Savinase as a function of time. The data are plotted in Fig. 11.3. We try and model these data with a simple exponential decay corresponding to a first-order reaction. Our interest is in the half-life, $t_{1/2}$, and we rewrite eq. (11.4) as:

$$A(t) = B \exp\{-(\ln 2)\, t/t_{1/2}\} \tag{11.39}$$

The linearization method to find the coefficients B and $t_{1/2}$ for eq. (11.39) will be worked out step by step.

TABLE 11.1

Activity (A) of Savinase as a function of time (t, in hours)

i	t	A
1	0	20.2
2	1	17.2
3	2	14.1
4	3	10.7
5	7	4.9
6	10	2.9
7	15	2.2
8	18	1.2

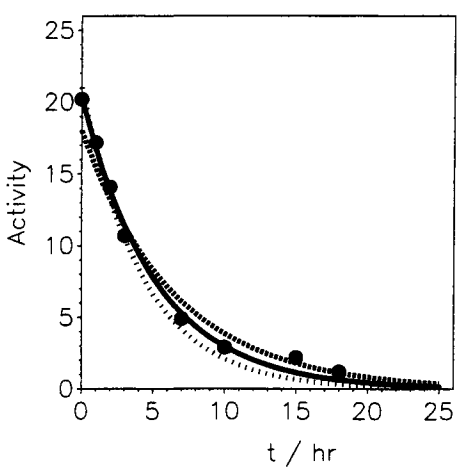


Fig. 11.3. Enzyme activity (A) as a function of storage time (t). Three fits to the data are shown: (—) non-linear regression, first-order kinetics; (- - -) non-linear regression, second-order kinetics; (· · ·) back-transformed linear regression of ln(A) on t.

(1) Initial estimates of the coefficients

From eq. (11.39) it can be derived that the activity, A, at time zero equals B. The parameter $t_{1/2}$ is defined as the time in which the activity has decreased to half of the original value. In this way, rough estimates can be made by inspecting the data of Table 11.1: $B^0 = 20$ and $t^0_{1/2} = 4$.

(2) The linearization procedure

Iteration 0

Step 0.1: Using eq. (11.39) and initial estimates B^0 and $t^0_{1/2}$ compute the predicted responses, $\hat{A}^0$:

i	t	$A(=y)$	$\hat{A}^0$	$\Delta y^0 = A - \hat{A}^0$
1	0	20.2	20.0000	0.2000
2	1	17.2	16.8179	0.3821
3	2	14.1	14.1421	–0.0421
4	3	10.7	11.8921	–1.1921
5	7	4.9	5.9460	–1.0460
6	10	2.9	3.5355	–0.6355
7	15	2.2	1.4865	0.7135
8	18	1.2	0.8839	0.3161
			$SS^0 =$	3.7159

Step 0.2: Calculate the partial derivatives of the expected response with respect to the two parameters in order to derive the Jacobian matrix $\mathbf{J}^0$ (eq. 11.26)

$$(dA/dB)^0 = \exp\{(-\ln 2)\, t/\, t^0_{1/2}\}$$

$$(dA/dt_{1/2})^0 = \{(B^0 \ln 2\, t)/(t^2_{1/2})^0\} \exp\{(-\ln 2) t/t^0_{1/2}\}$$

$$\mathbf{J}^0 = \begin{array}{c} \begin{matrix} (dA/dB)^0 & (dA/dt_{1/2})^0 \end{matrix} \\ \begin{bmatrix} 1.0 & 0.0 \\ 0.8409 & 0.7286 \\ 0.7071 & 1.2253 \\ 0.5946 & 1.5456 \\ 0.2973 & 1.8031 \\ 0.1768 & 1.5317 \\ 0.0743 & 0.9660 \\ 0.0442 & 0.6892 \end{bmatrix} \end{array}$$

Step 0.3: Minimize the term SS (eq. 11.32) to obtain the least squares solution for the correction $\Delta\mathbf{b}^0$ (eq. 11.30)

$$\Delta\mathbf{b}^0 = (\mathbf{J}^{0\,T}\mathbf{J}^0)^{-1}\mathbf{J}^{0\,T}\,\Delta\mathbf{y}^0 = \begin{bmatrix} 0.2653 \\ -0.3891 \end{bmatrix}$$

Step 0.4: Use these corrections to update the parameters by means of eq. (11.31)

$B^1 = B^0 + \Delta b_1^0 = 20 + 0.2653 = 20.2653$

$t_{1/2}^1 = t_{1/2}^0 + \Delta b_2^0 = 4 - 0.3891 = 3.6109$

Iteration 1

Step 1.1: Using the updated parameter estimates, compute new predictions, $\hat{A}^1$, and new residuals:

i	t	A	$\hat{A}^1$	$\Delta y^1 = A - \hat{A}^1$
1	0	20.2	20.2653	–0.06531
2	1	17.2	16.7258	0.4742
3	2	14.1	13.8045	0.2955
4	3	10.7	11.3934	–0.6934
5	7	4.9	5.2868	–0.3868
6	10	2.9	2.9723	–0.0723
7	15	2.2	1.1383	1.0617
8	18	1.2	0.6400	0.5600
			$SS^1 =$	2.3929

Step 1.2: Recalculate the Jacobian matrix $\mathbf{J}^1$ using the new estimates B^1 and $t_{1/2}^1$

$$\mathbf{J}^1 = \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} (dA/dB)^1 & (dA/dt_{1/2})^1 \end{matrix} \begin{bmatrix} 1.0000 & 0.0000 \\ 0.8253 & 0.8891 \\ 0.6812 & 1.4677 \\ 0.5622 & 1.8170 \\ 0.2609 & 1.9673 \\ 0.1467 & 1.5801 \\ 0.0562 & 0.9077 \\ 0.0316 & 0.6124 \end{bmatrix}$$

Step 1.3: Obtain the least-squares solution for $\Delta\mathbf{b}^1$

$$\Delta\mathbf{b}^1 = (\mathbf{J}^{1\,T}\mathbf{J}^1)^{-1}\mathbf{J}^{1T}\,\Delta\mathbf{y}^1 = \begin{bmatrix} 0.0591 \\ -0.0133 \end{bmatrix}$$

Step 1.4: Update the parameter estimates

$B^2 = B^1 + \Delta b_1^1 = 20.2653 + 0.0591 = 20.3244$

$t_{1/2}^2 = t_{1/2}^1 + \Delta b_2^1 = 3.6109 - 0.0133 = 3.5976$

Iteration 2

Step 2.1: Calculate the predicted responses, $\hat{A}^2$, with the new estimates:

i	t	A	$\hat{A}^2$	$\Delta y^2 = A - \hat{A}^2$
1	0	20.2	20.3244	–0.1244
2	1	17.2	16.7626	0.4374
3	2	14.1	13.8251	–0.2749
4	3	10.7	11.4023	–0.7023
5	7	4.9	5.2759	–0.3759
6	10	2.9	2.9598	–0.0598
7	15	2.2	1.1295	1.0705
8	18	1.2	0.6337	0.5663
			$SS^2 =$	2.3871

Step 2.2: Calculate the matrix $\mathbf{J}^2$ with the new estimates B^2 and $t_{1/2}^2$

$$\mathbf{J}^2 = \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} (dA/dB)^2 & (dA/dt_{1/2})^2 \end{matrix} \begin{bmatrix} 1.0000 & 0.0000 \\ 0.8248 & 0.8977 \\ 0.6802 & 1.4808 \\ 0.5610 & 1.8319 \\ 0.2596 & 1.9778 \\ 0.1456 & 1.5851 \\ 0.0556 & 0.9074 \\ 0.0312 & 0.6109 \end{bmatrix}$$

Step 2.3: Obtain the least squares solution for $\Delta \mathbf{b}^2$:

$$\Delta \mathbf{b}^2 = (\mathbf{J}^{2\,\mathrm{T}} \mathbf{J}^2)^{-1} \mathbf{J}^{2\,\mathrm{T}} \Delta \mathbf{y}^2 = \begin{bmatrix} 0.0014 \\ -0.0009 \end{bmatrix}$$

Step 2.4: Use the correction to update the parameters

$B^3 = 20.3244 + 0.0014 = 20.3257$

$t_{1/2}^3 = 3.5976 - 0.0009 = 3.5967$

Iteration 3

Step 3.1: Calculate the predicted responses, $\hat{A}^3$ with the new estimates

i	t	A	$\hat{A}^3$	$\Delta y^3 = A - \hat{A}^3$
1	0	20.2	20.3257	–0.1257
2	1	17.2	16.7629	0.4371
3	2	14.1	13.8247	0.2753
4	3	10.7	11.4014	–0.7014
5	7	4.9	5.2744	–0.3744
6	10	2.9	2.9586	–0.0586
7	15	2.2	1.1288	1.0712
8	18	1.2	0.6332	0.5668
			$SS^3 =$	2.3871

Remark: SS^3 equals SS^2; no improvement is made. This means that the method has converged. We will anyway calculate the Δb's for this step too.

Step 3.2 Calculate the matrix $\mathbf{J}^3$ with the new estimates B^3 and $t_{1/2}^3$

$$\mathbf{J}^3 = \begin{matrix} (dA/dB)^3 & (dA/dt_{1/2})^3 \end{matrix} \begin{bmatrix} 1.0000 & 0.0000 \\ 0.8247 & 0.8982 \\ 0.6802 & 1.4815 \\ 0.5609 & 1.8327 \\ 0.2595 & 1.9783 \\ 0.1456 & 1.5853 \\ 0.0555 & 0.9072 \\ 0.0312 & 0.6107 \end{bmatrix}$$

Step 3.3: Obtain the least squares solution for $\Delta \mathbf{b}^3$:

$$\Delta \mathbf{b}^3 = (\mathbf{J}^{3\,T}\mathbf{J}^3)^{-1}\mathbf{J}^{3\,T}\,\Delta \mathbf{y}^3 = \begin{bmatrix} 0.00009302 \\ -0.00006585 \end{bmatrix}$$

Step 3.4: Update the parameters

$B^4 = 20.3257 + 0.0001 = 20.3258$

$t_{1/2}^4 = 3.5967 - 0.0001 = 3.5966$

These values will be used as final estimates for the parameters in eq. (11.39). The final results then read:

i	t	A	$\hat{A}^{\text{final}}$	e
1	0	20.2	20.3258	–0.1258
2	1	17.2	16.7630	0.4370
3	2	14.1	13.8246	0.2754
4	3	10.7	11.4013	–0.7013
5	7	4.9	5.2743	–0.3743
6	10	2.9	2.9585	–0.0585
7	15	2.2	1.1287	1.0713
8	18	1.2	0.6331	0.5669
			$SS^{\text{final}} =$	2.3871

$$\mathbf{J}^{\text{final}} = \begin{bmatrix} 1.0000 & 0.0000 \\ 0.8247 & 0.8982 \\ 0.6802 & 1.4816 \\ 0.5609 & 1.8328 \\ 0.2595 & 1.9783 \\ 0.1456 & 1.5853 \\ 0.0555 & 0.9072 \\ 0.0311 & 0.6107 \end{bmatrix}$$

For the residual error variance we find:

$s_e^2 = 2.3871/6 = 0.40$

Since

$$(\mathbf{J}^{\text{final T}}\,\mathbf{J}^{\text{final}})^{-1} = \begin{bmatrix} 0.6141 & -0.1577 \\ -0.1577 & 0.1120 \end{bmatrix}$$

we find for the standard errors of the regression parameters:

$s(B) = (0.6141 \cdot 0.3978)^{1/2} = 0.49$

$s(t_{1/2}) = (0.1120 \cdot 0.3978)^{1/2} = 0.21$

With these estimates for the standard errors and a critical Student's t-value of 2.45 (df = 6), the 95% confidence interval estimates of the parameters are: $19.3 < B < 21.3$ and $3.2 < t_{1/2} < 4.0$. Tables 11.2 and 11.3 summarize the analysis in the form of a typical output of a non-linear regression computer program.

Figure 11.3 also shows the fit obtained via linear regression of $\log(A)$, back-transformed to the original scale. This alternative fit definitely is inferior to the fit just derived: it shows larger and more systematic deviations. We may also consider a different model (second-order kinetics). The fit obtained with this alternative model has a larger residual error sum of squares which can also be read from Fig. 11.3. Hence, the conclusion is that the experimental data are consistent with first-order kinetics.

TABLE 11.2

Evolution of parameter estimates and sum of squares during iterations

Iteration	Parameter estimates		Sum of squares
	B	$t_{1/2}$	
0	20.0	4.0	3.7159
1	20.2653	3.6109	2.3929
2	20.3244	3.5976	2.3871
3	20.3257	3.5967	2.3871
4	20.3258	3.5966	2.3871

TABLE 11.3

(a) ANOVA table of non-linear regression example

Source	df	Sum of squares	Mean square	F
Regression	2	10557.5	5278.8	13,268
Residual	6	2.3871	0.3978	

(b) Parameter estimates and asymptotic confidence intervals

Parameter	Estimate	Standard error	95% Confidence interval	
			Lower	Upper
B	20.3258	0.4943	19.3372	21.3144
$t_{1/2}$	3.5966	0.2111	3.1744	4.0188

11.2.5 Advanced topics

The statistical theory of non-linear regression modelling is considerably more complicated than for the linear case. Even when ideal assumptions are met, e.g. independence, normality and constant variance of the error, the estimators no longer have such desirable features as unbiasedness and normality. For that reason the standard deviations found in the previous section are only approximate.

Lacking exact theory, it is expedient to apply the same methods that are valid in linear regression theory. For example, one may use (approximate) t-tests to test for the significance of a parameter (compare Section 10.3.2). Replicate observations

can be used to provide a model-free estimate of the pure error and an approximate F-test can be used for testing model adequacy (compare Section 8.2). When the predictor variables are also subject to error one may try and apply orthogonal regression (compare Section 8.2.11), although this becomes much more complex. Also, methods for robust estimation (*cf* Chapter 12) can be applied in the non-linear case. The issue of experimental design in the case of a non-linear relation is another example where there is a large gap between the elegant theory and designs for the linear model and the complexity of the non-linear case [2]. In Section 11.2 we saw that the Jacobian matrix $\mathbf{J}$ depends on the parameters to be estimated. This is in contrast with linear regression where the Jacobian matrix, $\mathbf{X}$, is fixed. This has a direct bearing on experimental design. In a linear model situation we can design $\mathbf{X}$ (and hence $\mathbf{X}^T\mathbf{X}$) without knowing the corresponding parameters $\boldsymbol{\beta}$. In non-linear regression the role of $\mathbf{X}$ is taken by $\mathbf{J}$, which is not known at the start of the experiment. The conclusion then is that in order to design our experiments so as to measure the unknown parameters most precisely, we need to know their values! The practical way out is to cover the experimental region in a uniform way, with perhaps some additional experiments in those regions where 'things happen', i.e. where the response is expected to change rapidly.

Many of these advanced topics are still in an early stage of research and typical chemometric examples are scarce. For advanced and up-to-date textbooks on non-linear regression and design, see Refs. [3–5].

11.3 Empirical modelling

11.3.1 Polynomial regression

When the functional form is not known beforehand, the simplest approach to modelling curved functions is by fitting a polynomial function of a certain degree. The basis of this approach is the fact that any well-behaved mathematical function can be approximated by means of a higher-degree polynomial. Model estimation is relatively easy since the model is linear in the parameters and the regression analysis can be seen as a problem of multiple linear regression. For that reason the subject of polynomial regression or response surface modelling was already treated in Chapter 10. It will not be further discussed here, except to notice that polynomials can suffer from a serious drawback. At extreme values of the predictor variable x, polynomial functions tend to $+\infty$ or $-\infty$. This makes them unsuitable for fitting curves having horizontal asymptotes (or plateaus) at extreme values of x. An example of such a curve is shown in Fig. 11.1d. The use of *splines* is most appropriate in these cases.

11.3.2 Splines

Spline functions are constructed from joining pieces of local polynomials. The function values agree at the *knots*, i.e. the points at which the polynomial pieces are joined. Through a judicious choice of the knots and of the order of the piecewise polynomials one can fit functions of any shape. Thus, when the relationship between the independent variables and the response becomes complex, splines, because of their flexibility, can be used to advantage.

One distinguishes regression splines and smoothing splines. *Regression splines* (Section 11.3.2.1) are used to develop flexible yet parsimonious non-linear models that best fit observational data using a least-squares criterion. They are an alternative to the other regression methods described in this Chapter. *Smoothing splines* are used to regularize a set of data $\{x_i,y_i\}$. The objective is not so much to derive a model as to filter the noise from the data and to derive a smooth continuous curve through the data summarizing the main trend (Chapter 40 also discusses methods for smoothing data). We will briefly discuss the application of cubic smoothing splines in Section 11.3.2.2.

On a historical note it is interesting to observe that the very first scientific paper mentioning 'chemometrics' was a paper on the use of spline functions [6].

11.3.2.1 Regression splines

In spline regression the range of predictor values is subdivided in a number of intervals and in each interval a low-order polynomial is fitted. Usually one requires that the function is continuous at the junctions. In its simplest form a number of straight line segments is used to fit the data (see Fig. 11.4). By increasing the number of line segments it is clear that in this way complex curves can be approximated. When the line segments are replaced by quadratic or cubic functions a smooth curve can be obtained. One can then also ensure that not only the fitted function itself is continuous but also the first derivative or the second derivative (see Fig. 11.5 which gives an example of the data of Fig. 11.4 fitted with a spline of 1st, 2nd and 3rd degree). The piecewise nature of spline functions and the location of the knots is hardly visible with cubic splines and continuous second derivatives at the knots.

Let us consider the simple case of fitting a response curve by a number of piecewise linear functions. First one must decide on the number and the width of the intervals. This is done by selecting the knots or joint points. In Fig. 11.5 these are indicated by the points t_1, t_2 and t_3. The best positions of these knots can be identified after visual inspection of a scatter plot of the data. For some rules of thumb see Ref. [7]. Another approach is to spread the knots evenly over the range of the variables.

In spline regression the parameters of each polynomial segment must be estimated. We describe a simple procedure which consists in associating each knot ($k = 1,2,3$)

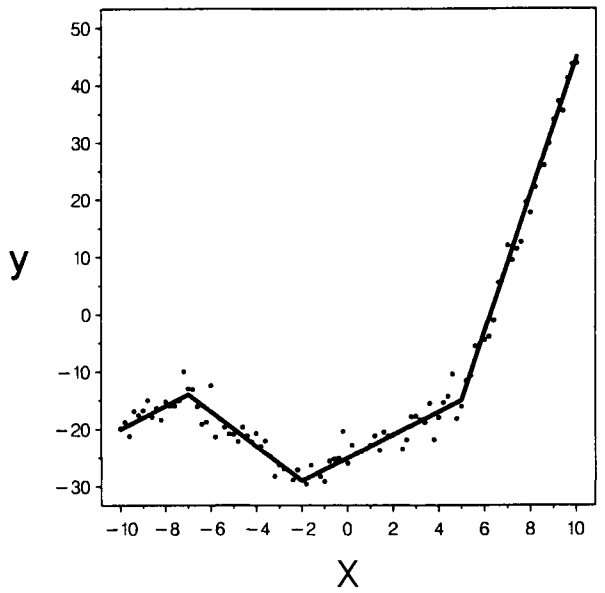


Fig. 11.4. Example of non-linear data fitted by pieces of straight lines (linear spline with 3 knots).

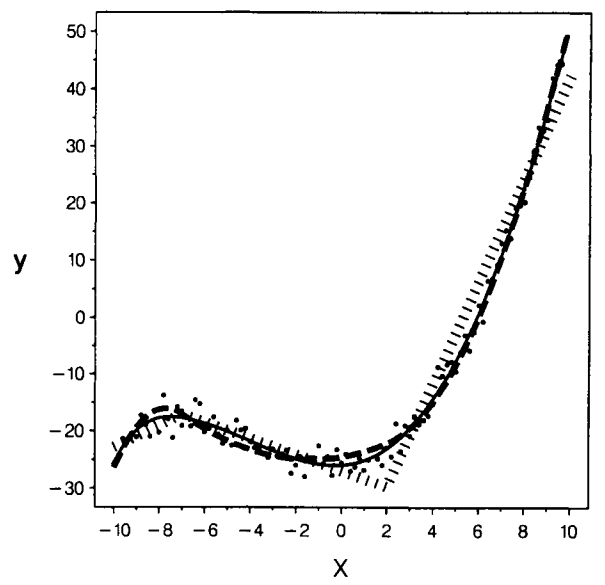


Fig. 11.5. Non-linear data fitted with spline functions of first (ıı ıı), second (- - -) and third degree (—). At the two knot positions the spline functions are continuous in the zeroth, first and second degree, respectively.

with an additional predictor variable, v_k. The value of this variable is 0 for all data points (x_i) situated on the x-axis before the knot, and is ($x_i - t_k$) for all other data points.

$$\begin{aligned} v_k &= 0 \qquad \text{for } x \le t_k \\ v_k &= x - t_k \quad \text{for } x > t_k \end{aligned} \tag{11.40a}$$

or in an implicit short-hand notation

$$v_k = (x - t_k)_+ \tag{11.40b}$$

Table 11.4 and Fig. 11.6 represent a fictitious example through which a jagged spline can be fitted. The equation of the spline curve is:

$$\hat{y} = b_0 + b_1 x + c_1 v_1 + c_2 v_2 + c_3 v_3 = 0 + 2x - 4v_1 + 3v_2 - 2v_3 \tag{11.41}$$

The coefficients c_k represent the change of slope of the line at the kth knot. This can be easily verified. Consider the first four points. Since only the x-variable is nonzero, the equation of the first line segment is:

$$y = 2x$$

From point 4 onwards the variable v_1 takes the value $(x - 4)$, while v_2 and v_3 still remain zero. The equation of the second line segment is:

$$\hat{y} = 2x - 4v_1 = 2x - 4(x - 4) = 16 - 2x$$

The equation of the third segment becomes:

$$\hat{y} = 2x - 4v_1 + 3v_2 = 16 - 2x + 3(x - 7) = -5 + 1x$$

In our example the spline consists of four line segments with slopes 2, –2, 1, and –1. The slope changes are –4, +3 and –2. These values correspond to c_1, c_2, and c_3, the coefficients of the additional variables that were included. This example is artificial since all data fit exactly the line. In practice a classical least squares procedure can be applied to fit the experimental data points using multiple regression of y on the four predictors x, v_1, v_2, and v_3. Just as in ordinary multiple regression all variables can be tested for significance. This implies that the coefficients, c_k, of all additional variables, v_k, can be tested. If one of the variables is not significant it means that the slope change at that specific joint is not significant, so that the knot can be deleted and the two neighbouring line segments combined into one. It must be noted that each additional knot or segment generates an additional variable and takes one degree of freedom away. This implies that the number of experiments required for spline regression is higher than for usual regression. As we have seen, estimation of the parameters in spline regression can be done easily using least squares regression provided the knot positions are known. In a certain sense spline regression is then a subset of multiple linear regression. However, this

TABLE 11.4

Data for the spline fit of Fig. 11.6

i	y	x	$v_1=(x-4)_+$	$v_2=(x-7)_+$	$v_3=(x-11)_+$
1	2	1	0	0	0
2	4	2	0	0	0
3	6	3	0	0	0
4	8	4 (t_1)	0	0	0
5	6	5	1	0	0
6	4	6	2	0	0
7	2	7 (t_2)	3	0	0
8	3	8	4	1	0
9	4	9	5	2	0
10	5	10	6	3	0
11	6	11 (t_3)	7	4	0
12	5	12	8	5	1
13	4	13	9	6	2
14	3	14	10	7	3
15	2	15	11	8	4

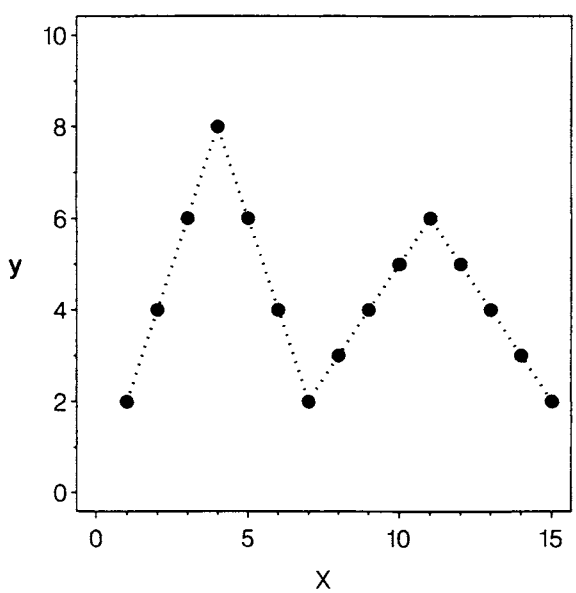


Fig. 11.6. Artificial data fitting exactly to a linear spline with three knots.

only holds true when the knot positions are fixed to known positions. If these positions have to be estimated as well then the addition of these unknown parameters renders the model estimation into a non-linear regression problem. This can be appreciated by realizing that one cannot construct a fixed design matrix as in Table 11.4 when the knot positions are not known beforehand.

The use of straight line segments is not so useful in practice since only jagged lines can be fitted. When instead of simple line segments higher-degree polynomials are fitted in each segment a smooth curve can be obtained. When fitting a quadratic polynomial for each segment, two additional parameters have to be estimated for each additional segment, one for each term in the polynomial. Just as in the case of fitting line segments, the value of these quadratic variables is zero for data points, positioned before the specific knot position. For data points after the knot the value equals

$$v_{k1} = (x - t_k)_+$$

$$v_{k2} = (x - t_k)_+^2$$

The first index refers to the segment while the second index refers to the order of the polynomial term. Each segment requires two additional degrees of freedom. It is clear that with this method arbitrarily complex curves can be fitted by increasing the number of segments or the degree of the polynomial or both. It must be remembered, however, that each additional term in the polynomial requires an additional degree of freedom per segment. If one employs the added variables of highest order only, the function becomes continuous in all derivatives except the highest derivative. For example, a superposition of an overall quadratic function and 'local' cubic terms v_{k3} yields a spline function with a continuously varying slope and a continuously varying curvature. This results from the behaviour of the individual terms $(x - t_k)_+^3$ whose first derivative, $3(x - t_k)_+^2$, and second derivative, $6(x - t_k)_+$, are continuous at the knot position t_k. Only the third derivative changes suddenly from 0 (for $x < t_k$) to 6 (for $x > t_k$) at the knot position.

The above description of developing a regression spline model was given because of its intuitive simplicity. Essentially the same model can be derived through the use of so-called *B*-splines [8]. This alternative method, which is beyond the scope of this book, has better numerical properties.

11.3.2.2 Smoothing splines

The aim of smoothing splines is to derive a realistic looking curve through a set of data points. For example, a spectroscopic measurement may generate a set of discrete points (Fig. 11.7a). Rather than just plotting the sequence of dots, we want to portray the continuous nature of the spectrum. Simply joining the dots yields a continuous curve (Fig. 11.7a), but it does not bring across the smooth nature of the spectrum. Technically, we may consider the 'model' shown in Fig. 11.7a as a

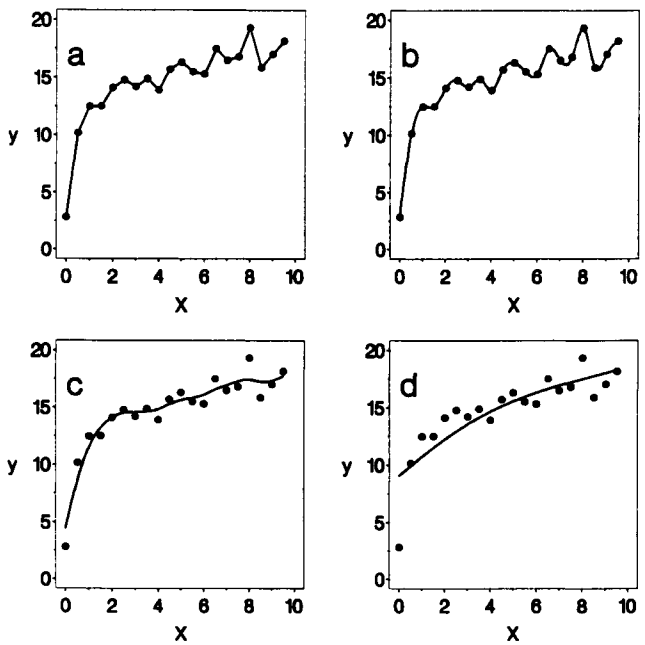


Fig. 11.7. Four spline models fitted to the same data: (a) interpolating linear spline model; (b) interpolating cubic spline model ($\lambda = 0$); (c) smoothing cubic spline model (optimum λ); (d) oversmoothed cubic spline model (large λ).

first-degree spline function with a knot at every observed x-value. Since it fits each observed data point, this type of function is known as an *interpolating spline function*. Although the function itself is continuous, its first derivative is not. We may obtain a quadratic interpolating spline function by applying the method discussed in the previous section, treating each observed x-value as a knot. Now, the function as well as its first derivative are continuous. The second derivative is discontinuous, which may be visible as abrupt changes in the curvature of the spline function at the knot positions. Therefore, we may go one step further and obtain an interpolating cubic spline function by fitting the data with a general intercept, a general linear term (x), a general quadratic term (x^2), and a separate third-order term ($v_{i3} = (x - x_i)_+^3$, $i = 1,...,n$) for each observed x-value. This fit is shown in Fig. 11.7b. It has continuous second derivatives, so that the curvature changes in a gradual manner, giving the curve quite a natural and realistic appearance.

In one sense, though, interpolating splines are not realistic. We know that the measured data is noisy, so we may relax the requirement that the curve passes exactly through the observed data points. We may allow some deviation of the curve at the observed data points if this results in a smoother curve. Mathematically, this boils down to minimizing a composite optimization criterion that has two contributions: the usual error sum of squares, $SS_E = \sum\{y_i - f(x_i)\}^2$, and a term

representing the total amount of curvature, CURV = $\int\{\partial^2 f/\partial x^2\}^2 dx$. In the composite criterion the second term is weighted with a smoothing parameter λ, i.e. one minimizes $SS_E + \lambda$ CURV. For small values of λ the emphasis is on minimizing SS_E, hence on obtaining a close fit. The perfectly fitting interpolating spline of Fig. 11.7b corresponds to the limit $\lambda = 0$. As λ takes larger values smoother curves are obtained. These are called *cubic smoothing splines*. Figure 11.7c shows a cubic smoothing spline that displays a good balance between the two desiderata: fit and smoothness. As the value of λ grows larger the curves become smoother and deviate more from the data (Fig. 11.7d). In the limit of $\lambda \rightarrow \infty$ one approaches the ultimate smooth curve that has no curvature at all, namely a linear fit. Of course, this may represent a severe misfit of the data. The leave-one-out cross-validation technique discussed in Section 10.3.4 may be applied to determine a best value for the smoothing parameter avoiding overfitting (too small λ) and underfitting (too large λ).

11.3.3 Other techniques

Recently some other techniques have been used to model in an empirical way non-linear relationships. We will give a short qualitative description of two such techniques, ACE and MARS. A full technical description of these techniques is outside the scope of this book.

11.3.3.1 Alternating Conditional Expectation (ACE)

ACE, which stands for Alternating Conditional Expectations, is a method for multiple regression where each of the predictors x_j is optimally transformed into a new variable $z_j = f_j(x)$, $j = 1,...,p$, which allows a better fit of the response y. As a simple example, if the true (unknown) relation reads $y = b_1 \sqrt{x_1} + b_2 x_2^3 + b_3 \log(x_3)$ then ACE aims to uncover the three non-linear transformations from the data. The transformation functions f_j, $j = 1,...,p$, which in principle will all be different, are not given in analytical form (a formula), but in tabular form (x_{ij} vs. $f_j(x_{ij})$, $i = 1,...,n$). Plotting the transformed vs. the original values and inspecting the scatter plot may suggest the nature of the non-linear transformation for each predictor variable. The response itself may also be transformed, say into $z_0 = f_0(y)$. In the latter case, the predictors x_j and the response y take equivalent roles, so that ACE becomes a correlation technique rather than a regression technique. Here, we only discuss the regression variant in which y is not transformed. The main idea is that the linear additive model

$$y = \sum z_j + e \tag{11.42}$$

may, for certain choices of transformations f_j, $j = 0,1,...,p$, be better satisfied than the linear model in terms of the original variables

$$y = \sum b_j x_j + e \tag{11.43}$$

The criterion for assessing and comparing the linear relationships is the multiple coefficient of determination, R^2. Thus, the task of ACE is to find non-linear transformations, not necessarily in parametric form, of all the variables involved that maximize R^2. The algorithm starts with a multiple linear regression (eq. 11.43), and we use the simple linear transformations ($z_1 = b_1x_1$, $z_2 = b_2x_2$, ...) as a first approximation to the optimal non-linear transformations.

$$y = z_1^{(1)} + z_2^{(1)} + \ldots + z_p^{(1)} + e \tag{11.44}$$

where $z_j^{(1)} = b_j x_j$ ($j = 1,..,p$), the superscript indicating the iteration number. Notice that we have absorbed the proportionality constants b_j into the transformed variable. A still better fit can be obtained if the x-variables are also transformed non-linearly. This is done in turn for each x_j, giving the x-variable in question the temporary status of a response variable. Starting with x_1, we rearrange eq. (11.44) into

$$z_1^{(1)} = y - z_2^{(1)} - z_3^{(1)} - \ldots - z_p^{(1)} - e \tag{11.45}$$

This now shows z_1 as the 'dependent' variable and y as a predictor. Next we try and obtain an update for z_1 through a process called *back-fitting*. For this, a smooth curve is fitted through the scatterplot of $z_1^{(1)}$ against $(y - z_2^{(1)} - z_3^{(1)} - \ldots - z_p^{(1)})$. In its simplest form we can move a window (e.g. spanning 20% of the data points) along the x_1 axis and compute the expectation of z_1 for each data point as the (local) average of $(y - z_2^{(1)} - z_3^{(1)} - \ldots - z_p^{(1)})$. Using the updated transformation for z_1 we proceed to fit z_2 to the other variables, i.e.

$$z_2^{(1)} = y - z_1^{(2)} - z_3^{(1)} - \ldots - z_p^{(1)} - e \tag{11.46}$$

When all variables have been transformed according to this back-fitting procedure we start again with a new cycle ($z_j^{(2)}$, $j = 1,...,p$). Such cycles of alternatingly updating the variable transformations are repeated until convergence.

Estimating the expectation functions $f_j(.)$ can be done in different ways. In the original ACE algorithm [9] it is done by a local smoothing operation. In the related MORALS algorithm (MORALS = Multiple Optimal Regression by Alternating Least Squares) spline regression is used [10]. It is also possible to restrict the non-linear transformations to a certain class, e.g. monotone transformations which preserve ranking order. Another closely related method is AVAS [11]. In this method also care is taken to stabilize the variance. AVAS (= Additivity and VAriance Stabilization) is claimed to be better suited for predictive modelling than the correlation-based ACE method.

As an example Fig. 11.8 shows the transformation plots (z_j versus x_j) for the predictor variables in a QSAR investigation [12]. The x-variables represent structural parameters of a set of 6-anilinouracils. The response y was the enzyme inhibitor activity. Non-linear transformations (quadratic or piecewise linear) of the predictors are clearly indicated. The best transformation of the response appeared to be nearly linear.

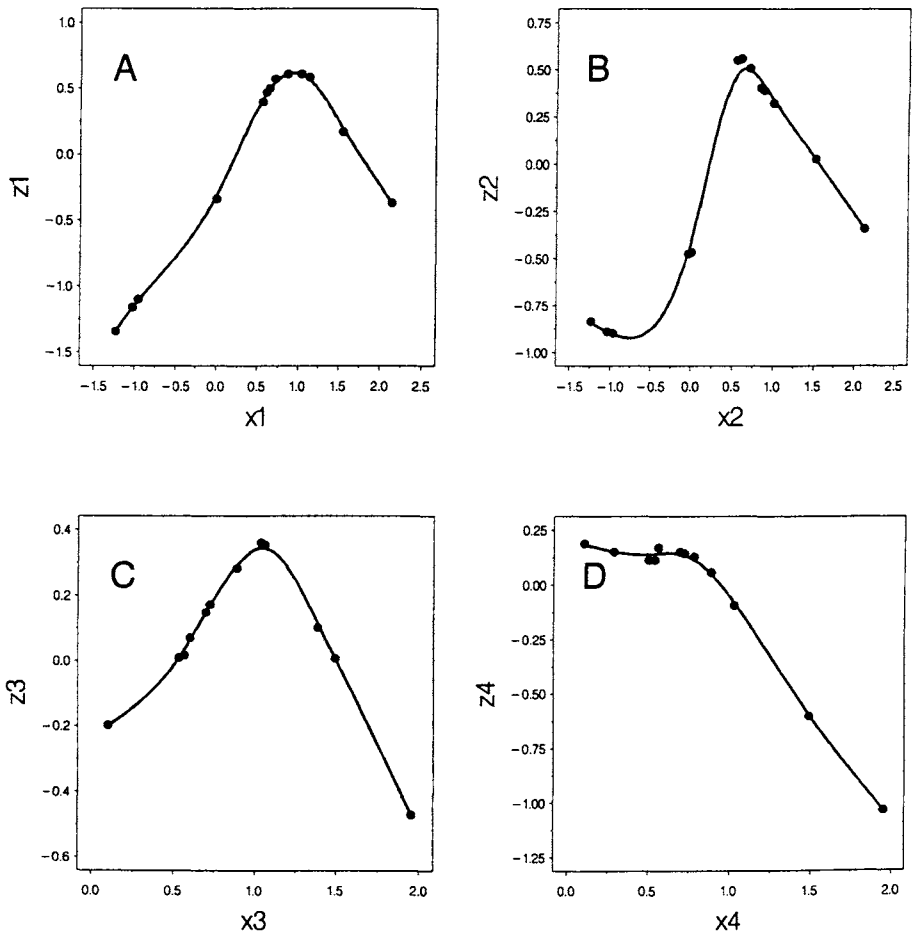


Fig. 11.8. ACE transformations of the predictor variables in QSAR data set [9].

The optimal transformation approach of ACE (or MORALS) is best applicable when there are many observations compared to the number of predictors. The technique is especially suitable for data which have been measured on non-linear scales. For example, with rank scales one only has information on the relative ordering of items, the distances between successive categories are not necessarily equal. One may consider applying monotone transformations which affect the distances along the scale, which improve the fit in a regression model and still preserve the original ranking of the items. Notice that the transformations can be very flexible but that the model remains additive, i.e. it does not accommodate for interactions between variables. A general introduction to these additive models is the book of Hastie and Tibshirani [13].

11.3.3.2 Multivariate Adaptive Regression Splines (MARS)

In principle the idea of spline fitting can also be applied in a multivariate setting. In practice this is a difficult task since the problem of placing the knots becomes much harder. Consider a problem with 5 variables and 2 knots per variables. This already generates 3 regions per variable, or $3^5 = 243$ subregions in 5-dimensional predictor space. If we fit each region with a constant value, i.e. the average response for that region, we need at least one data point per region, i.e. at least 243 observations in total. Such a model would be equivalent to an ANOVA model containing all high-order interactions. It would be discontinuous at the boundaries. As a first step toward a continuous response surface, we should fit each subregion with a first-degree model requiring 6 parameters to be estimated. A very modest number of 10–20 observations would be necessary to estimate such a local model. Thus, we would need at least 1000, preferably over 3000, observations to fit the overall model. More often than not, such a large number of data points is not available. A possible way out of the dilemma is to search for a multivariate spline model that is more parsimonious in the number of subregions.

The recent technique of *Multiple Adaptive Regression Splines* (MARS) combines forward variable selection with spline fitting to develop a non-linear multiple regression model [14]. The model can be written as a summation of a few *basis functions*. Each basis function is a piecewise polynomial associated with a variable and a certain range of that variable. It is possible to include interaction terms involving two or three variables and which are active in a localized region of the variables involved. The task of MARS is to select the important variables and to determine the subregions for each variable by optimal location of a knot. Like many other advanced non-linear multiple regression methods, MARS can only be applied when a large number of observations is available.

Figure 11.9 gives a simple two-dimensional response surface that can be fitted by a superposition of the two contributions shown separately and a localized interaction term. The model can be written as:

$$y = f_1(x_1 - t_1)_+ + f_2(x_1 - t_1)_- + f_3(x_2 - t_2)_+ + f_4(x_2 - t_2)_- + f_5((x_1 - t_1)_+(x_2 - t_3)_+) \quad (11.47)$$

where the notation $(.)_+$ indicates that the result is set to zero when the argument is not positive (see eqs. (11.40a and b)). Likewise, the minus suffix in $(.)_-$ indicates that the result is left unaltered for negative values of the argument and set to zero for positive values. The parameters t_1, t_2 and t_3 are knot positions. The knot position t_1 belongs to x_1, whereas t_2 and t_3 are knot positions for x_2. Each knot splits the experimental range of the predictor involved in two sub-regions that are separately modelled.

As drawn in Fig. 11.9, the functions f_1 to f_4 are simple linear functions making the response surface not smooth at the knot positions. Notice that the last term in eq. (11.47) is an interaction term that is only active in a local region, viz. in the

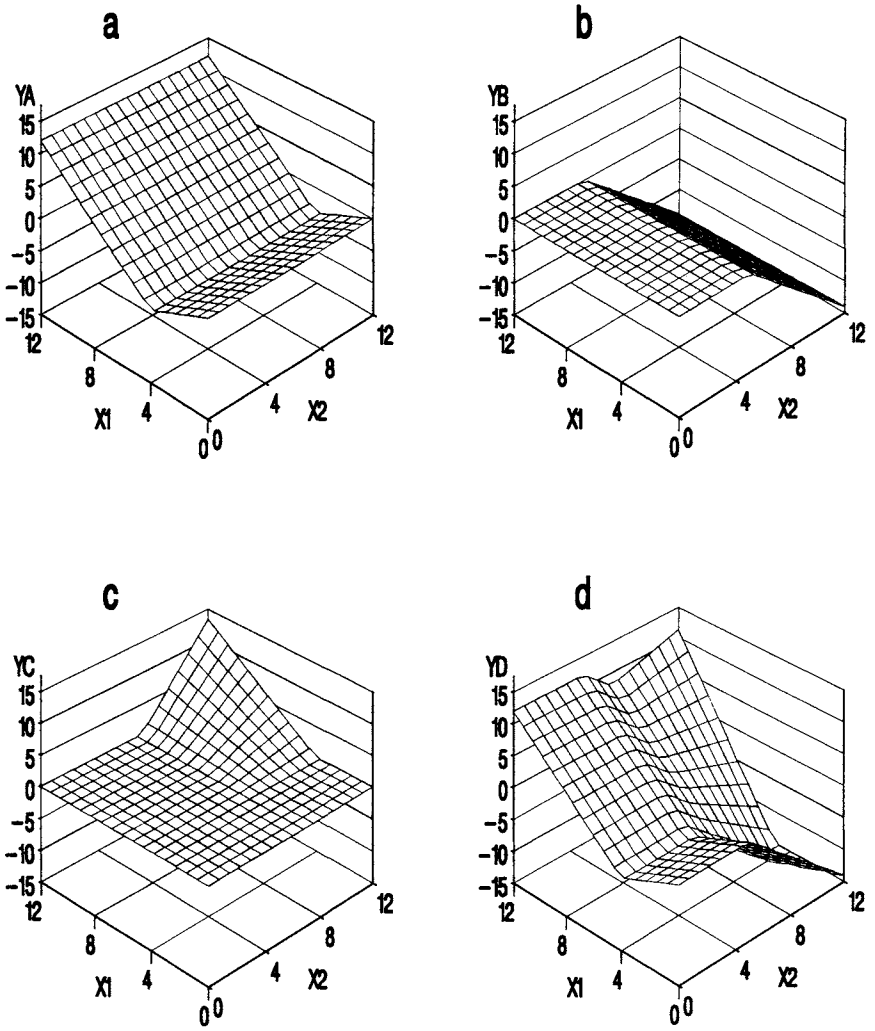


Fig. 11.9. MARS model as a superposition of univariate and bivariate spline models. (a) linear spline along x_1; (b) linear spline along x_2; (c) local $x_1 \times x_2$ interaction spline; (d) total model combining the three contributions.

corner for which $x_1 > t_1$ and $x_2 > t_3$. By going from a first-degree spline to a spline of degree 2 or 3 a smoother impression of the fitted surface is obtained. Quadratic or cubic functions give continuity in the first or second derivative. There are several levels of complexity of a MARS model having to do with the number of terms (basis functions) in the model, the degree of the piecewise polynomials (splines) and the level of interactions allowed (no interactions, only two-variable interactions, or three-variable interactions).

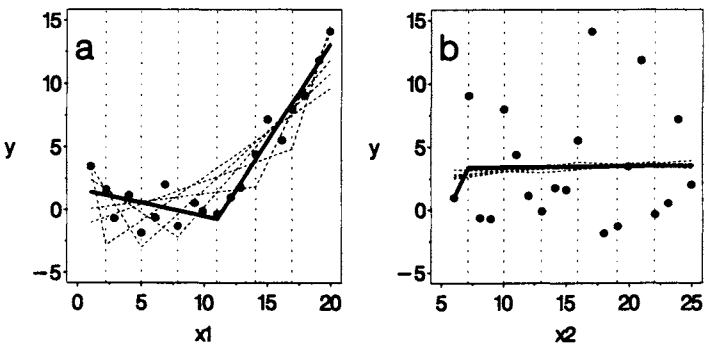


Fig. 11.10. Search for the best variable and its optimal knot position for explaining the response y. Predictor variable x_1 (a) explains the response better than variable x_2 (b). For each predictor variable the solid line presents the best fitting single-knot linear regression spline.

How are the variables and the knot positions selected? The method starts by choosing the first predictor variable, x_1, as a candidate for the first pair of basis functions. Each value x_{i1}, $(i = 1,..,n)$ observed in the data set for variable x_1 is considered as a candidate for placing the first knot t. The knot splits variable x_1 into two segments: observations with a lower value ($x_{i1} \leq t$) and observations with a higher value than the knot ($x_{i1} > t$). A simple regression of y on x_1 is done for the two subsets of data, i.e. for the two segments of data space. The error sum of squares is recorded. The best knot position is the one corresponding to the best fit, i.e. smallest error sum of squares (see Fig. 11.10a). This is done for each variable in turn. With p predictor variables and n observations one must consider p (number of variables) times n (number of knot positions for each variable) pairs of simple regressions. The variable x_j and knot position $t = x_{ij}$ giving the best fit among the np candidates is selected (compare Figs. 11.10a and b). This then establishes the first two basis functions.

This strategy can be seen as a search for the best way to split the experimental region in two parts which are then separately modelled. The process is repeated with the residual values of the response, now searching for a different way of splitting the experimental region in two parts that best explains the remaining variation in y. Quite likely a different variable will be selected, although it is possible that the same variable is chosen but then at a different knot location.

Once subregions are formed, one may also consider splitting a subregion only (*cf* the last interaction term in the example given). The model is expanded in this forward manner to an extent that it is deliberately overfitting. Then it is checked whether certain terms in the model can be dropped or neighbouring (sub)regions be merged. In either case the complexity of the model is reduced leading to more reliable predictions. This backward elimination proceeds until the criterion for determining the optimal model complexity minimum is reached. The criterion

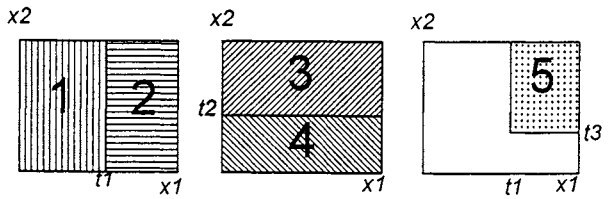


Fig. 11.11. Optimal sequential splits of the experimental (x_1,x_2)-region for developing a MARS model.

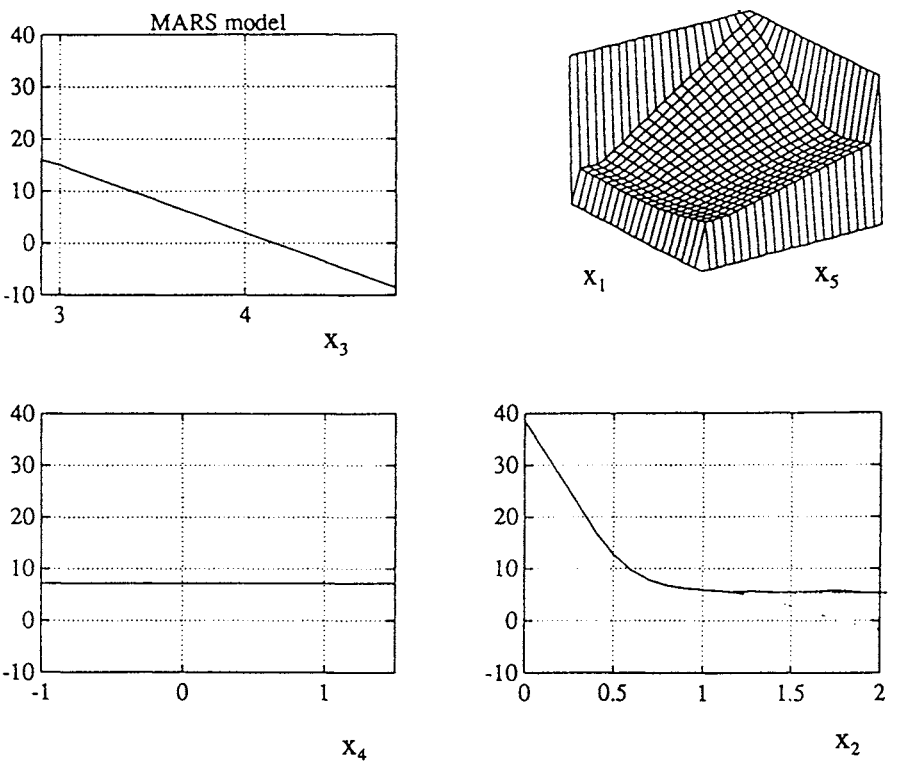


Fig. 11.12. Additive components of MARS model for polymer property data.

equals $(SS/n)/(1 - kM/n)^2$, where SS is the error sum of squares, M is the (final) number of basis functions and $k(\approx 4)$ represents an empirically determined factor that puts a penalty on each additional basis function.

Figure 11.11 shows a simple example of two predictor variables only. The first split is on x_2, the second on x_1 and the third on x_1 again, but only for the high x_2 subregion. Notice that the last splitting introduces a strong interaction term as it involves the levels of two variables. The advantage of the MARS approach is that

it allows a representation of the model as a number of additive terms. For example, the final model of a five variable system may be presented as in Fig. 11.12. The model relates a physical property of liquid detergents to molecular structural parameters of the polymeric system. Variable 3 enters the model linearly, variable 2 as a non-linear function with a plateau, and variables 1 and 5 show a strong interaction. Variable 4 has very little effect and does not enter the model.

11.3.3.3 Recent developments

The estimation of nonparametric non-linear models involving many predictors is an area of great research interest [15,16]. Artificial neural networks provide a powerful technique for modelling non-linear relations between multivariate **X** and multivariate **Y** (see Chapter 43). Whereas such ANN models often have good predictive properties their interpretation is quite difficult. There are also non-linear versions of popular multivariate regression techniques e.g. quadratic partial least squares (PLS, see Chapter 35) [17], splines-PLS [18] or locally weighted regression (LWR) in conjunction with PCR [19]. Finally, there is a growing tendency to exploit the growing computing power and the insight offered by interactive computer graphics, leading to methods such as Projection Pursuit regression [20]. Genetic algorithms have been used to advantage in ill-determined curve fitting problems [21]. Another computer-intensive natural computation approach is to assemble models from a set of basic functions (constant, x, sin, log, exp) and elementary operators ($*$, /, + , −, $\sqrt{}$, exp, log, power) and to utilize genetic algorithms (Chapter 27) in the search for plausible and well-fitting models among the vast number of possibilities [22].

References

1. R.D.B. Fraser and E. Suzuki, Resolution of overlapping bands: functions for simulating bandshapes. Anal. Chem., 41 (1969) 37–39.
2. A.C. Atkinson, Beyond response surfaces: recent developments in optimum experimental design. Chemom. Intell. Lab. Syst., 28 (1995) 35–47.
3. D.M. Bates and D.G. Watts, Non-linear Regression Analysis and its Applications. Wiley, New York, 1988.
4. G.A.F. Seber and C.J. Wild, Non-linear Regression. Wiley, New York, 1989.
5. A.C. Atkinson and A.N. Donev, Optimum Experimental Designs. Clarendon Press, Oxford, 1992.
6. S. Wold, Spline-funktioner — ett nytt verktyg i data-analysen. Kem. Tidskr., no. 3 (1972) 34–37.
7. S. Wold, Spline functions in data analysis. Technometrics, 16 (1974) 1–11.
8. C. de Boor, A Practical Guide to Splines. Springer, New York, 1978.
9. L. Breiman and J.H. Friedman, Estimating optimal transformation for multiple regression and correlation (with discussion). J. Am. Statist. Assoc., 80 (1985) 580–619.

10. F.W. Young, J. de Leeuw and Y. Takane, Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. Psychometrika, 41 (1976) 505–528.
11. R. Tibshirani, Estimating transformations for regression via additivity and variance stabilization. J. Am. Statist. Assoc., 83 (1988) 394–405.
12. I.E. Frank and S. Lanteri, ACE: a non-linear regression model. Chemom. Intell. Lab. Syst., 3 (1988) 301–313.
13. T.J. Hastie and R.J. Tibshirani, Generalized Additive Models. Chapman, London, 1990.
14. J.H. Friedman, Multivariate adaptive regression splines (with discussion). Ann. Statist., 19 (1991) 1–141.
15. S. Sekulic, M.B. Seasholtz, Z. Wang, B.R. Kowalski, S.E. Lee and B.E. Holt, Non-linear multivariate calibration methods in analytical chemistry. Anal. Chem., 65 (1993) 835A–844A.
16. I.E. Frank, Modern non-linear regression methods. Chemom. Intell. Lab. Syst., 27 (1995) 1–19.
17. A. Höskuldsson, Quadratic PLS regression. J. Chemom., 6 (1992) 307–334.
18. S. Wold, Non-linear partial least squares modelling. II. Spline inner relation. Chemom. Intell. Lab. Syst., 14 (1992) 71–84.
19. T. Næs, T Isaksson and B.R. Kowalski, Locally weighted regression and scatter correction for near-infrared reflectance data. Anal. Chem., 62 (1990) 664–673.
20. J.H. Friedman and W. Stuetzle, Projection pursuit regression. J. Am. Statist. Assoc., 76 (1981) 817–823.
21. A.P. de Weijer, L. Buydens, G. Kateman and H.M. Heuvel, Spectral curve fitting of infrared spectra obtained from semi-crystalline polyester yarns. Chemom. Intell. Lab. Syst., 28 (1995) 149–164.
22. V. Babovich and M.B. Abbott, The evolution of equations from hydraulic data. J. Hydraulics Res., 35 (1997) 15–21.

Additional recommended reading

Books

N.R. Draper and H. Smith, Applied Regression Analysis, 2nd edn. Wiley, New York, 1981.
D.A. Ratkowsky, Nonlinear Regression Modelling: a Unified Practical Approach. Dekker, New York, 1983.
G.E.P.Box and N.R. Draper, Empirical Model Building and Response Surfaces. Wiley, New York, 1987.
R.L. Eubank, Spline Smoothing and Nonparametric Regression. Dekker, New York, 1988.
A. Gifi, Nonlinear Multivariate Analysis. Wiley, New York, 1990.
G.J.S. Ross, Nonlinear Estimation. Springer, Berlin, 1990.
G. Whaba, Spline Models for Observational Data. SIAM, Philadelphia, 1990.
J.M. Chambers and T.J. Hastie, eds., Statistical Models in S. Chapman and Hall, London, 1993.

Articles

K.R. Beebe and B.R. Kowalski, Nonlinear calibration using projection pursuit regression: application to an array of ion-selective electrodes. Anal. Chem., 60 (1988) 2273–2278.
W.S. Cleveland and S.J. Devlin, Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Statist. Assoc., 83 (1988) 596–640.
R. Danielsson and G. Malmquist, Multi-dimensional simplex interpolation. An approach to local models for prediction. Chemom. Intell. Lab. Syst., 14 (1992) 115–128.

R.D. DeVeaux, D.C. Psichogios and L.H. Ungar, A comparison of two nonparametric estimation schemes: MARS and neural networks. Comput. Chem. Eng., 17 (1993) 813–837.

R.I. Jenrich and M.L. Ralston, Fitting nonlinear models to data. Ann. Rev. Biophys. Bioeng., 8 (1979) 195–238.

W.S. Cleveland, Robust locally weighted regression and smoothing scatter plots. J. Am. Statist. Assoc., 74 (1979) 828–836.

P. Geladi, D. McDougall and H. Martens, Linearization and scatter correction for near-infrared reflectance spectra of meat. Appl. Spectrosc., 39 (1985) 491–500.

P.J. Gemperline, Development in non-linear multivariate calibration. Chemom. Intell. Lab. Syst., 15 (1992) 115–126.

S. Sekulic and B.R. Kowalski, MARS: a tutorial. J. Chemom., 6 (1992) 199–216.

A.P. de Weijer, L. Buydens, G. Kateman, and H.M. Heuvel, Neural networks used as a soft modelling technique for quantitative dscription of the relation between physical structure and mechanical properties of poly(ethylene terephthalenes) yarns. Chemom. Intell. Lab. Syst., 16 (1992) 77–86.

Chapter 12

Robust Statistics

12.1 Methods based on the median

12.1.1 Introduction

All tests described so far have been based on the normal distribution. In applying these tests it is assumed that the mean and the standard deviation are representative measures of the central tendency and of the dispersion of the data examined, respectively.

Here we introduce some methods in which no assumptions about the distribution of the population is made. Therefore they are called *non-parametric or distribution-free methods*. Since they are also resistant to outlying observations, which have a large effect on the mean and the standard deviation, these tests are also identified as *robust methods*.

We start with a discussion of some descriptive robust statistics and their application for a visual inspection of the data. Different methods are then discussed which are based on a ranking of the observations and make use of the median. In Section 12.2 some other approaches are described.

12.1.2 The median and the interquartile range

The *median* is the value such that 50% of the observations are smaller (or larger). It is obtained by ranking the n data. When n is odd the median is the observation with rank $(n + 1)/2$; when n is even it is the mean of the observations with rank $n/2$ and rank $(n + 2)/2$.

As an example, let us consider the data of Table 12.1. To obtain the median for these data the measurements are ranked:

1.1	1.2	1.5	1.6	1.8	1.9	2.0	2.2	2.7	2.8
2.9	2.9	2.9	3.0	3.1	3.3	3.4	3.4	3.5	3.8
3.8	3.9	4.0	4.2	4.3	4.5	4.5	4.6	4.9	5.3
5.5	5.5	5.8	6.0	6.2					

TABLE 12.1

The determination of aflatoxin M in 7 laboratories (from Ref. [1])

Laboratory						
a	b	c	d	e	f	g
1.6	4.6	1.2	1.5	6.0	6.2	3.3
2.9	2.8	1.9	2.7	3.9	3.8	3.8
3.5	3.0	2.9	3.4	4.3	5.5	5.5
1.8	4.5	1.1	2.0	5.8	4.2	4.9
2.2	3.1	2.9	3.4	4.0	5.3	4.5

There are 35 values and consequently the median has rank $(35 + 1)/2 = 18$. Thus the median is 3.4 (the mean is 3.5). Sometimes the median gives a better idea of the central tendency than the mean because it is rather insensitive to the skewness of the distribution and to extreme values. The mean as well as the median of 15, 16, 17, 18, 19 equals 17. By the addition of the value 100 to this small data set the mean increases to 30.8. This is obviously not a good representative of the central tendency of the data since it exceeds 5 of the 6 observations. The median, on the other hand, hardly changes to 17.5.

The *first quartile* or *lower fourth*, F_L, is the value so that 25% of the observations are smaller. Similarly the *third quartile* or *upper fourth*, F_U, corresponds to the value that is exceeded by 25% of the observations. The second quartile is the median. The *fourth spread* or *interquartile range* (IQR) is computed as the difference between F_U and F_L. It represents the range containing the middle 50% of the data and therefore is a measure of spread. Note that in a normal distribution 50% of the observations are contained in a 1.35σ range. The IQR is less sensitive to extreme values than the standard deviation since it is not affected by values that lie beyond F_U and F_L.

In our example the median is the measurement with rank 18 and the lower fourth is obtained as the median of the first 18 ranked observations. Consequently, the lower fourth, F_L, is the mean of the observations with rank 9 and 10. These are respectively 2.7 and 2.8. Therefore $F_L = 2.75$. In a similar way the median of the last 18 observations, being the mean of observations with rank 26 and 27, corresponds to the upper fourth. Therefore $F_U = 4.5$ and the interquartile range is obtained as:

$$IQR = F_U - F_L = 4.5 - 2.75 = 1.75$$

A possible approach for identifying extreme values makes use of the IQR. The IQR is multiplied by 1.5 and the result is taken on both sides of the interquartile range. Values outside this interval (or acceptable range) are considered to be

outliers or at least extreme values that deserve close scrutiny. In our example values beyond $4.50 + 1.75 \times 1.5 = 7.13$ and $2.75 - 1.75 \times 1.5 = 0.13$ would be regarded as extreme. There are no such values in this case. Since in a normal distribution the IQR almost corresponds to $\frac{4}{3}\sigma$ the interval calculated above corresponds to about 5σ.

12.1.3 Box plots

All the parameters introduced in the previous section can be used to construct a *box and whisker plot* (or simply *box plot*) which allows a visual representation of the data. One constructs a box with ends corresponding to the lower and upper fourths in which the median is represented by a horizontal bar. From each end of the box a vertical line is then drawn to the most remote data point that is not an outlier. These most remote, non extreme values are pictured with a small horizontal line, called "whisker". For our example of the previous section, the box plot is represented in Fig. 12.1a. Since no outliers were identified the whiskers correspond to the lowest and the highest value in the data set, i.e. 1.1. and 6.2.

Outliers are indicated by a cross outside the whiskers. If in our example the highest value, 6.2, was replaced by the value 7.5 the resulting box plot would be the one represented in Fig. 12.1b. The box itself would be the same since neither the median nor the IQR would be affected by this change. Only the upper

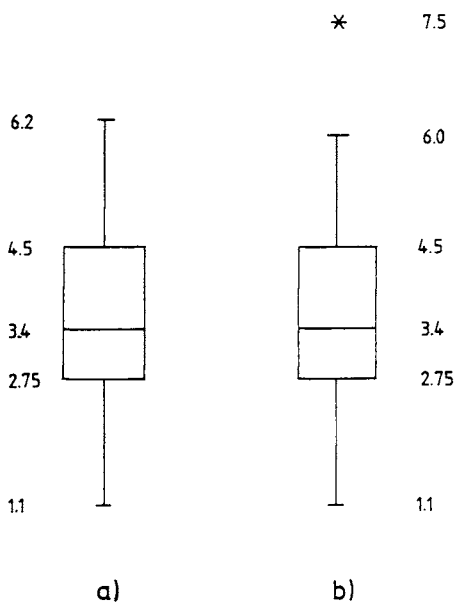


Fig. 12.1. An example of box plots.

whisker changes since the highest value which is not an outlier is now 6.0. The value 7.5 being larger than 7.13 (which was calculated in Section 12.1.2 as being the upper limit of the acceptable range) is indicated as an extreme value or outlier. Box plots allow a visual interpretation of the data. They contain information concerning the range (characterized by the whiskers), the spread (characterized by the length of the box) and the distribution of the observations (characterized by the position of the median and the box). A horizontal bar (representing the median) situated out of the middle of the box, for example, is an indication of a skewed distribution. The latter is illustrated in Fig. 12.2a, obtained

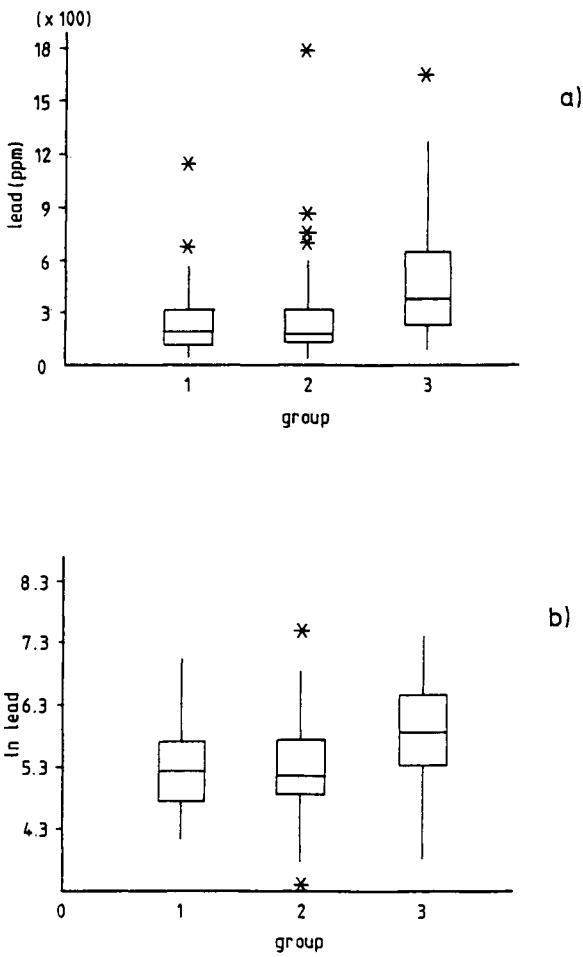


Fig. 12.2. Box and whisker plots of (a) the lead contents (in ppm) in surface enamel. Group 1, *in vivo* samples; group 2, *in vitro* urban group; and group 3, *in vitro* indust group. Asterisks indicate outlying lead values; (b) the natural logarithm of the lead data displayed in (a).

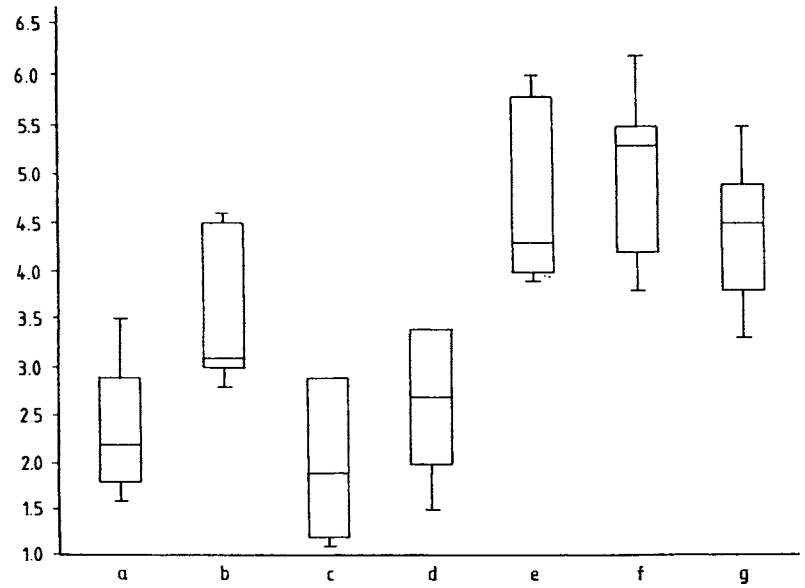


Fig. 12.3. Box plots used in the comparison of groups of data.

from Cleymaet et al. [2]. The example concerns the lead content of surface enamel. Acid etch surface enamel microbiopsies were taken from extracted permanent teeth from persons living in an urban area (urban groups) and from persons living within a distance of 10 km from a lead-polluting nonferrous-metal industrial plant (indust group). A third group of samples was obtained *in vivo* from adult volunteers living in an urbanized area (*in vivo* group). The box plots indicate that the three groups are characterized by a skewed distribution and that the indust group shows higher lead contents and a larger variation than the urban groups. The box plots of Fig. 12.2b show the effect of a logarithmic transformation of the lead content: the variance in the different groups is comparable and the skewness of the distribution is reduced.

Box plots are also useful for the comparison of different groups of data. Fig. 12.3 summarizes the box plots for the laboratories of Table 12.1. It is immediately obvious that some of the groups are very different from the others (e.g. c from f). Moreover, this plot also indicates that the spread in each of the groups is similar which means that the within laboratory precision is more or less the same. There seems to be homogeneity of variance so that the classical ANOVA can be applied for a further analysis of the data.

12.1.4 Hypothesis tests based on ranking

12.1.4.1 The sign test for two related samples

The sign test is a non-parametric alternative to the paired *t*-test (see Section 5.2) which makes use of positive and negative signs. To illustrate the test let us take the data from Table 12.2 in which the results obtained with a test method, x_T, are compared with those of a reference method, x_R. For each sample, the sign of the difference between x_R and x_T is considered. Differences equal to zero are not taken into account since they have no sign. If there is no true difference between the two paired samples the number of positive signs can be expected to be almost equal to the number of negative signs. In our example, seven out of the 10 differences are positive and three are negative. To test whether too few negative differences occur the binomial distribution with $p = q = 1/2$, which is discussed in Chapter 15, is used. However, statistical tables are available which contain the necessary information to perform the test. In Table 12.3 r represents the number of fewer signs and n the number of total signs. Since differences that are zero are not taken into account, n is smaller or equal to the number of paired observations. The table gives the probability that out of n (positive and negative) signs, the smaller number of like signs (here the – signs) is equal to or smaller than r. The probabilities given are for a one-sided test. They should be doubled for a two-sided test.

In our example $n = 10$ and $r = 3$. Since we only want to know whether there is a difference between both methods, the test is two-sided. For $n = 10$ the two-tailed probability that $r \leq 3$ is 0.344. This figure has to be compared with 0.05 if the test is performed at the 5% significance level. Since it indicates non-significance, the null hypothesis that there is no difference between both methods cannot be rejected.

TABLE 12.2

Data to illustrate the sign test and the Wilcoxon signed rank test for two paired samples

Sample	x_R	x_T	d_i	Sign	Rank	Signed rank
1	114	116	− 2	−	1	−1
2	49	42	+ 7	+	7.5	+7.5
3	100	95	+ 5	+	4	+4
4	20	10	+10	+	9.5	+9.5
5	90	94	− 4	−	2.5	−2.5
6	106	100	+ 6	+	5.5	+5.5
7	100	96	+ 4	+	2.5	+2.5
8	95	102	− 7	−	7.5	−7.5
9	160	150	+10	+	9.5	+9.5
10	110	104	+ 6	+	5.5	+5.5

TABLE 12.3

The sign test. The table gives the probability that out of n positive and negative signs, the smaller number of like signs is equal to or smaller than r. The values are for a one-sided test. They should be doubled for a two-sided test.

n\r	0	1	2	3	4	5	6	7	8	9	10
4	0.063	0.313	0.688								
5	0.031	0.188	0.500								
6	0.016	0.109	0.344	0.656							
7	0.008	0.062	0.227	0.500							
8	0.004	0.035	0.145	0.363	0.637						
9	0.002	0.020	0.090	0.254	0.500						
10	0.001	0.011	0.055	0.172	0.377	0.623					
11		0.006	0.033	0.113	0.274	0.500					
12		0.003	0.019	0.073	0.194	0.387	0.613				
13		0.002	0.011	0.046	0.133	0.291	0.500				
14		0.001	0.006	0.029	0.090	0.212	0.395	0.605			
15			0.004	0.018	0.059	0.151	0.304	0.500			
16			0.002	0.011	0.038	0.105	0.227	0.402	0.598		
17			0.001	0.006	0.025	0.072	0.166	0.315	0.500		
18			0.001	0.004	0.015	0.048	0.119	0.240	0.407	0.593	
19				0.002	0.010	0.032	0.084	0.180	0.324	0.500	
20				0.001	0.006	0.021	0.058	0.132	0.252	0.412	0.588

For large samples ($n > 25$) the binomial distribution can be approximated by a normal distribution [3] with:

mean = $\mu_x = 1/2\ n$

and

standard deviation = $\sigma_x = (1/2)\sqrt{n}$

The null hypothesis is then tested by computing

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{x - (1/2)\,n}{(1/2)\sqrt{n}}$$

and using one of the tables in Section 3.4.

12.1.4.2 The Wilcoxon signed rank test or the Wilcoxon t-test for two paired samples

A more powerful alternative to the paired t-test is the signed rank test. Besides the direction of the deviation between the observations, which is the only information used in the previously described sign test, the signed rank test also considers the magnitude of the deviation. Its main limitation is that it cannot be applied for a two-tailed test if $n \leq 6$. For the example of Table 12.2 the absolute values of d_i, the

TABLE 12.4
Critical values of the Wilcoxon signed rank test ($\alpha = 0.05$)

n	One-tailed	Two-tailed
6	2	0
7	3	2
8	5	3
9	8	5
10	10	8
11	13	10
12	17	13
13	21	17
14	25	21
15	30	25
16	35	30
17	41	35
18	47	40
19	53	46
20	60	52
21	67	59
22	75	66
23	83	73
24	91	81
25	100	89

differences for each pair of measurements are ranked. When ties are present the mean of the ranks is computed. For example here the value four occurs twice; they are both given the rank (2 + 3)/2 = 2.5. The next value, 5, is then given rank 4. Afterwards each rank is attributed the same sign as the original difference. If there is no true difference between the two paired samples there should not be a large difference between the sum of positive ranks (T^+) and that of negative ranks (T^-). The test consists in comparing $T = \min(T^+, T^-)$ to a critical value. The critical values for one and two tailed tests of significance at $\alpha = 0.05$ are given in Table 12.4. The null hypothesis is rejected if the calculated T is less or equal to the tabulated T. *Notice that in the parametric tests the null hypothesis is rejected if the calculated test-statistic is larger than the tabulated critical value.* In the example $T^+ = 44.0$ and $T^- = 11.0$ and therefore $T = 11$. For a two-sided test and $n = 10$, the critical value of T at $\alpha = 0.05$ is 8. It is concluded that the null hypothesis can be accepted and that there is no significant difference between the results of the two methods.

For large samples ($n > 25$) it can be shown [3] that the sum of ranks, T, is approximately normally distributed with mean

$$\mu_T = \frac{n(n+1)}{4}$$

and standard deviation

$$\sigma_T = \sqrt{\frac{n(n+1)\,(2n+1)}{24}}$$

The null hypothesis is then tested by computing

$$z = \frac{T - \mu_T}{\sigma_T}$$

and using one of the tables in Section 3.4

12.1.4.3 Mann–Whitney U-test for two independent samples

A powerful alternative to the parametric *t*-test for independent samples (see Section 5.1) is the Mann–Whitney *U*-test. As an example consider the following two groups of measurements that are to be compared:

A: 11.2; 13.7; 14.8; 11.1; 15.0; 16.1; 17.3; 10.9; 10.8; 11.7 $n_1 = 10$

B: 10.9; 11.2; 12.1; 12.4; 15.5; 14.6; 13.5; 10.8 $n_2 = 8$

First, all data are taken together and are ranked. When ties are present again the mean of the ranks is computed. This yields the ranking as given in Table 12.5.

If there is no true difference between both samples the ranks for A and B measurements should appear at random in the above list. The test consists in comparing the smaller of the following two test-statistics with the critical value for *U* in Table 12.6:

TABLE 12.5

Ranking of the measurements for groups A and B in the Mann–Whitney *U*-test

Group	Result	Rank
A	10.8	1.5
B	10.8	1.5
A	10.9	3.5
B	10.9	3.5
A	11.1	5
A	11.2	6.5
B	11.2	6.5
A	11.7	8
B	12.1	9
B	12.4	10
B	13.5	11
A	13.7	12
B	14.6	13
A	14.8	14
A	15.0	15
B	15.5	16
A	16.1	17
A	17.3	18

TABLE 12.6

Tables for the Mann–Whitney test. The following tables contain critical values of the U statistic for significance levels α equal to 5% and 10% for a two-sided test. If an observed U value is less than or equal to the value in the table, the null hypothesis may be rejected at the level of significance of the table.

$n_1 \backslash n_2$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Critical values of U for α equal to 5%																			
1																			
2							0	0	0	0	1	1	1	1	1	2	2	2	2
3				0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4			0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5		0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6		1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7		1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
Critical values of U for α equal to 10%																			
1																		0	0
2		0		0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
3		0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
4		1	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5	0	2	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6	0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7	0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
8	1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
9	1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
10	1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
11	1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
12	2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
13	2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
14	2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
15	3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
16	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
17	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
18	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
19	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
20	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad (12.1)$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where n_1 and n_2 = the smaller and the larger sample size, respectively; and R_1 and R_2 = the sum of the ranks for the group with sample size n_1 and n_2, respectively.

The null hypothesis is rejected if the test-statistic (the smaller of U_1 and U_2) is less or equal to the tabulated U.

For our example $n_1 = 8$, $n_2 = 10$, $R_1 = 70.5$ and $R_2 = 100.5$. Consequently, $U_1 = 45.5$ and $U_2 = 34.5$. The smaller of these values, i.e. 34.5, has to be compared with the critical value of U. For a two-sided test with $n_1 = 8$ and $n_2 = 10$ the 5% level of U, as obtained from Table 12.6, is 17. Therefore the null hypothesis is accepted and one concludes that there is no evidence for a difference between the two groups of measurements.

For large samples ($n_2 > 20$) U is approximately normally distributed with mean

$$\mu_u = \frac{n_1 n_2}{2}$$

and standard deviation

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

The null hypothesis is then tested by computing

$$z = \frac{U - \mu_u}{\sigma_u}$$

and using one of the tables in Section 3.4. A corrected standard deviation when a large amount of ties are present can be found in Siegel [3]. Different alternatives, requiring the use of different tables with critical values, for this test have been proposed.

12.1.4.4 Kruskal–Wallis one-way analysis of variance by ranks

In this section a non-parametric test is introduced for the comparison of k independent samples. To illustrate the method, the data from Table 12.1 will be used as an example. As in the previous test all data are first taken together and they are ranked. For ties the mean of the ranks is computed. In the table the original data are then replaced by their corresponding rank and the sum of the ranks in each column (= R_i with $i = 1, ..., k$) is calculated. For our example this results in Table 12.7. With this information the following test-statistic is calculated:

TABLE 12.7

Ranks for the data of Table 12.1

	a	b	c	d	e	f	g
	4	28	2	3	34	35	16
	12	10	6	9	22	20.5	20.5
	19	14	12	17.5	25	31.5	31.5
	5	26.5	1	7	33	24	29
	8	15	12	17.5	23	30	26.5
R_i	48	93.5	33	54	137	141	123.5

$$H = \frac{12}{N(N+1)} \left(\sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3\,(N+1) \tag{12.2}$$

where k = the number of samples; n_i = the number of observations in the ith sample; and $N = \sum n_i$, the total number of observations.

Since H is distributed approximately as χ^2 with $k-1$ degrees of freedom [4] the test consists in comparing the calculated H value with the tabulated χ^2 given in Table 5.4. The null hypothesis is rejected at the chosen level of significance if H is equal to or larger than the tabulated χ^2 value.

For our example

$$H = \frac{12}{30(30+1)} \left(\frac{48^2}{5} + \frac{93.5^2}{5} + \ldots + \frac{123.5^2}{5} \right) - 3(50+1)$$

$$= 24.94$$

Since $\chi^2_{0.05;6} = 12.59$ the null hypothesis is rejected. It is concluded that the results obtained by the seven laboratories differ significantly. In this example all samples are of equal size (all $n_i = 5$) but eq. (12.2) applies equally well with samples of different size.

The χ^2 approximation to the distribution of H is only valid if there are at least 5 observations in the different groups. Moreover with less than 5 observations the test should not be used at a significance level lower than 1%.

12.1.4.5 The Spearman rank correlation coefficient

This non-parametric correlation coefficient for measuring the degree of association between two variables y_1 and y_2 in a sample is calculated in the following way:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{12.3}$$

TABLE 12.8

Calculation of d_i^2 to obtain Spearman rank correlation between Cu and Zn data from Table 8.8

Brain structure	Cu	Zn	Rank Cu	Rank Zn	d_i	d_i^2
	($\mu g\ g^{-1}$ dry weight)					
1	25.8	78.0	9	11	−2	4
2	24.2	81.8	8	12	−4	16
3	27.3	69.4	10.5	9	1.5	2.25
4	32.8	76.1	12	10	2	4
5	27.3	62.5	10.5	8	2.5	6.25
6	17.9	60.1	7	7	0	0
7	14.0	34.2	5	2	3	9
8	13.3	35.5	4	3	1	1
9	10.0	33.3	1	1	0	0
10	10.9	38.9	3	4	−1	1
11	10.7	40.8	2	5	−3	9
12	16.0	46.4	6	6	0	0

where n = the number of paired observations; and d_i = the difference between the ranks given separately to the variables y_1 and y_2.

It can be shown [3] that r_s is the Pearson product–moment correlation coefficient r, as defined by eq. (8.58), between the ranks of y_1 and y_2.

Consider, for example, the Cu and Zn concentrations determined in 12 different structures of the human brain from Table 8.8. The calculation of r_s is illustrated in Table 12.8. In the case of ties, tied values have been given the average rank. For the example r_s is found to be 0.816 whereas the product–moment correlation coefficient calculated in Section 8.3.1. was 0.898. The significance of r_s (H_0: $\rho_s = 0$; H_1: $\rho_s \neq 0$) can be deduced from Table 12.9 which tabulates critical values of r_s. Since r_s is larger than the critical value at the 5% significance level, a significant correlation between Cu and Zn has been detected.

When n is larger than 25, r_s can also be tested as described in Section 8.3.2 for the Pearson correlation coefficient, r. There will be little error in using eq. (8.61) or Table 8.9.

12.1.4.6 Detection of trends by the runs test

In order to be able to draw conclusions about a population from a sample taken from that population the sample must be random. The *runs test* can be used to test the random sampling assumption if the original order in which the observations were obtained is known.

It is especially useful in testing the random sequence of observations. In Chapters 8 (Section 8.2.2.1) and 10 (Section 10.3.1.3) on regression we concluded

TABLE 12.9

Critical values of the Spearman rank correlation coefficient for a two-tailed test at different p values

n	$p = 0.10$	$p = 0.05$	$p = 0.01$
6	0.829	0.886	1.000
7	0.714	0.786	0.929
8	0.643	0.738	0.881
9	0.600	0.700	0.833
10	0.564	0.648	0.794
11	0.536	0.618	0.755
12	0.503	0.587	0.727
13	0.484	0.560	0.703
14	0.464	0.538	0.675
15	0.443	0.521	0.654
16	0.429	0.503	0.635
17	0.414	0.485	0.615
18	0.401	0.472	0.600
19	0.391	0.460	0.584
20	0.380	0.447	0.570
21	0.370	0.435	0.556
22	0.361	0.425	0.544
23	0.353	0.415	0.532
24	0.344	0.406	0.521
25	0.337	0.398	0.511

that a random sequence of positive and negative residuals $(y_i - \hat{y}_i)$, when plotted against $\hat{y}_i$, is an indication for the adequacy of the model used to fit the data. For the residuals plot in Fig. 8.5b a non-random arrangement of residuals was detected. Here we will show how we came to that conclusion by using the runs test. The following pattern of positive and negative residuals was obtained:

$$- - - - - - + + + + + + + + + - + - -$$

There are 19 residuals ($n = 19$), 9 of which are negative ($n_1 = 9$) and 10 of which are positive ($n_2 = 10$). A run being a sequence of identical signs, 5 runs ($r = 5$) are observed in these data: a run of 5 negative residuals is followed by a run of 9 positive residuals, a run of one negative residual, a run of one positive residual and finally a run of 2 negative residuals. Table 12.10 gives the critical values of r at $\alpha = 0.05$ for n_1 and n_2 less or equal to 20. For each combination of n_1 and n_2 two critical values are listed. An observed r value which is less than or equal to the smaller critical value or greater than or equal to the larger critical value results in a rejection of the hypothesis of a random arrangement at the 5% significance level. For our example with $n_1 = 9$ and $n_2 = 10$ a non-random sample would contain 5 or less runs or 16 or more runs. Since only 5 runs are observed a non-random arrangement of positive and negative residuals has been detected.

TABLE 12.10

Critical values of r in the runs test [4]

$n_1 \backslash n_2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2										2	2	2	2	2	2	2	2	2
										–	–	–	–	–	–	–	–	–
3				2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
				–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4			2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
			9	9	–	–	–	–	–	–	–	–	–	–	–	–	–	–
5		2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
		9	10	10	11	11	–	–	–	–	–	–	–	–	–	–	–	–
6	2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
	–	9	10	11	12	12	13	13	13	13	–	–	–	–	–	–	–	–
7	2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
	–	–	11	12	13	13	14	14	14	14	15	15	15	–	–	–	–	–
8	2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
	–	–	11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9	2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
	–	–	–	13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10	2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
	–	–	–	13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11	2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
	–	–	–	13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
	–	–	–	13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
	–	–	–	–	15	16	17	18	19	19	20	20	21	21	22	22	23	23
14	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
	–	–	–	–	15	16	17	18	19	20	20	21	22	22	23	23	23	24
15	2	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
	–	–	–	–	15	16	18	18	19	20	21	22	22	23	23	24	24	25
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	12	12
	–	–	–	–	–	–	17	18	19	20	21	21	22	23	23	24	25	25
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
	–	–	–	–	–	–	17	18	19	20	21	22	23	23	24	25	25	26	26
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
	–	–	–	–	–	–	17	18	19	20	21	22	23	24	25	25	26	26	27
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
	–	–	–	–	–	–	17	18	20	21	22	23	23	24	25	26	26	27	27
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14
	–	–	–	–	–	–	17	18	20	21	22	23	24	25	25	26	27	27	28

When either $n_1 > 20$ or $n_2 > 20$ a normal approximation may be used [3] with

$$\text{mean} = \mu_r = \frac{2n_1n_2}{n} + 1$$

where $n = n_1 + n_2$, and

$$\text{standard deviation} = \sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}$$

The null hypothesis is then tested by computing

$$z = \frac{r - \mu_r}{\sigma_r}$$

and using one of the tables in Section 3.4.

The runs test can also be used when the observations can be *dichotomized* (i.e. converted into two categories). Consider, for example, 20 successive measurements performed on a sample. To test whether there is a drift in the results, the runs test above and below the median can be used. Observations that are lower than the median are denoted by a negative sign and observations that are larger than the median by a positive sign. Observations that are equal to the mean are either disregarded [5] or are all given a positive or a negative sign [6].

In the following example the median is 8 (the average of the 10th and 11th measurement after ranking) and there are 9 runs:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Result	5	9	9	10	7	3	7	10	9	7	9	9	4	9	10	9	4	6	2	3
Sign	−	+	+	+	−	−	−	+	+	−	+	+	−	+	+	+	−	−	−	−

Since it follows from Table 12.10 that with $n_1 = 10$ and $n_2 = 10$ a non-random sample would contain 6 or less runs or 16 or more runs at the 5% significance level there is no drift in the results.

12.1.5 Median-based robust regression

The classical least-squares regression, which consists of minimizing the sum of the squared residuals assumes among others a normal error distribution. Consequently, the presence of outliers can have a large influence on the estimated parameters. The lack of robustness of the regression parameters is illustrated in Fig. 12.4. The hypothetical data consist of six points (0.0, 0.0), (1.0, 1.1), (2.0, 2.0), (3.0, 3.1), (4.0, 3.8) and (5.0, 10.0). It is clear that the last point is not representative for the linear model fitted by the rest of the data. The outlier in this straight line relationship attracts the regression line, computed by least squares, to such an extent that the estimated line is unacceptable. It could be argued that outliers can

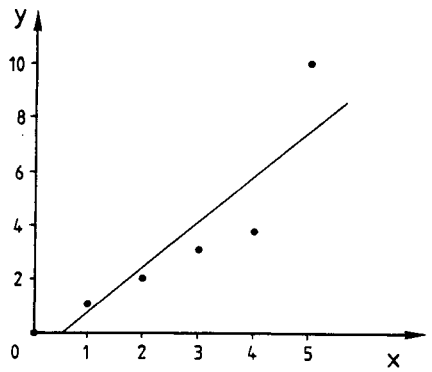


Fig. 12.4. Hypothetical data with an outlier. The line is the least squares line. The data are given in the upper part of Table 12.2.

be discovered by examining the least-squares residuals. Unfortunately this is not always true: the outlying point in Fig. 12.4 does not have a particularly larger residual than some other data points (see the upper part of Table 12.12).

The concept of robust estimation has been extended to regression analysis and different robust regression methods that resist the violations of the classical assumptions have been described. The resistance of a procedure to outliers is measured by means of the breakdown point. Hampel [7] defined the breakdown point as the smallest percentage of contaminated data (outliers) that can cause the regression estimators to take on arbitrarily large aberrant values. Since even one outlier can have a large influence on the least squares line, the least squares method has a breakdown point of 0%. Here some median-based robust regression methods will be described. Other robust methods in which weighting procedures are introduced to downweight the influence of outlying observations are given in Section 12.2.

12.1.5.1 Single median method

The single median method (SM) proposed by Theil [8] is the simplest median-based robust regression method for a straight line relationship. The slope, b_1, is estimated as the median of the $n(n-1)/2$ slopes between all pairs of data:

$$b_1 = \underset{i,j}{\text{med}}\,((y_j - y_i)/(x_j - x_i)) \qquad 1 \le i < j \le n \tag{12.4}$$

For the data in Fig. 12.4, $n = 6$ and $n(n-1)/2 = 15$. The slopes for all pairs of data are given in Table 12.11. The median of these slopes is 1.03. Consequently the single median slope, b_1, equals 1.03. The estimator of the intercept, b_0, can be obtained as the median of the intercepts calculated with this robust slope for all data points:

$$b_0 = \underset{i}{\text{med}}\,(y_i - b_1 x_i) \tag{12.5}$$

TABLE 12.11

Slopes for all pairs of data in Fig. 12.4

i	j	Slope
1	2	1.10
1	3	1.00
1	4	1.03
1	5	0.95
1	6	2.00
2	3	0.90
2	4	1.00
2	5	0.90
2	6	2.23
3	4	1.10
3	5	0.90
3	6	2.67
4	5	0.70
4	6	3.45
5	6	6.20

For our example the intercepts calculated are:

Point	1	2	3	4	5	6
Intercept	0.00	0.07	–0.06	0.01	–0.32	4.85

Consequently the single median intercept, b_0, equals 0.00 (0.005 rounded to 0.00) and the SM regression line is:

$y = 0.00 + 1.03\,x$

It can be shown [9] that this method has a breakdown point of 29%. In Table 12.12 (Data 1) notice the large residual from the robust SM fit for the outlying point. This indicates that the line is less influenced by the outlying point than the least squares line. However, if two outliers exist in these data (see Data 2 of Table 12.12) the contamination by outliers is too large to obtain correct estimators.

12.1.5.2 Repeated median method

The repeated median method (RM) is an improvement of the single median since the breakdown point is increased to 50%. In this method developed by Siegel [10] the slope and the intercept are obtained as:

$$b_1 = \underset{i}{\text{med}} \left(\underset{j \neq i}{\text{med}} \left((y_j - y_i) / (x_j - x_i) \right) \right) \tag{12.6}$$

$$b_0 = \underset{i}{\text{med}} \left(y_i - b_1 x_i \right) \tag{12.7}$$

TABLE 12.12

Comparison of least-squares and median-based robust regression methods

	x	y	LS	Residual		
				SM	RM	LMS
DATA 1	0.0	0.0	0.90	0.00	–0.03	0.00
(1 outlier)	1.0	1.1	0.30	0.07	0.06	0.07
	2.0	2.0	–0.49	–0.07	–0.06	–0.07
	3.0	3.1	–1.08	0.00	0.03	0.00
	4.0	3.8	–2.07	–0.33	–0.29	–0.33
	5.0	10.0	2.44	4.83	4.89	4.83
	regression	b_0	–0.90	0.00	0.03	0.00
	parameters	b_1	1.69	1.03	1.02	1.03
DATA 2	0.0	0.0	1.19	0.45	0.00	0.00
(2 outliers)	1.0	1.1	0.07	–0.45	0.00	0.07
	2.0	2.0	–1.26	–1.55	–0.20	–0.07
	3.0	3.1	–2.38	–2.45	–0.20	0.00
	4.0	10.0	2.30	2.45	5.60	5.87
	5.0	10.0	0.08	0.45	4.50	4.83
	regression	b_0	–1.19	–0.45	0.00	0.00
	parameters	b_1	2.22	2.00	1.10	1.03

First, for each of the n data points the median of the $(n-1)$ slopes between that point and all other points is calculated. Thus n medians are obtained and the median of these n medians is the repeated median estimator of the slope. The procedure is explained in Fig. 12.5 for the data from Fig. 12.4 which are also given in the upper part of Table 12.12. The lowest median is 0.90 for point 5 and the highest median is 2.67 for point 6. In Fig. 12.5b the 6 medians are ranked and the RM estimator of the slope, b_1, is the mean of the third and fourth ranked median values. It equals 1.02.

The estimation of the RM intercept is identical to the SM estimation of the intercept described in the previous section. Consequently, the RM line calculated for our example is

$$y = 0.03 + 1.02\,x$$

which again is not influenced by the outlying point. That the RM method is more robust than the SM method follows from the lower part of Table 12.12. Even with two outliers the repeated median method behaves well.

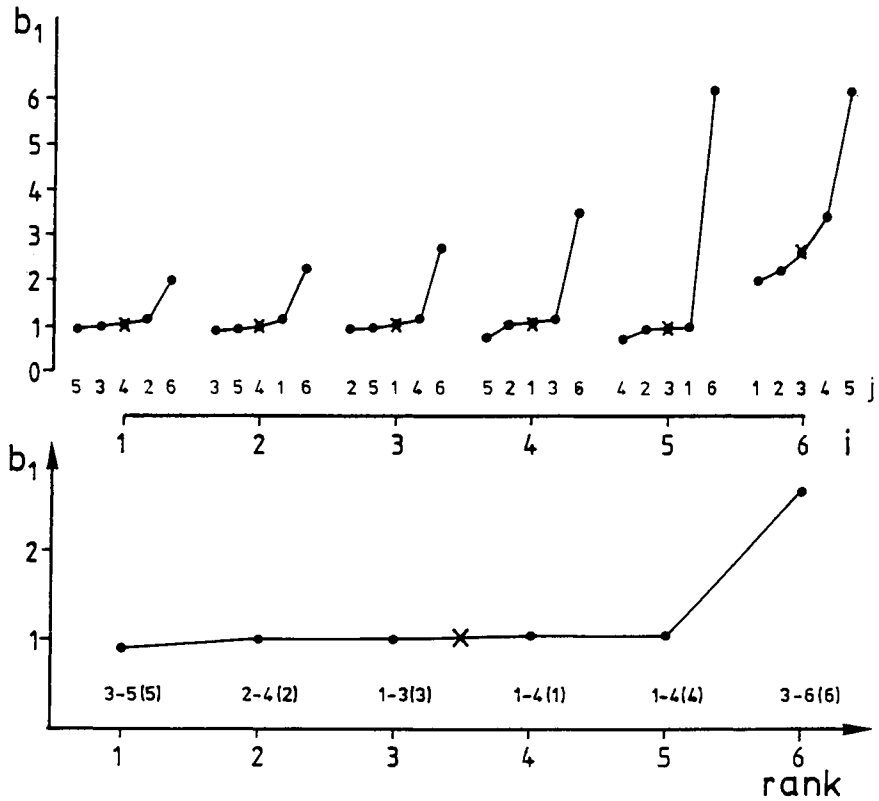


Fig. 12.5. Illustration of the repeated median method for the data in Table 12.2. (upper part). (a) Ranked slope b_1 for each point i, joined by a line to each of the other points; and (b) ranked median slopes selected from (a). Medians are indicated by a cross and 3–5(5)indicates that for point 5 the median slope is that of the line between points 3 and 5, etc.

12.1.5.3 The least median of squares (LMS) method

Another important median-based robust regression method, which is also applicable to the multiple regression situation, is the least median of squares method proposed by Rousseeuw [9] and first introduced in chemometrics by Massart et al. [11].

The LMS method is based on the minimization of the median of the squared residuals. For the straight line relationship this means:

$$\text{minimize med } (y_i - b_1 x_i - b_0)^2$$

with the median defined here as the ([n/2] + 1)th ranked value; [n/2] denotes the integer part of n/2. Notice that this definition of the median differs slightly from the one given in Section 12.1.2 if n is even.

In its simplest form, the slope and the intercept are estimated as follows: the lines between all possible pairs of points are calculated; with n data points this

TABLE 12.13

Least median of squares regression for the example of Fig. 12.4

i	j	b_1	b_0	Residuals (r)						med (r^2)
				1	2	3	4	5	6	
1	2	1.100	0.000	0.000	0.000	–0.200	–0.200	–0.600	4.500	0.040
1	3	1.000	0.000	0.000	0.000	0.000	0.100	–0.200	5.000	0.010
1	4	1.033	0.000	0.000	0.067	–0.067	0.000	–0.333	4.833	0.004
1	5	0.950	0.000	0.000	0.150	0.100	0.250	0.000	5.250	0.023
1	6	2.000	0.000	0.000	–0.900	–2.000	–2.900	–4.200	0.000	4.000
2	3	0.900	0.200	–0.200	0.000	0.000	0.200	0.000	5.300	0.040
2	4	1.000	0.100	–0.100	0.000	–0.100	0.000	–0.300	4.900	0.010
2	5	0.900	0.200	–0.200	0.000	0.000	0.200	0.000	5.300	0.040
2	6	2.225	–1.125	1.125	0.000	–1.325	–2.450	–3.975	0.000	1.756
3	4	1.100	–0.200	0.200	0.200	0.000	0.000	–0.400	4.700	0.040
3	5	0.900	0.200	–0.200	0.000	0.000	0.200	0.000	5.300	0.040
3	6	2.667	–3.333	3.333	1.767	0.000	–1.567	–3.533	0.000	3.121
4	5	0.700	1.000	–1.000	–0.600	–0.400	0.000	0.000	5.500	0.360
4	6	3.450	–7.250	7.250	4.900	2.350	0.000	–2.750	0.000	7.563
5	6	6.200	–21.000	21.000	15.900	10.600	5.500	0.000	0.000	112.360

yields $n(n-1)/2$ trial estimates for b_0 and b_1 each; for each line the squared residuals for all n data points are calculated; finally the line is retained for which the median of the squared residuals is minimal.

For our example of Fig. 12.4 (Data 1 in Table 12.12) the slopes and intercepts of the 15 lines between all pairs of the 6 data points as well as the residuals for 6 data points and the median of the squared residuals for each line are summarized in Table 12.13. From this table it follows that the median of the squared residuals (med(r^2)) is minimal for the line between point 1 and point 4. Consequently, the LMS line is

$$y = 0.00 + 1.03\,x$$

It should be noted that in this way the LMS line always exactly fits two of the data points. Rousseeuw [9] proposes an adjustment of the intercept by replacing the intercept term by the LMS location estimate of the n values:

$$b_{0(i)} = y_i - b_1\,x_i \qquad i = 1, \ldots, n \tag{12.8}$$

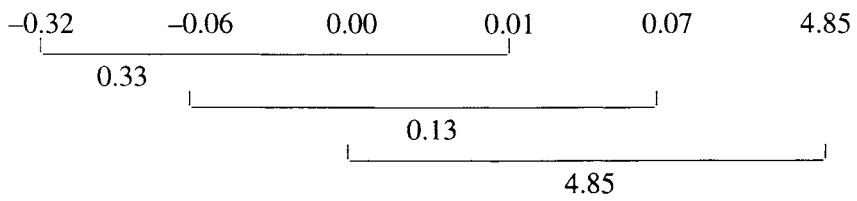
where b_1 represents the robust estimate of the slope determined as described above.

The LMS location estimate is the midpoint of the shortest half of the sample which is obtained as follows. After ranking the intercepts, by finding the smallest of the differences

$$b_{0(h)} - b_{0(1)},\ b_{0(h+1)} - b_{0(2)},\ \ldots,\ b_{0(n)} - b_{0(n-h+1)}$$

where $h = [n/2] + 1$.

In our example the ranked intercepts are:



The halves of the sample are indicated by the horizontal lines. The shortest half is between –0.06 and 0.07. Therefore the midpoint of this interval, namely 0.00, is the LMS location estimate.

Rousseeuw and Leroy [9] indicate that the standardized residuals resulting from a robust fit such as LMS can be used to diagnose outliers. The procedure first involves the calculation of an initial scale estimator, s^0:

$$s^0 = 1.4826(1 + 5/(n - 2))\sqrt{\text{med } e_i^2} \tag{12.9}$$

The expression is based on the median of the squared LMS residuals where 1.4826 is an asymptotic correction factor for the case of normal errors and $(1 + 5/(n - 2))$ is a finite sample correction factor. The latter is necessary to make s^0 unbiased when errors are normally distributed. If $|e_i/s^0| \leq 2.5$, the data point is retained, otherwise it is rejected. The final scale estimate, s^*, for LMS regression is then calculated:

$$s^* = \sqrt{\sum_{i=1}^{n^*} e_i^2 /(n^* - 2)} \tag{12.10}$$

where n^* represents the number of data points retained.

For the ultimate identification of an outlier, each observation is evaluated again: if $|e_i/s^*| \leq 2.5$ the data point is retained, otherwise it is rejected.

For our example where the LMS line is $\hat{y} = 0.00 + 1.03\,x$ this procedure yields the following results:

x_i	y_i	$\hat{y}_i$	e_i	e_i^2	e_i/s^*
0.0	0.0	0.00	0.00	0.0000	0.00
1.0	1.1	1.03	0.07	0.0049	0.36
2.0	2.0	2.06	–0.06	0.0036	–0.31
3.0	3.1	3.09	0.01	0.0001	0.05
4.0	3.8	4.12	–0.32	0.1024	–1.67
5.0	10.0	5.15	4.85	23.5225	25.26

The median of the squared residuals, here the 4th ranked value (see earlier), is 0.0049. Therefore from eq. (12.9), $s^0 = 1.4826\ (1 + 5/(6 - 2))\ \sqrt{0.0049} = 0.234$. Since for the last data point $|e_i/s^0| = 4.85/0.234 > 2.5$ this point is deleted and the

final scale estimate (eq. (12.10)) is calculated as $s^* = \sqrt{0.111/(5-2)} = 0.192$. The outlier test reveals that the last point is an outlier since $e_i/s^* = 25.26 > 2.5$.

In reweighted least squares based on LMS, proposed by Rousseeuw and Leroy [9], the outlier is given a zero weight and the classical least squares procedure is applied to the remaining data points. Reweighted least squares for the example yields the following regression equation: $\hat{y} = 0.08 + 0.96\,x$.

LMS can also be applied in multiple regression. For further information the reader is referred to the book by Rousseeuw and Leroy [9].

12.1.5.4 Comparison of least squares and different median based robust regression procedures

Hu et al. [12] used simulated data contaminated with outliers to compare different regression methods. For data that do not contain outliers least squares provides the best results, i.e. the least biased estimates of slope and intercept and the least variance. The robust regression methods also behave well: the estimated regression parameters are similar to the ones obtained by LS; however their dispersion is larger. When outliers are present the performance of LS degrades rapidly with increasing magnitude of the outlying observations. The effect is largest for outliers situated at the extreme points. Robust methods are then better suited, the best results (in terms of bias) being obtained for the LMS procedure.

The authors also applied the outlier diagnosis, as described in the previous section for LMS, to the other median-based robust regression methods. SM and RM detect only part of the outliers while in some situations LMS treats too many points as outliers. They propose the use of these methods in the exploratory validation of linear calibration lines and in suitability checks in routine calibration. Robust regression is applied to detect outliers in calibration lines found to have a bad quality, after which reweighted least squares is performed.

For the least-squares method confidence intervals for the regression parameters can be easily obtained (see Chapter 8). For the median based methods this seems less evident. Rousseeuw and Leroy [9] proposed (complicated) approximate confidence intervals for the LMS parameters. These confidence intervals are, of course, not needed if a reweighted least squares procedure is used.

12.2 Biweight and winsorized mean

Another approach to robust estimation is Mosteller and Tukey's *biweight* approach [13]. It is representative for a class of methods that use iterative weighting procedures to downweight the influence of outlying observations. It is computationally more complex, but, as shown later, it has other advantages. The biweight, x^*, is defined as:

$$x^* = \frac{\sum w_i x_i}{\sum w_i}$$

where

$$w_i = \begin{cases} \left(1 - ((x_i - x^*)/cS)^2\right)^2 & \text{when } ((x_i - x^*)/cS)^2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

with S a measure of spread such as half the interquartile range (1/2 IQR); and c a constant, usually 6.

For a normal distribution, where $S = (2/3)\sigma$ (see Section 12.1.2), $6S$ corresponds to 4σ and therefore observations that are more than 4σ away from the mean are given a zero weight.

Iterative calculations are required since, to obtain the different weights, w_i, one needs x^* and to obtain x^* the values of w_i must be known. As starting value for x^* the mean or the median can be used. Iteration proceeds until a stable value for x^* is obtained. An example adapted from Mosteller and Tukey [13] is given in Table 12.14. The biweight determined for the observations 7, 3, 3, –2, –5, –6, –21, is –0.79 (it can be checked that after a fifth iteration, not shown here, a stable value of –0.79 is obtained). This is quite different from the value for the median (–2) and of the mean (–3). Both the median and the biweight are less affected by the outlier than the mean, but the median is affected here by the fact that there is a rather large difference between the middle values 3, –2, and –5. For the example the biweight is probably a better measure of central tendency of these data.

To describe the influence of an outlier on the different measures of central location discussed here, we will consider a somewhat larger series of numbers [13] –8, –6, –5, –5, –2, 3, 3, 3, 7, 10 for which the mean, the median and the biweight are equal or very close to zero. Let us see how these measures of location behave when an 11th measurement, x, which takes different values is added. In Fig. 12.6 the effect of x on the mean ($\bar{x}$), the median (represented as $\tilde{x}$) and the biweight (x^*) is shown. One observes that the biweight performs best: it remains closest to zero and when x becomes an outlier it does not have any influence on the biweight. The median of course is most influenced by changes in the middle values of the data set. Once x reaches a value outside this middle range, the median is no longer influenced. The largest influence of a single observation is observed on the mean. If the observation is sufficiently outlying the mean becomes $+\infty$ or $-\infty$.

Mosteller and Tukey [13] recommend the use of

– median-based estimations in explorative data analysis

– the biweight (or related estimates such as the trimmed mean (see further)) when higher performance is needed for data that are not normally distributed or for which normality has not been verified;

TABLE 12.14

Example of biweight mean computation

First iteration: $x_1^* = \bar{x} = -3$

$S = 4.5$ = half the distance between 3 and –6 (in fact –5.5, but –6 is used for ease of computation)

x_i	$x_i - x_1^*$	$\lvert x_i - x_1^* \rvert / 6S = u_i$	u_i^2	$1 - u_i^2$	w_i	$w_i x_i$
7	10	0.37	0.14	0.86	0.74	5.18
3	6	0.22	0.05	0.95	0.90	2.70
3	6	0.22	0.05	0.95	0.90	2.70
–2	1	0.04	0.00	1.00	1.00	–2.00
–5	–2	0.07	0.00	1.00	1.00	–5.00
–6	–3	0.11	0.01	0.99	0.98	–5.88
–21	–18	0.67	0.45	0.55	0.30	–6.30
					$\Sigma =$ 5.82	$\Sigma =$ –8.60
						$x_2^* =$ –1.48

Second iteration:

x_i	$x_i - x_2^*$	$\lvert x_i - x_2^* \rvert / 6S = u_i$	u_i^2	$1 - u_i^2$	w_i	$w_i x_i$
7	8.48	0.31	0.10	0.90	0.81	5.67
3	4.48	0.17	0.03	0.97	0.94	2.82
3	4.48	0.17	0.03	0.97	0.94	2.82
–2	–0.52	0.02	0.00	1.00	1.00	–2.00
–5	–3.52	0.13	0.02	0.98	0.96	–4.80
–6	–4.52	0.17	0.03	0.97	0.94	–5.64
–21	–19.52	0.72	0.52	0.48	0.23	–4.83
					$\Sigma =$ 5.82	$\Sigma =$ –5.96
						$x_3^* =$ –1.02

Third iteration:

x_i	$x_i - x_3^*$	$\lvert x_i - x_3^* \rvert / 6S = u_i$	u_i^2	$1 - u_i^2$	w_i	$w_i x_i$
7	8.02	0.30	0.09	0.91	0.83	5.81
3	4.02	0.15	0.02	0.98	0.96	2.88
3	4.02	0.15	0.02	0.98	0.96	2.88
–2	–0.98	0.04	0.00	1.00	1.00	–2.00
–5	–3.98	0.15	0.02	0.98	0.96	–4.80
–6	–4.98	0.18	0.03	0.97	0.94	–5.64
–21	–19.98	0.74	0.55	0.45	0.20	–4.20
					$\Sigma =$ 5.85	$\Sigma =$ –5.07
						$x_4^* =$ –0.87

Fourth iteration:

x_i	$x_i - x_4^*$	$\lvert x_i - x_4^* \rvert / 6S = u_i$	u_i^2	$1 - u_i^2$	w_i	$w_i x_i$
7	7.87	0.29	0.08	0.92	0.85	5.95
3	3.87	0.14	0.02	0.98	0.96	2.88
3	3.87	0.14	0.02	0.98	0.96	2.88
–2	–1.13	0.04	0.00	1.00	1.00	–2.00
–5	–4.13	0.15	0.02	0.98	0.96	–4.80
–6	–5.13	0.19	0.04	0.96	0.92	–5.52
–21	–20.13	0.75	0.56	0.44	0.19	–3.99
					$\Sigma =$ 5.84	$\Sigma =$ –4.60
						$x_5^* =$ –0.79

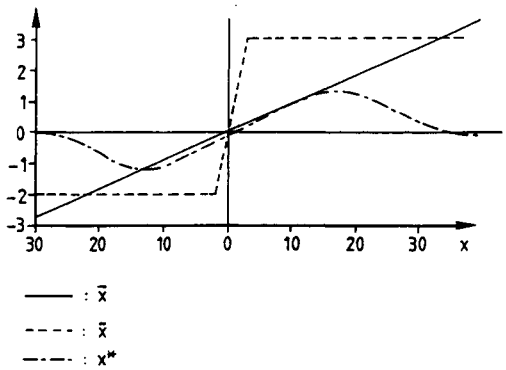


Fig. 12.6. The effect of an eleventh value x, added to the numbers –8, –6, –5, –5, –2, 3, 3, 3, 7, 10, on the mean ($\bar{x}$), the median ($\tilde{x}$) and the biweight (x^*).

– mean-based calculations when careful studies indicate that all aspects of the normal distribution are verified.

A somewhat similar approach is the *winsorized mean* [14] for which different proposals exist. Thompson [15] applied one of these to method validation in analytical chemistry. The robust estimate of the mean, x_w, is obtained as:

$$x_w = \sum x_{i(w)} / n$$

where the $x_{i(w)}$ are the winsorized values of x_i. They are obtained as follows:

$$x_{i(w)} = x_i \qquad \text{if } |x_i - x_w| < cS$$

$$x_{i(w)} = x_w + cS \qquad \text{if } x_i - x_w > cS$$

$$x_{i(w)} = x_w - cS \qquad \text{if } x_i - x_w < -cS$$

The value of c depends on the amount of outliers expected and a value of 1.5 is often used. S is a robust estimate of the standard deviation for which, with $c = 1.5$, Thompson uses the expression:

$$S^2 = \text{variance } x_{i(w)}/0.778$$

This means that values within a spread of $3S$ around the winsorized mean are used as such, while more outlying observations are given the less extreme value of $x_w + 1.5\,S$ or $x_w - 1.5\,S$.

As for the biweight approach, an iterative calculation of x_w and S is required. Initial estimates of these parameters used by Thompson are:

$$x_{w(0)} = \text{med } x_i$$

$$S_{(0)} = 1.483 \text{ med } (|x_i - \text{med } x_i|)$$

12.3 Iteratively reweighed least squares

In Section 12.1.5 median based robust regression methods, that make use of the robustness of the median as location estimator, have been described. In other robust methods weighting procedures are introduced in order to downweight the influence of possible outliers in the regression data. Different weight functions have been proposed among which the biweight described in the previous section. Iteratively reweighted least squares (IRLS) is a least-squares method in which at each iteration the observations are weighted. Weighted least squares is applied iteratively. At each iteration the regression coefficients are estimated and new weights based on the residuals are calculated. If the biweight is used the observations are weighted according to:

$$w_i = \begin{cases} 1 - (e_i / cS)^2)^2 & \text{when } |e_i| < cS \\ 0 & \text{otherwise} \end{cases}$$

where e_i is the residual, the deviation of the ith observation from its value predicted by the regression model ($= y_i - \hat{y}_i$); S is a robust measure of spread (the median of the absolute residuals is frequently used; then $S = (\text{med}\,|r_i|)$); c is a constant, usually 6.

Initially, all observations are given a weight equal to one. Consequently the starting values for the regression parameters are obtained from a simple least-squares procedure. From the least-squares fit new weights are calculated which are used to estimate new regression parameters by means of a weighted regression procedure (see Section 8.2.3.2). Iteration is continued until stable regression coefficients are obtained.

The technique of iteratively reweighted least squares will be illustrated with the hypothetical data of Section 12.1.5. Only the calculations for the first three steps are given in Table 12.15. A summary of the complete results is given in Table 12.16. The intermediate results were rounded to two decimal places. Stable regression coefficients are obtained after 6 iterations. The weight of the outlying observation is then zero while all good points reach weights very close to one. Therefore the regression equation obtained corresponds to the line through the first five data points.

Philips and Eyring [16] have proposed a correction for the initial least-squares estimates used as starting values since they suffer from lack of robustness and according to the authors can lead to incorrect results. The correction is based on a winsorizing of the residuals from the least squares fit. The authors [16] compared iteratively reweighted least squares and classical least squares regression (LS). They propose approximate confidence intervals for the IRLS regression parameters which are based on an estimate of the variance for the biweight, given in Mosteller and Tukey [13]. Philips and Eyring conclude that IRLS is superior to LS when errors are not normally distributed or when normal data are contaminated

TABLE 12.15

Computation of the first three steps in iteratively reweighted least squares applied to the hypothetical data of Section 12.1.5

Step 1	x	y			
	0.0	0.0			
	1.0	1.1			
	2.0	2.0	$\hat{y} = -0.90 + 1.69\,x$		
	3.0	3.1			
	4.0	3.8			
	5.0	10			
Step 2	x	$\hat{y}$	$\lvert e_i \rvert$	$\lvert e_i \rvert / 6S$	w_i
	0.0	−0.90	0.90	0.15	0.96
	1.0	0.79	0.31	0.05	1.00
	2.0	2.48	0.48	0.08	0.98
	3.0	4.17	1.07	0.18	0.94
	4.0	5.86	2.06	0.35	0.77
	5.0	7.55	2.45	0.41	0.69
		$S = (0.90 + 1.07)/2 = 0.99$			
		$\hat{y} = -0.78 + 1.62\,x$			
Step 3	x	$\hat{y}$	$\lvert e_i \rvert$	$\lvert e_i \rvert / 6S$	w_i
	0	−0.78	0.78	0.15	0.96
	1	0.84	0.26	0.05	1.00
	2	2.46	0.46	0.09	0.98
	3	4.08	0.98	0.19	0.92
	4	5.70	1.90	0.36	0.76
	5	7.32	2.68	0.51	0.55
		$S = (0.78 + 0.98)/2 = 0.88$			
		$\hat{y} = -0.66 + 1.54\,x$			

TABLE 12.16

Summary of the iteratively reweighted least squares applied to the hypothetical data of Section 12.1.5

x_i	y_i	w_i							
		Step: 1	2	3	4	5	6	7	8
0.0	0.0	1	0.96	0.96	0.96	0.96	0.96	0.96	0.96
1.0	1.1	1	1.00	1.00	1.00	1.00	0.98	0.98	0.98
2.0	2.0	1	0.98	0.98	0.98	0.98	1.00	1.00	1.00
3.0	3.1	1	0.94	0.92	0.92	0.92	0.92	0.92	0.90
4.0	3.8	1	0.77	0.76	0.74	0.71	0.86	0.92	0.92
5.0	10.0	1	0.69	0.55	0.34	0	0	0	0
	b_0	−0.90	−0.78	−0.66	−0.47	0.07	0.09	0.08	0.08
	b_1	1.69	1.62	1.54	1.39	0.97	0.96	0.96	0.96

with outliers. LS however is to be preferred in the ideal situation of normally distributed observations. This conclusion therefore is similar to the one made for median-based robust regression methods in Section 12.1.5.4.

12.4 Randomization tests

In a randomization test the probability (p) of falsely rejecting the null-hypothesis when in fact it is true, is not determined from statistical tables. Significance is determined from a distribution of the test statistic generated by randomly assigning the experimental data to the different conditions (i.e. groups, treatments, methods, etc.) studied.

An example of the randomized independent t-test will illustrate this. The example is obtained from Ref. [17]. Two treatments A and B which yield the following results: A: 18; 30; 54 and B: 6; 12 are compared. If the null hypothesis (H_0: $\mu_A = \mu_B$) is true one could randomly interchange the results for A and B without affecting the conclusion. The example is small to allow an illustration of the complete procedure. The t-value (see eq. (5.8)) calculated for these data is 1.81. The significance of this calculated t-value is determined by computing t for all permutations of the data. These are given in Table 12.17. For a one-tailed test (H_0: $\mu_A = \mu_B$; H_1: $\mu_A > \mu_B$) the ordered theoretical distribution of t: –3.00, –1.22, –0.83, –0.52, 0.00, 0.25, 0.52, 0.83, 1.22, 1.81 is considered. A t-value as high as the one obtained with the experimental data (1.81) occurs only once in the ten possible permutations. In this example $t \geq 1.81$ has a probability of 0.10 ($p = 0.10$). Therefore, if H_0 is true, the probability that the random assignment performed would result in a t-value as large as the one obtained with the experimental results is 0.10. If the pre-established level of significance $\alpha = 0.20$, H_0 is rejected.

TABLE 12.17

Data combinations obtained by permutation of the original data [17]

	A	B	A	B	A	B	A	B	A	B
	6	30	6	18	6	18	6	12	6	12
	12	54	12	54	12	30	18	54	18	30
	18		30		54		30		54	
$\bar{x}$	12	42	16	36	24	24	18	33	26	21
t		–3.0		–1.22		0.00		–0.83		0.25
	6	12	12	6	12	6	12	6	18	6
	30	18	18	54	18	30	30	18	30	12
	54		30		54		54		54	
$\bar{x}$	30	15	20	30	28	18	32	12	34	9
t		0.83		–0.52		0.52		1.22		1.81

TABLE 12.18

Comparison of two series of measurements; an outlier is present in the data of treatment B [17]

	Treatment A		Treatment B
	0.33		0.28
	0.27		0.80
	0.44		3.72
	0.28		1.16
	0.45		1.00
	0.55		0.63
	0.44		1.14
	0.76		0.33
	0.59		0.26
	0.01		0.63
$\bar{x}_A =$	0.412	$\bar{x}_B =$	0.995
	$\lvert t \rvert = 1.78$		

In general, the one-tailed probability for t for a randomization test is defined as the probability, if H_0 is true, to obtain a t-value at least as large as the obtained value. For a two-sided test (H_0: $\mu_A = \mu_B$, H_1: $\mu_A \neq \mu_B$) the probability for t is defined as the probability, if H_0 is true, to obtain a value of $|t|$ as large as the obtained value of $|t|$. For our example the following ordered theoretical distribution of $|t|$: 0, 0.25, 0.52, 0.52, 0.83, 0.83, 1.22, 1.22, 1.81, 3.00 is obtained. Two values are at least as large as the $|t|$ for the experimental data. Therefore $p = 0.20$ and for $\alpha = 0.20$ H_0 is rejected.

The data from Table 12.18 show that the test is robust. The outlier in B (3.72) causes the (two-tailed) parametric independent t-test to yield a non-significant result ($p = 0.092$). The randomization test yields $p = 0.026$. The presence of the outlier increases the standard deviation for group B, s_B, and reduces the value of t. Because s_B increases more than $|\bar{x}_A - \bar{x}_B|$ the t-value is lower.

A randomization test requires a great amount of computation, even for small samples. For the analysis of the results of Table 12.18 e.g. 184756 t values have to be computed, requiring a computer. For large samples the computer time can be reduced by using random data permutation programs [17]. With random data permutations the test statistics are only calculated for a given number of permutations from all those possible.

This has been applied in the next example which illustrates the randomization test procedure for one-way analysis of variance. The data are given in Table 12.19. Four laboratories analyzed the same sample containing trifluoperazine with the same titrimetric method. Each laboratory obtained 10 replicate results. The question is then whether there is a significant difference between the mean results obtained.

TABLE 12.19

Comparison of 4 laboratories for the analysis of trifluoperazine with the same titrimetric method

	1 (%)	2 (%)	3 (%)	4 (%)
	100.25	100.83	99.58	100.89
	100.29	100.82	99.76	100.99
	100.09	100.60	100.53	99.98
	100.49	99.43	100.1	100.41
	101.18	100.73	99.1	100.53
	101.32	100.85	100.3	101.05
	100.63	100.17	99.7	101.92
	100.90	101.22	99.2	100.12
	100.76	100.21	99.6	100.35
	100.46	98.96	99.6	100.08
$\bar{x}$ =	100.64	100.38	99.75	100.63
s =	0.41	0.71	0.45	0.60

The ANOVA takes the form of Table 6.3 with $k = 4$ and $n = 40$. Systematic data permutation would result in about 5×10^{21} permutated data sets. The random data permutation method was used to select 3000 permutations. To perform the randomization ANOVA test one could compute F for each of these permutated data sets and compute the probability for F as the proportion of the 3000 permutations that yield an F-value as large as the F-value for the original data set. Edgington [17] shows that one can advantageously use $(\sum(T_i^2/n_i)$, with T_i and n_i the sum and the number of experimental results for a particular condition (here method), to test the significance of F. Since for the example only 8 out of the 3000 permutations provide a value for $\sum(T_i^2/n_i)$ as large as that for the obtained data the probability of obtaining such a large F is 0.0026. The selection of another 3000 permutations confirmed the differences between the laboratories.

Besides the fact that randomization tests do not require the assumptions of normality and homoscedasticity they have an additional advantage. Randomization tests can also be applied when the random sampling assumption, which is the basis of all classical statistics, is violated. One such violation is systematic selection which may occur in some intercomparison studies in which only good laboratories participate.

12.5 Monte Carlo methods

Monte Carlo (MC) methods are part of the field of numerical simulation (Part B, Chapter 42) and play an important role in mathematics and statistics since their formal development in 1945. They originated within the context of the Manhattan

project which dealt with the design of the first atomic bomb and nuclear reactor. During that period the approach was code named Monte Carlo, which is also reminiscent of gambling and casinos. The founders of modern MC methods were Ulam, Metropolis, Fermi and von Neumann [18,19].

We will discuss two aspects of the MC approach. The first is referred to as *deterministic MC* and aims to determine theoretical quantities, that arise from differential equations and integrals, by means of simulated random events. The other is called *probabilistic MC* and is used for the simulation of properties of stochastic processes, such as the distribution of a random variable and the robustness of a statistical procedure.

12.5.1 Probabilistic MC for statistical methods

Due to the steady improvement in speed of computers and the decreasing cost of computing, Monte Carlo methods constitute a robust alternative to physical and statistical models that often have to introduce simplifying assumptions in order to obtain a manageable solution. The role of Monte Carlo methods in statistics can be compared to that of experimentation in the natural sciences [20]. Consequently, there also is a need for proper conduct and reporting of MC experiments.

The scope of the statistical MC also includes the permutation and randomization tests (Section 12.3) and of the resampling tests which are based on *bootstrapping* and *jack-knifing*. In bootstrapping one produces a number of samples of size n from an original sample of the same size n by means of random selection with replacement. From these bootstrapped data one can then compute various statistics, such as the confidence intervals for the median, interquartile range, etc. of the original sample. Jack-knifing employs a similar technique, but with the difference that resampling produces a predetermined number of samples of size $m < n$ by means of random selection without replacement. The statistics computed from the jack-knifed data are then corrected for the loss of degrees of freedom that resulted from drawing samples whose size is smaller than that of the original sample [21].

One of the earliest applications of this approach is attributed to Student (W.S. Gosset) for the study of the t-distribution for small samples, before the analytical form of the distribution was known [22]. The particular shape of the t-distribution can be studied empirically by repeatedly taking two samples of a given size n from the normal distribution. One then computes the t-statistic in the usual way:

$$t = (\bar{x}_1 - \bar{x}_2) / (s_p / \sqrt{n}) \text{ with } s_p^2 = (s_1^2 + s_2^2) / 2$$

where $\bar{x}_1, \bar{x}_2$ and s_1, s_2 represent the means and standard deviations of the two samples, respectively.

This operation is repeated a large number of times, after which the distribution of the t-values is plotted. Figure 12.7 shows the result of an MC simulation of the

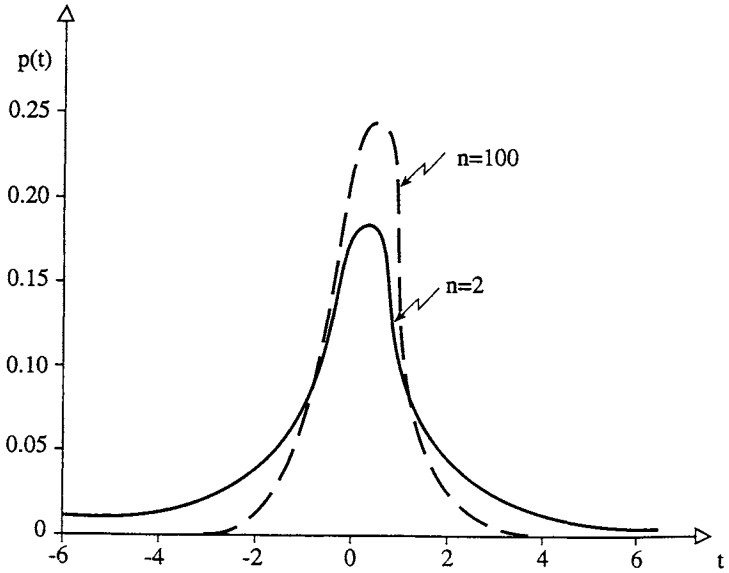


Fig. 12.7. Student's t-distributions for the difference of the means from two samples with size 2 and 100, respectively, drawn from a normal population. The distributions have been derived from 1,000 runs in a probabilistic Monte Carlo simulation.

t-distribution for n equal to 2 and 100, using 1000 runs in each case. One can observe the protracted tails of the t-distribution for small sample sizes and the asymptotic convergence toward the normal distribution for larger sample sizes.

The MC procedure can be employed for the study of distributions whose analytical form is difficult to obtain. Practical applications of this approach can be found in problems that involve waiting lines (queues) or random walks where the basic transition probabilities are not stationary, such as occurs in self-regulating and adaptive systems.

In MC models one often makes the simplifying assumption that the phenomena are normally distributed, with a given mean and standard deviation. An expedient way to generate *pseudo-normally distributed* random numbers follows from the central limit theorem. (The term pseudo-normal indicates that the numbers are only approximately normally distributed.) To this effect, 12 random numbers u_1, u_2, ... u_{12} are drawn from the uniform distribution between 0 and 1. It can be shown that the pseudo-normal variable z is given by:

$$z = \left(\sum_{i=1}^{12} u_i \right) - 6 \tag{12.11}$$

A practical application of probabilistic MC is the robust determination of the minimal sample size n for a minimal detectable difference d in a test for the

comparison of two means, when the distribution of the test statistic x is known but not normal [23]. The procedure is as follows. First, one constructs an alternative distribution by shifting the original distribution of x by the amount d. Then one takes a sample of a relatively small size n from each of the two distributions. The preferred two-sample test is performed and one notes whether the outcome is significant at the stated level of significance α (e.g. 0.05) or not. The sampling is repeated a large number of times (of the order of 10000) and the fraction of significant outcomes is determined. This fraction is the power $1 - \beta$ of the test for the given distribution of x, level of significance a, difference d and small sample size n. Usually, the power thus obtained will be smaller than the required one (e.g. 0.80). The procedure is repeated again for the same settings of d and α, but for a substantially larger sample size n. The resulting power may turn out to be larger than required. If not, the procedure is repeated once again until a power of at least equal to the prescribed one is obtained. Finally, the power $1 - \beta$ is plotted against the sample size n. From this plot one can determine the minimal sample size by means of interpolation for a given power $1 - \beta$, significance α, difference d and distribution of x.

A drawback of the Monte Carlo approach is that the number of simulation runs that must be performed can be excessively high. As a general rule one can state that the error d between a theoretical and MC-estimated distribution decreases with the square root of the number of runs N [21]. In order to obtain an accuracy δ with a certainty of 99% one must have at least:

$$N = (1.63 / \delta)^2 \tag{12.12}$$

For example, in the case of 3-digit accuracy ($\delta = 0.0005$) the minimal requirement is about 10^7 runs.

12.5.2 Probabilistic MC for physical systems

Generally in statistics, phenomena with a random component are studied by means of a model. Sometimes, however, the model is too complex to be solved either analytically or numerically. With Monte Carlo methods, the model itself is studied by means of simulated random events.

The latter was the case with the design of the first atomic bomb and nuclear reactor. The physical model which accounted for production, scattering and absorption of secondary neutrons produced by fission of uranium was intractable by ordinary mathematical methods. Hence, no reliable estimates for the design parameters could be obtained, which either guaranteed rapid explosion or controlled operation.

The Monte Carlo approach consisted in modelling the chains of random events that could take place. These are represented schematically in Fig. 12.8 which shows two types of events following fission of an U_{235} atom in the core of a reactor.

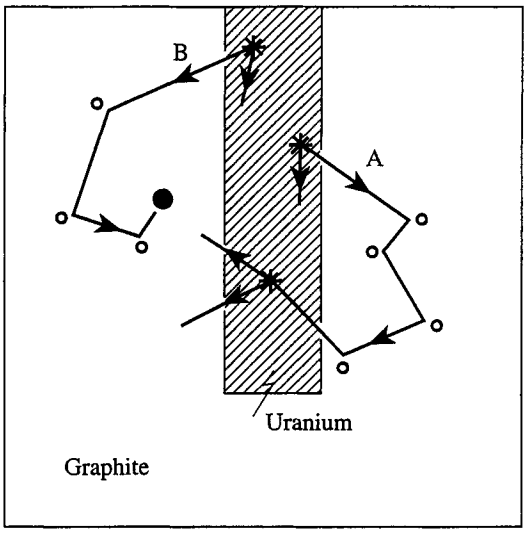


Fig. 12.8. Random walks of secondary neutrons produced by fission of uranium inside the core of a nuclear reactor. The mean free path length of the neutrons can be estimated by means of a probabilistic Monte Carlo simulation of the physical processes.

In event A, a neutron is produced by fission of U_{235} and is scattered outside the fuel rod into the surrounding graphite mantle. The neutron is slowed down by successive collisions and eventually reenters the fuel rod as a thermal neutron where it contributes to the chain reaction by fission of another U_{235} atom. In event B, a secondary neutron is also scattered into the graphite moderator, but is absorbed in the graphite and thus lost for a sustained chain reaction. The successive events constitute a *random walk*, which is characterized by the *mean free path length* of the neutrons. This is a critical parameter for the controlled operation of a nuclear reactor. In reality the model is much more complicated than we have described here. The point is, however, that the random processes inside a reactor can be simulated, given the relevant physical characteristics and design parameters, such as the cross section for scattering and absorption of neutrons by graphite and uranium, the proportion of U_{235} to U_{238}, the dimensions of the fuel rods, the spacing between rods, etc. For various settings of the design parameters, one can then obtain statistical estimates for the operating characteristics, such as the yield of secondary neutrons, heat production, etc.

The phenomena of photon scattering and absorption inside a photographic film have been studied likewise by means of MC calculations [24]. Three basic types of events are represented schematically in Fig. 12.9. In event A, a photon is absorbed directly in the photographic emulsion on top of the film. In event B, the photon is scattered into the supporting film and is subsequently lost, while in C it

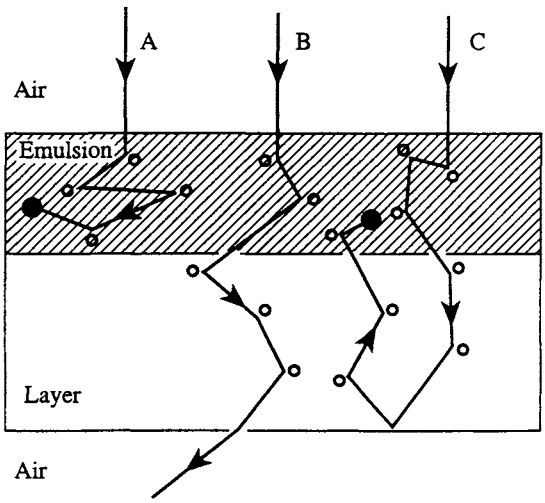


Fig. 12.9. Random walks of photons that interact with the sensitive emulsion and plastic support of a photographic film. The sensitivity of the film can be determined by means of the probabilistic Monte Carlo approach.

is back scattered into the sensitive emulsion layer. In this case, the description of the physical system also led to intractable mathematical equations. The MC approach, however, allowed to design photographic films with controllable sensitivity for various types of emulsion and support.

12.5.3 Deterministic MC

In deterministic applications of MC a deterministic quantity is expressed as a parameter of some random distribution, and then that distribution is simulated. A classical illustration of the deterministic approach is the so-called *needle game of Buffon*, which was designed around 1750 for the determination of the number π [25]. In this game, parallel lines, separated by a distance d are drawn on a sheet of paper (Fig. 12.10). A needle with length l, smaller than d, is thrown nt times on the sheet, and the number of times nc that the needle crosses one of the lines is recorded. The value of π can be determined to any desirable degree of accuracy, depending upon the number of throws nt, by means of the formula:

$$\pi = 2\, l/(d \cdot k) \quad \text{with } k = nc/nt \tag{12.13}$$

Deterministic MC is also used for the calculation of *high-dimensional integrals* for which analytical or numerical solutions are difficult to obtain. The procedure for the one-dimensional case is illustrated in Fig. 12.11. We assume that the function $f(x)$ to be integrated ranges between the values f_{min} and f_{max} on the interval

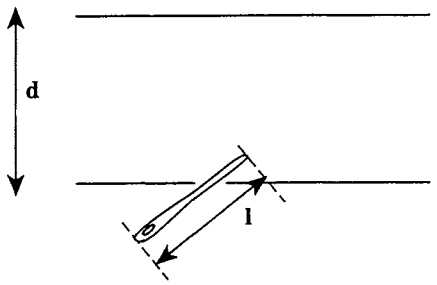


Fig. 12.10. The needle game of Buffon, a classical illustration of the deterministic Monte Carlo method.

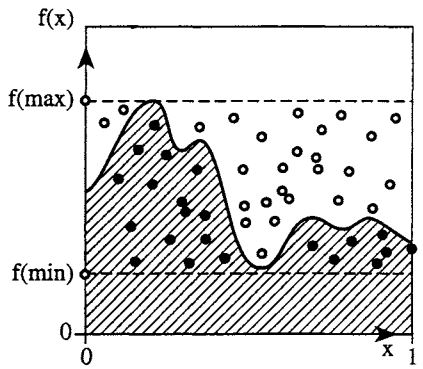


Fig. 12.11. Integration of a function by the deterministic (hit-or-miss) Monte Carlo approach.

of x between 0 and 1. The rectangle defined by the limits f_{min}, f_{max} and 0, 1 is then seeded randomly by nt points and the number of points nb that lie below the curve of $f(x)$ is recorded. The integral is then determined to any degree of accuracy, depending on nt, by means of:

$$\int_0^1 f(x)dx = f_{min} + k(f_{max} - f_{min}) \text{ with } k = nb/nt \tag{12.14}$$

The above approach is also referred to as the *hit-or-miss Monte Carlo* method.

In a broad sense one can also regard the so-called natural computing techniques, such as genetic algorithms and simulated annealing (Chapter 27), as modern developments of the Monte Carlo approach. While the crude Monte Carlo method uses only random sampling of the space of possible solutions, natural computing algorithms make use of both random variation and selection rules in order to arrive at the solution.

References

1. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, Chemometrics: a Textbook. Elsevier, Amsterdam, 1988, p. 51.
2. R. Cleymaet, E. Quartier, D. Slop, D.H. Retief, J. Smeyers-Verbeke and D. Coomans, Model for assessment of lead content in human surface enamel. J. Toxicol. Environ. Health, 32 (1991) 111–127.
3. S. Siegel, Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, Kogakusha, 1956.
4. S. Siegel and N.J. Castellan, Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York, 1988.
5. R.R. Sokal and F.J. Rohlf, Biometry — The Principles and Practice of Statistics in Biological Research. W.H. Freeman, New York, 1981.
6. CETAMA, Statistique appliquée à l'exploitation des mesures. Masson, Paris, 1986.
7. F.R. Hampel, A general qualitative definition of robustness. Ann. Math. Stat., 42 (1971) 1887–1896.
8. H. Theil, A rank-invariant method of linear and polynomial regression analysis (Parts 1–3). Nederlandse Akademie van Wetenschappen Proceedings, Scr. A., 53 (1950) 386–392; 521–525; 1397–1412.
9. P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection. John Wiley, New York, 1987.
10. A.F. Siegel, Robust regression using repeated median. Biometrika, 69 (1982) 242–244
11. D.L. Massart, L. Kaufman, P.J. Rousseeuw and A. Leroy, Least median of squares: a robust method for outlier and model error detection in regression and calibration. Anal. Chim. Acta, 187 (1986) 171–179.
12. Y. Hu, J. Smeyers-Verbeke and D.L. Massart, Outlier detection in calibration. Chemom. Intell. Lab. Syst., 9 (1990) 31–44.
13. F. Mosteller and J.W. Tukey, Data Analysis and Regression. Addison-Wesley, Reading, 1977.
14. M. Meloun, J. Militky and M. Forina, Chemometrics for Analytical Chemistry — Volume 1: PC-aided Statistical Data Analysis. Ellis Horwood, New York, 1992.
15. M. Thompson, Robust statistics and functional relationship estimation for comparing the bias of analytical procedures over extended concentration ranges. Anal. Chem., 61 (1989) 1942–1945.
16. G.R. Phillips and E.R. Eyring, Comparison of conventional and robust regression in analysis of chemical data. Anal. Chem. 55 (1983) 1134–1138.
17. E.S. Edgington, Randomization Tests. Marcel Dekker, New York, 1987.
18. N. Metropolis and S. Ulam, The Monte Carlo method. J. Am. Statist. Assoc., 44 (1949) 335–341.
19. W.F. Bauer, The Monte Carlo method. SIAM J., 6 (1958) 438–451.
20. D.C. Hoaglin and D.F. Andrews, The reporting of computation-based results in statistics. Am. Statist., 29 (1975) 122–126.
21. J.E. Gentle, Monte Carlo Methods, Encyclopedia of Statistical Sciences (S. Kotz and N.L. Johnson, Eds.). Wiley, New York, 1985.
22. Student (W.S. Gosset), The probable error of the mean. Biometrika, 6 (1908) 1–25 and 302–310.
23. P.J. Lewi, R.C.A. Keersmaekers and F. Awouters, Monte Carlo method for the estimation of statistical significance in pharmacological testing. Drug Devel. Res., 8 (1986) 103–108.

24. J. De Kerf, Monte-Carlo methods: random numbers and applications (In Dutch). Het Ingenieursblad (KVIV), 43 (1974) 3–11.
25. M.D. Perlman and M.J. Wichura, Sharpening Buffon's needle. Am. Statist., 29 (1975) 157–163.

Additional recommended reading

K. Danzer, Robuste Statistik in der analytischen Chemie. Fresenius Z. Anal. Chem., 335 (1989) 869–875.
S.C. Rutan, Comparison of robust regression methods based on least-median and adaptive Kalman filtering approaches applied to linear calibration data. Anal. Chim. Acta, 215 (1988) 131–142.
M.O. Moen, K.J. Griffin and A.H. Kalantar, Simple regression and outlier detection using the median method. Anal. Chim. Acta, 277 (1993) 477–487.

Chapter 13

Internal method validation

13.1 Definition and types of method validation

Method validation in analytical chemistry is the last step in method development. Once a candidate method has been obtained, it has to be shown to meet the requirements of the user, namely to measure a specific substance with a given precision, accuracy, detection limit, etc. Method validation is carried out to ensure the quality of a method. It is therefore an essential part of any quality assurance program in the laboratory. *Quality assurance* (in general) has been defined [1] as: "A system of activities whose purpose is to provide to the producer or user of a product or a service the assurance that it meets defined standards of quality with a stated level of confidence."

Chemical analysis can also be considered as a service. To "meet defined standards of quality" requires, among other things, that the analyst should define the performance characteristics that a method must meet and the "stated level of confidence" requires a statistical approach to measuring those performance characteristics. This then leads to the following definition [1]:

> "Method validation consists of documenting the quality of an analytical procedure, by establishing adequate requirements for performance criteria, such as accuracy, precision, detection limit, etc. and by measuring the values of these criteria."

The word "document" is important. Several regulatory bodies require that one details the analytical procedures used as standard operating procedures (SOP). They also require proof that one has indeed carried out a validation of such methods and this means one must document the validation as part of a quality assurance program. Method validation requires different experimental set-ups according to the purpose and the context and must always be concluded with a statistical analysis of the data produced by the method validation experiments. The statistics applied are also very diverse and require knowledge of a large part of the statistical techniques described in the preceding chapters. It is therefore one of the main fields of application of chemometrics.

Two types of method validation can be distinguished. The first will be called *internal method validation*. It consists of the validation steps carried out within one laboratory, for instance, to validate a new method that has been developed in-house or to verify that a method adopted from some other source is applied sufficiently well. In many cases the method validation stops here. When a company is preparing a method for the determination of a drug in blood, it is often not necessary to collaborate with other laboratories in doing this. However, there are many situations in which two or more parties are involved, e.g. the laboratory of a manufacturer and that of a third party. Also, there are many instances where analytical results are of interest to the general scientific community. In such cases agreement about analytical results requires *interlaboratory validation*. Moreover, confronting many laboratories is the best way of thoroughly testing a method. Internal validation is described in this chapter and the interlaboratory approach in Chapter 14.

Primary and secondary performance criteria can be distinguished. The primary criteria are *precision*, which describes the size of random errors, *bias*, *accuracy* and/or *trueness*, which measure the magnitude of systematic errors (see Sections 2.5 and 2.6) and the *detection limit*, which determines the lowest quantity of a substance that can still be distinguished from the background. Secondary criteria are criteria that have an influence on the primary ones. An example is *linearity*. In many cases the determination requires a calibration step and the calibration line is often a straight line. If the method is based on the linearity of the calibration line, then deviation from this postulated relationship will lead to bias. Other secondary criteria are as follows.

- The *range* (i.e. the interval between upper and lower analyte levels) in which the linear relationship or any other calibration relationship used is correct.
- The *quantification limit*, which is the lowest concentration of the analyte that can be determined with sufficient precision and accuracy.
- The *selectivity*, which ensures that the signal measured is not influenced by concomitant substances or, at least, that the contribution of other substances is removed.
- The *sensitivity*, which gives an indication of how much the signal changes with concentration. As discussed in Section 13.8, this term is also used in a very different context together with *specificity* and terms such as *false positive rate* and *false negative rate* to describe the performance of qualitative analysis procedures.
- The *ruggedness*, which measures to what extent a method is sensitive to small changes in procedure or circumstances.

All the terms given above are used in this section in a colloquial sense. Definitions will be given and discussed in the sections devoted to each of the performance criteria. It should be noted immediately that terminology is a major problem. At the time of writing this chapter (1996), there are for instance

nomenclature guidelines by IUPAC [2–4], by ISO [5,6] and a proposed terminology by AOAC [7]. These definitions do not agree on several points. Moreover, many other guidelines, norms, and definitions exist in specific areas. Where possible, we will follow ISO, IUPAC and AOAC guidelines. Where they disagree, we will say so and state our preference for one or other term.

13.2 The golden rules of method validation

There are three very important rules which must always be kept in mind:

– *Validate the whole method*: Quite often, one validates only the actual determination (e.g. the atomic absorption measurement). One must however validate also the preparatory steps, such as dissolution and digestion of the sample. Where relevant, attention must be paid to the sampling and the storage of the sample. These, however, are not part of the validation of the analytical determination as such. In other words, one assumes that the sampling is correct and this assumption is tested separately.

– *Validate over the whole range of concentrations*: A method may work very well at high concentration but be inadequate at low concentration. It is also known that precision depends on concentration (see further Sections 13.4.2 and 14.2.5).

– *Validate over the whole range of matrices*: It is evident that a method for moisture in cheese does not necessarily work for the same determination in chocolate. However, even "cheese" consists of sufficiently different types of matrices to require that one should consider several representative kinds of cheese in the validation. This is also true for "urine" (include several urines from different patients) or "waste waters" (identify the different types of waste water and include a representative set in the validation procedure).

13.3 Types of internal method validation

There are several types of internal laboratory validation:

– *Prospective validation*. This is carried out when a new method is introduced. The method must then be fully tested for its performance characteristics. Prospective validation can often be divided in an exploratory phase and a full validation. In the *exploratory validation* stage one determines with a limited number of samples whether the method can be considered to be a good candidate for its purpose. Very often this initial phase will focus on those aspects of a method which are known to be the more delicate ones (e.g. selectivity of a chromatographic method, freedom from matrix interferences of an atomic absorption method) and a cursory determination of repeatability. When the results are considered acceptable, a more detailed *full validation* follows, the extent of which is determined by the context in which the analysis is carried out (e.g. is the method to be used for a short

period or over many years? In the first case, there is no sense in determining the ruggedness of the method — see Section 13.4.5).

– *Suitability checks.* These can be applied when transferring a method from one laboratory (where it was fully tested) to another. This is then called *transfer suitability check* and requires the receiving laboratory to do a reduced amount of testing. Method bias (see Section 13.5.1) has been eliminated as a source of error in the prospective validation phase, but laboratory bias may exist. This means that one will no longer need to study, e.g., freedom from matrix effects, but the receiving laboratory will need to analyze a few samples which have also been analyzed by the developing laboratory. The receiving laboratory will certainly also need to determine its own repeatability values (see Section 13.4.1).

System suitability checks are used to investigate whether the instruments, reagents, etc. are functioning correctly before starting a new series of determinations. It often consists of checking whether certain key characteristics of the method are respected. This will nearly certainly involve an evaluation of the calibration line (is it still straight? Has the sensitivity changed?) and, where this is relevant, of the blank. Method specific characteristics are also used. For instance, in chromatography one will require that a certain resolution (often ≥ 1.5) is obtained. This type of suitability checking is often included when working under Good Laboratory Practices (GLP) rules or under a quality assurance program and should be specified in the standard operating procedure (SOP).

– *Retrospective validation.* One can collect over a period of time the results of a certain number of determinations. These are then used to determine precision over long periods.

– *Quality control.* Running one or a few samples with known composition, preferably in a blind way, permits the preparation of charts for both the mean result (detection of bias) and the range (repeatability). This was discussed in detail in Chapter 7 and will not be considered further here.

Which type of method validation has to be carried out depends on the application field of the laboratory. Because there are so many different contexts, it is impossible to give an exhaustive enumeration. However, one can distinguish more or less three types of situations, namely:

– The laboratory develops its own methods, to a large extent for its own use. A typical example is a pharmaceutical company that develops and produces its own active molecules and requires analytical methods for content and stability in formulations, to investigate metabolization, etc. The methods are essentially meant for use in the company, but must be validated and be available to regulatory bodies. Such a laboratory will essentially carry out full validation about all the performance characteristics described in Section 13.2. It may also develop suitability checks for transfer to other laboratories of the same group or contract laboratories and will certainly prepare suitability checks for inclusion in SOPs.

– The laboratory develops new methods for general use. An example might be a research institute for the agro-food industry. This laboratory will need to carry out a full validation and prepare suitability checks for the SOP. If the method is successful and thought to be of more general use, the laboratory will take steps to have some official organization, such as the Association of Official Analytical Chemists (AOAC), organize an interlaboratory study of the method-performance type (see Section 14.1).

– The laboratory uses standard methods. Examples here are control laboratories, both governmental and industrial. Since the method has been validated its performance characteristics are known. The laboratory should concentrate on proving that it is generally proficient in its chosen area, for instance by analyzing reference materials, and, where possible, by participating in interlaboratory studies of the lab-performance type (see Section 14.2). When the laboratory carries out routine analyses on a regular basis, it will also need suitability checks for daily use and quality control procedures.

13.4 Precision

13.4.1 Terminology

The precision is a measure for the size of the random errors. Random errors are discussed in Sections 2.2.2 and 2.5. From a statistical point of view, precision measures the dispersion of the results around the mean, irrespective of whether that mean is a correct representation of the true value. Therefore, it requires the measurement of the standard deviation. How this is done depends on the context.

Two extreme types of precision are usually distinguished, namely the *repeatability* and the *reproducibility*. Reproducibility, as defined by ISO [5,6], can be determined only with interlaboratory experiments and for this reason, we define these terms in Section 14.2.1 and recommend that the reader should read that section together with the present section.

In short, repeatability is the precision obtained in the best possible circumstances (same analyst, within one day when possible) and reproducibility in the most adverse possible circumstances (different laboratories, etc.). Intermediate situations may and do occur.

A protocol about collaborative studies prepared under the auspices of IUPAC [8] also considers what it calls preliminary estimates of precision. Among these it defines the *total within-laboratory standard deviation*. It includes both the *within-run* (= repeatability) and the *between-run* variation. This means that one has measured on different days and preferably used different calibration curves. The total within-laboratory standard deviation can be considered as a *within-laboratory*

reproducibility. These estimates are preliminary when the experiments are carried out as a prelude to an interlaboratory method performance study. Other terms, such as intra-assay (= within-run) and inter-assay (= between-run) precision are also used. The ISO-standard [6] also gives some definitions in this context. A laboratory cannot determine reproducibility as such, because this has to be done in interlaboratory experiments, but it can determine *intermediate precision conditions* (i.e. intermediate between reproducibility and repeatability). ISO recognizes what is called *M*-factor different intermediate precision conditions ($M = 1, 2$ or 3), where $M = 1$ means that only one of the three factors (operator, equipment or time) is different, or the equipment is recalibrated between successive determinations. $M = 2$ or 3 means that two or all three factors differ between successive determinations. The term intermediate precision has been accepted for instance by the ICH [9], the International Committee for Harmonization that regulates terminology in pharmaceutical analysis.

A third term used in the context of precision is *robustness* or *ruggedness*. An analytical procedure consists of a set of instructions, such as "Adjust the pH to 5 by adding acetic acid 1 N," or "Heat during 5 minutes at a temperature of 100°C." Small departures from these details often occur when one carries out the procedure in practice and one may wonder how rugged the procedure is to such variations. In the same way, the analyst developing a method is faced with the question of how strictly instructions should be stated. Should a pH of 5 ± 0.05 be required or is 5 ± 0.5 sufficient? The question will also be how rugged the new method is in relation to departures from the nominal values put in. In this case, one needs to measure the robustness or ruggedness of the method. The determination of the ruggedness is sometimes carried out to detect possible critical experimental parameters, that have a larger effect on the results than other parameters. Controlling such parameters may lead to better reproducibility or to avoid sources of laboratory bias (see also Section 13.5.1).

13.4.2 Repeatability

A laboratory can measure its own performance for a certain application in terms of repeatability, or several laboratories together can measure the repeatability of the method by carrying out an interlaboratory experiment. The latter is explained in Section 14.2 and we will confine the discussion here to the former. The basic procedure is simple. Six to eight replicate determinations are carried out, when possible within a single run by the same analyst, and the standard deviation is determined. The result can also be reported as a *relative standard deviation* (sometimes symbolized as RSD) or the *coefficient of variation* (% CV) (see also Section 2.1.4.3). IUPAC [2] prefers the term *percentage standard deviation* instead of coefficient of variation, but recommends that the relative standard deviation be

reported. Some guidelines suggest carrying out the repeatability measurement three times, to pool the variances (see Section 2.1.4.4) and obtain the standard deviation from the pooled variance. Care should be taken that the replicates are true replicates and not only measurement replicates, i.e. it should be ensured that all steps are replicated. If the blank is a possibly important source of variation, then it, too, should be replicated. Because precision often depends on concentration, it should not only be determined at the standard or specification values of concentration but also at the upper and lower limits (see also quantification limit — Section 13.7.3) of the concentration range if this is not very limited. Of course, it may be useful to measure separately the repeatability of a certain step in the procedure. For instance, repeatability of the injection is often measured by chromatographers. This permits the steps responsible for important parts of the total variation to be identified and a decision made as to which step should be better controlled. To make a distinction, certain organizations [10] distinguish between what they call the *precision of a method* and the *precision of a system.* The former requires repetition of the whole procedure, while the latter results from replicate measurements of a standard preparation "in a form ready for direct measurement of the analyte (e.g. no further sample treatment is required)".

There are situations which require more elaborate experimental designs. Consider, for instance, the example of Table 13.1. The left part of this table has already been given as Table 2.3. A method for measuring moisture in cheese was developed. It is not acceptable to validate it for only one type of cheese. This would violate the third rule of method validation (Section 13.2). We then need to select a certain number of cheeses that covers sufficiently well the scope of the method. Let us suppose that this was achieved by selecting the seven first types in Table 2.3 (reprinted as Table 13.1). In that case we can analyze a number of replicates n_i of

TABLE 13.1

Comparison of the repeatabilities of two methods for moisture in cheese

Type of cheese	Karl Fischer		Oven	
	s_i	n_i	s_i	n_i
Processed cheese food	0.29	10	0.01	2
Processed cheese food	0.31	10	0.01	2
Monterey jack	0.35	8	0.12	2
Cheddar	0.24	8	0.13	2
Processed american	0.30	8	0.13	2
Swiss	0.31	8	0.25	2
Mozzarella	0.24	9	0.01	2
s_r	0.293		0.126	
df		54		7

each type as described in the table. It is not always necessary that as many replicates of each sample are analyzed as for the Karl Fischer method. In fact, we will see that in certain applications duplication can be sufficient.

When the replicates of one cheese are analyzed within the shortest possible time and, when possible within one run, a measure of repeatability is obtained. It does not matter that there is a time lapse between the analysis of each different cheese. Indeed, the calculation procedure is such that one determines precision under repeatability conditions for each separate cheese and then pools them to obtain an average measure of repeatability. How to do this was shown in Section 2.1.4.4. We conclude that the repeatability standard deviation, $s_r = 0.293$. When the number of replicates is only two then we have paired data and eq. (2.8) can be applied.

A question which can be asked in method development is whether a certain method is more precise than another. Let us consider an example. We want to compare the Karl Fischer method, which was used for the data of Table 2.1 with another method, namely an oven method. This leads to the data of Table 13.1.

In Chapter 5 we learned that two standard deviations can be compared using the F-test. We make use of this here for the pooled standard deviations. The calculated F is given by $F = (0.293)^2/(0.126)^2 = 5.41$.

For H_0: σ_r (Fischer) = σ_r (oven) and H_1: σ_r (Fischer) $\neq \sigma_r$ (oven), i.e. a two-sided test, and $\alpha = 0.05$, $F_{crit} = F_{0.05;54,7} = 4.27$

Since $F > F_{crit}$ we reject H_0 or, in other words, the Fischer method has a different repeatability from that of the oven and, in view of the results obtained, we conclude that the oven shows better repeatability.

Note that it is not possible to show that the Karl Fischer and the oven method have a different repeatability for the first processed cheese food (or any other cheese) specifically. Indeed, for the first type of cheese $F = (0.29)^2/(0.01)^2 = 841$ and $F_{crit} = F_{0.05;9,1} = 963$. This is due to the small number of replications (and therefore of degrees of freedom) for the oven method. We should remember that the β-error (not finding a difference, when that difference is real) depends on the number of replicates (see Chapters 4 and 5). An ISO norm [11] gives graphs that allow to determine β at a given α and number of replicates n or the n needed to reach a given β at a certain level of α, both for the comparison of an experimental s with a given σ or the comparison of two standard deviations. Unfortunately, for the comparison of two standard deviations the graphs are given only for situations where the number of replicates is the same for both standard deviations. It will be clear that the larger n is and therefore the degrees of freedom, the smaller β will be. Pooling variances as we did here, is useful in such cases, because the number of degrees of freedom increases. For cheese 1 we are not able to decide whether the Karl Fischer method is worse than the oven method, but we can decide that this is so for cheese samples on the whole. Part 6 of ISO norm 5725 [6] gives numbers of

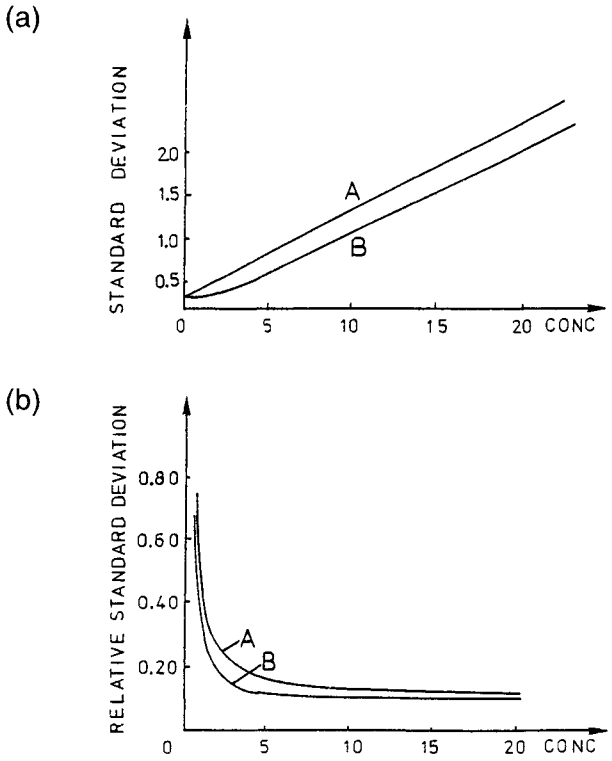


Fig. 13.1. Standard deviation (a) and relative standard deviation (b) as a function of concentration. Models derived for ICP [12]; A: $\sigma = 1/3 + 0.1\ C$, B: $\sigma^2 = 1/3 + 0.12\ C^2$.

measurements required to detect a difference between precisions as a function of what is called the *detectable ratio*. This is defined as the minimum ratio of precision measures that the experimenter wishes to detect with high probability from the results of experiments using two methods.

A final question that can be asked is how the repeatability changes with concentration. The data of Table 13.6 are typical. They show that the repeatability standard deviation increases with the concentration. The coefficient of variation usually decreases, but may stabilize for higher concentrations. When the concentrations cover a large range, this phenomenon of heteroscedasticity is often noted. Several studies have evaluated repeatability as a function of concentration. For instance, Thompson [12] studied repeatability in 700 geochemical materials for 25 elements by ICP. He investigated several models, the most adequate of which was also the simplest, namely $\sigma = \sigma_0 + b_1\ C$. Good results were also obtained with the relationship $\sigma^2 = \sigma_0^2 + b'^2_1\ C^2$, where C is concentration. The authors consider this to be theoretically more satisfactory because variances are additive and standard deviations are not. The relationships are shown in Fig. 13.1. It should be noted that

these relationships are studied also in the context of interlaboratory studies (see Section 14.2.5).

In Chapter 8 it was explained that when the variance is not constant over the range studied weighted regression may be preferable when constructing calibration lines, because it will lead to results with better precision. It follows that in many cases weighted regression should indeed by preferred.

13.4.3 An intermediate precision measure: within-laboratory reproducibility

In this section we will discuss the determination of (time-different) intermediate precision or within-laboratory reproducibility as defined in Section 13.4.1.

The basic experimental set-up is again straightforward. One measures the standard deviation of a set of replicate measurements, for instance by analyzing one replicate each day for a certain number of days (often $n = 5$ to $n = 8$). More complex set-ups are possible. For instance, one can estimate within-laboratory reproducibility and repeatability in a single experimental set-up with duplicates. Consider, for example, Tables 13.2 and 13.3. Two measurements are carried out on 7 days. The ANOVA table learns that the residual mean squares, which here are termed the mean squares within-days, is 0.0414. This represents s_r^2 under repeatability conditions ($s_r = 0.20$). To determine the variance due to the between-day effect, one uses eq. (6.20) and substitutes $s^2_{between}$ and $MS_{between}$ for s_A^2 and MS_A respectively

TABLE 13.2

Experiment for the determination of within-laboratory reproducibility and repeatability from a single experimental set-up

Day	Replicate 1	Replicate 2
1	31.2	31.7
2	30.9	30.9
3	30.7	30.9
4	31.1	31.5
5	31.3	31.6
6	31.4	31.6
7	31.4	31.4

TABLE 13.3

ANOVA Table for the data of Table 13.2

Source	SS	df	MS
Between days	0.9843	6	0.1640
Within days	0.2900	7	0.0414
Total	1.2743	13	

$$s^2_{between} = (MS_{between} - MS_{within}) / n_j \quad (13.1)$$

For $MS_{between} = 0.1640$, $MS_{within} = 0.0414$ and $n_j = 2$, this yields $s^2_{between} = 0.0613$.

The within-laboratory reproducibility s^2_{WR} is equal to:

$$s^2_{WR} = s^2_r + s^2_{between} = 0.103 \quad (13.2)$$

$$s_{WR} = 0.32$$

The conclusion is that the repeatability standard deviation is 0.20 and the within-laboratory reproducibility standard deviation is 0.32.

The number of replicates and days depends on the situation. Some guidelines give minimal requirements. For instance, the Société Française des Sciences et Techniques Pharmaceutiques (SFSTP) [13] requires 6 replicates and 3 days or laboratories or operators or instruments, depending on the type of intermediate precision one needs to measure. The National Committee for Clinical Standards (NCCLS) [14] recommends 20 days and determinations in duplicate. In all cases, one should remember that the repeatability and within-laboratory reproducibility are estimates of the true values of these parameters and that the estimate becomes better when n increases. The NCCLS procedure is more equilibrated than the SFSTP one. In the latter there are 15 df for the repeatability compared to only 2 for the between-day component. In the former there are 20 df for the repeatability and nearly as many (19) for the between-day component.

13.4.4 Requirements for precision measurements

The precision required depends on the application. However, we can ask what precision should reasonably be expected. Important work in this context has been done by Horwitz [15] in the context of interlaboratory studies (see Section 14.2). Some guidelines have been proposed by several organizations in specific areas. For instance, in the area of pharmacokinetics, a committee [16] proposed that precision is acceptable if it is smaller than 15% relative standard deviation, as measured with $n \geq 5$ replicates, except at the quantification limit (see Section 13.7.3) where it should not exceed 20%. Strangely, these values are given both for repeatability and within-laboratory reproducibility. As the latter is usually worse than the former, one can infer that the criteria given above are meant for within-laboratory reproducibility. The Canadian Acceptable Methods guidelines [10] expect a method intra-day and inter-day relative precision of 1% for drug substances and less than 2% for active substances in dosage forms. For minor components (impurities/related substances) less than 5% system relative precision is expected at the 0.2% concentration level. In all these guidelines it would seem that one does not distinguish between true precision and precision, measured with the recommended number of replicates (5 to 8), which is only an estimate of the true precision. If one

finds that the estimated precision measure is 4.8% and 5% is the accepted limit, should one then accept the method because $4.8 < 5$ or should one reject it, because the upper confidence limit around 4.8 exceeds 5? It would seem that the first position is adapted, but this is not made clear. This is certainly a weakness of such guidelines.

In Section 2.4 we studied the quality of a measurement in relation to the quality of a process. Relating the capability of a process and the tolerance limits of a process to the acceptable precision for a measurement method is possible, but seems to have been performed only rarely in the analytical literature.

A very interesting new development is the use of precision clauses based on repeatability or reproducibility standard deviations. For the moment, this is only advocated for standard methods resulting from interlaboratory experiments. However, there is no reason why they should not be included more generally in suitability checks, SOPs, QC programs, etc. An example of such a clause is: "The absolute difference between two single test results obtained under repeatability conditions should not be greater than 0.5 mg/kg". This is described in Chapter 14.2.

13.4.5 Ruggedness

There are no definitions of ruggedness by the more general authorities such as ISO or IUPAC, but there are some in the pharmaceutical world, such as in the US Pharmacopeia [17], the Canadian Acceptable Methods [10] and the SFSTP document [13]. In the chemical literature the term ruggedness or robustness is used when one measures the influence of small changes in the stated procedure on the result. If the change induced is considered to be acceptably low, then the procedure is considered to be rugged. The French definition comes close to this. It states that "the ruggedness of an analysis procedure is its capacity to yield exact results in the presence of small changes of experimental conditions such as might occur during the utilization of these procedures." It continues by defining that by small changes in experimental conditions is meant "any deviation of a parameter of the procedure compared to its nominal value as described in the method of analysis."

The US Pharmacopeia, on the other hand, defines ruggedness as follows: "The ruggedness of an analytical method is the degree of reproducibility of test results obtained by the analysis of the same samples under a variety of normal test conditions, such as different laboratories, different analysts, different instruments, different lots of reagents, different elapsed assay times, different assay temperatures, different days, etc.". In short, this is a definition of reproducibility. It should be noted that the definitions of the US Pharmacopeia often do not follow general usage in method validation. For instance, the term repeatability is not known by them. The Canadian document [10] follows the US Pharmacopeia, but includes a paragraph hinting at the French definition by including different levels of ruggedness

testing. One level "requires verification of the basic insensitivity of the method to minor changes in environmental and operational conditions", while another level is very similar to the US Pharmacopeia definition. We consider the French definition as the most apt. As a consequence, a ruggedness test as we understand it here consists of a set of experiments according to an experimental design to study how an analytical method is affected by small changes in the implicit or explicit procedural details. By explicit, we mean a factor mentioned in the procedure, for instance the time during which one has to boil a solution or the molarity of the hydrochloric acid to be added. The implicit factors are not mentioned as such but may have an influence. For instance, if no temperature is mentioned at which certain steps in the procedure have to be carried out, then one will work at ambient temperature. We may then wonder whether carrying out these steps at 15°C or at 25°C will have an effect on the end result.

The term "ruggedness" was introduced by Youden and Steiner [18] into analytical chemistry. They recommend that for each factor one defines a nominal and an extreme level. The nominal level is the level given in the procedure or the most probable level of an implicit factor, the extreme level is the one which exceptionally might be attained in practice. Usually, one exaggerates a little in defining the extreme level to make sure that one measures the maximum effect possible. For instance, if a procedure states: "Boil the solution during 10 minutes", then one could reason that it is unlikely that anyone would boil it for longer than 15 minutes. The nominal level would be 10 minutes, the extreme level 15 minutes. One can also consider two extreme levels around the nominal level. For instance, the nominal level for boiling a solution being 10 minutes, one could consider that the extremes are 7 minutes and 15 minutes and try to determine the effect on a response between those two levels.

As there are two levels of each variable and one does not want to perform too many experiments, the experimental design used is often one of the screening designs, described in Chapter 23, i.e. either a saturated fractional factorial or a Plackett Burman design. Different articles concerning the measurement of ruggedness using designs of this type were published by Vander Heyden et al. [19] and van Leeuwen et al. [20].

It is not possible to go into the details of exactly how these designs are applied and interpreted yet but a few examples should give an idea. Table 13.4 is an example of the simplest possible application namely a design consisting of four experiments to examine the ruggedness of a procedure towards three factors. Let us suppose that we have developed a colorimetric procedure and are concerned about the effects of the factors pH (A), temperature (B) and concentration (C) of a reagent on the absorbance. The nominal values are pH = 8.0, concentration = 0.10 M and the temperature is not specified. We could decide to investigate the experimental region from pH 7.8 to 8.2, concentration from 0.09 M to 0.11 M and

TABLE 13.4

Ruggedness determination for three variables: (a) actual values; (b) coded values; (c) computation of effects

(a)	Exp.	pH	$t°$	Conc.	Result
	1	8.2	25	0.11	1.00
	2	7.8	25	0.09	0.90
	3	8.2	18	0.09	1.01
	4	7.8	18	0.11	0.89
(b)	Exp.	A	B	C	y
	1	+	+	+	1.00
	2	–	+	–	0.90
	3	+	–	–	1.01
	4	–	–	+	0.89
(c)	Effect A = $[(y_1 + y_3) - (y_2 + y_4)]/2 = 0.11$				
	Effect B = $[(y_1 + y_2) - (y_3 + y_4)]/2 = 0.0$				
	Effect C = $[(y_1 + y_4) - (y_2 + y_3)]/2 = -0.01$				

temperature from 18°C to 25°C. We will call the lowest value the – level and the higher one the + level. Referring to Table 13.4, this means that one should carry out the first experiment at pH +, i.e. 8.2, concentration +, i.e. 0.11 M, and temperature +, i.e. 25°C.

Note that for each factor there are two experiments at the + and two at the – level. For instance, for pH the experiments 1 and 3 are at the +, 2 and 4 at the – level. One reasons that by subtracting the sum of the two – experiments from the two + experiments and dividing by 2, one estimates the effect of that factor. Thus, one obtains the estimates given in Table 13.4. The effect of A (in absolute values) is higher than that of B and C. How to treat such data is described in more detail in Chapter 23. However, it is clear that, if the standard deviation on the four experiments is not appreciably higher than that obtained for the repeatability, one may conclude that the method is rugged. Also, if one variable needs to be better controlled, it is variable A.

More complex designs are sometimes required. One such design is the so-called reflected design. An example is given in Table 13.5. This is applied when one considers that effects may be asymmetric, for example a higher pH than the nominal one may have an effect but a lower one not. Two designs are then made, one with the upper level (1) and the nominal one (0) and one with the lower level (– 1) and the nominal one. The experiments 1–12 make up the first design and experiments 12–23 the second. It should be noted that experiment 12 is common to both designs. The statistical interpretation of the results will be discussed further in Chapter 23.

TABLE 13.5

Reflected design for 11 factors (F1–F11) (from [20])

Exp.	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
1	1	0	1	0	0	0	1	1	1	0	1
2	1	1	0	1	0	0	0	1	1	1	0
3	0	1	1	0	1	0	0	0	1	1	1
4	1	0	1	1	0	1	0	0	0	1	1
5	1	1	0	1	1	0	1	0	0	0	1
6	1	1	1	0	1	1	0	1	0	0	0
7	0	1	1	1	0	1	1	0	1	0	0
8	0	0	1	1	1	0	1	1	0	1	0
9	0	0	0	1	1	1	0	1	1	0	1
10	1	0	0	0	1	1	1	0	1	1	0
11	0	1	0	0	0	1	1	1	0	1	1
12	0	0	0	0	0	0	0	0	0	0	0
13	−1	0	−1	0	0	0	−1	−1	−1	0	−1
14	−1	−1	0	−1	0	0	0	−1	−1	−1	0
15	0	−1	−1	0	−1	0	0	0	−1	−1	−1
16	−1	0	−1	−1	0	−1	0	0	0	−1	−1
17	−1	−1	0	−1	−1	0	−1	0	0	0	−1
18	−1	−1	−1	0	−1	−1	0	−1	0	0	0
19	0	−1	−1	−1	0	−1	−1	0	−1	0	0
20	0	0	−1	−1	−1	0	−1	−1	0	−1	0
21	0	0	0	−1	−1	−1	0	−1	−1	0	−1
22	−1	0	0	0	−1	−1	−1	0	−1	−1	0
23	0	−1	0	0	0	−1	−1	−1	0	−1	−1

When one interprets the ruggedness as proposed by the US Pharmacopeia [17] (see above) and would like to quantify the effects of, for instance, different laboratories and different instruments, it is not possible to apply designs such as those of Tables 13.4 and 13.5. Supposing that one lab is situated in the US, the other in Japan, this would require the Japanese instrument to be moved to the US and the US instrument to Japan to carry out experiments for combinations of the variables "country" and "instrument" as required in a factorial design. In such cases, one would prefer to carry out nested designs (see Chapter 6).

13.5 Accuracy and bias

13.5.1 Definitions

Systematic errors are characterized by terms such as *trueness* and *bias* and related to the term *accuracy*. Unfortunately, there is quite a lot of confusion about

them, because the definitions by different organizations are sometimes contradictory. ISO [5,6] defines accuracy as "the closeness of agreement between test result and the accepted reference value" and adds as a note that the term accuracy describes a combination of random components and a common systematic error or bias component. A test result can be a single result or the average of a set of results. IUPAC [2] and AOAC [7] give definitions that are very similar. Probably the AOAC definition is clearest. It states that the accuracy is the difference of individual values from the "true" or "assigned" or "accepted" value.

ISO [6] defines the trueness as "the closeness of agreement between the average result obtained from a large series of test results and the accepted reference value". The definition adds that the measure of trueness is expressed in bias. In other words, trueness is the concept and bias is the measure. Bias itself is defined as "the difference between the expectation of the test results and an accepted reference value". In practical terms, this means that to ISO bias and trueness essentially mean the same thing. It should be noted that IUPAC [2] gives the same meaning to bias, but does not recognize the term trueness. AOAC also accepts bias in the same sense. It states that bias is the "long term" or expected difference from an average of many groups of individual values from the "true" or "assigned" or "accepted" value. AOAC defines trueness on the contrary as the difference of an average for a group of individual values from the "true" or "assigned" or "accepted" value. It thereby creates a hierarchy such that accuracy is the difference of an individual result from the true value, the trueness that of a single average and the bias that of many averages.

Although the wording of the definitions is different one should note that all three organizations seem to agree about the terms accuracy and bias. As the terminological situation stands now, it therefore seems reasonable to avoid the term trueness and use only the others.

It is only recently that the term accuracy was accepted by the chemical community as having the meaning given in the above definition. Indeed, as ISO writes in its introduction "accuracy was at one time used to cover only the component now named trueness". It was an ill-advised move of ISO not to have kept the term accuracy as it was used originally and introduced trueness, because then all organizations would have agreed without difficulty. Indeed, in the 1990 draft to its present document IUPAC still defined accuracy (of the mean) as "The closeness of agreement between the true value and the limiting or population mean result which would be approached by applying the experimental procedure a very large number of times". The ICH [9], for instance, still states that the "accuracy of an analytical procedure expresses the closeness of agreement between the value which is accepted either as a conventional true value or an accepted reference value and the value found" and add as an afterthought. "This is sometimes termed trueness". The reader should be warned therefore that in many textbooks and

documents the terms discussed above will be defined or used differently and that, moreover, it is probable that further changes will occur in the terminology.

Let us now try to clarify the situation with a simple example. Suppose the true value, μ_0, is known to be 100. For a single measurement yielding 92 one would then say that the accuracy is –8. If that measurement were to be replicated a number of times, say 5 to 8 times as would be the case for a repeatability measurement and yield an average of 89, then in AOAC terminology the trueness would be –11, ISO would still call this accuracy and IUPAC does not seem to have a specific term for this situation. Probably it would be best to call this an estimated bias or, in analogy, with precision measurements, an intermediate estimate of bias. If many sets of averages are obtained for instance by several labs, and this would yield 90, then IUPAC, AOAC and ISO would say that the bias is (estimated to be) –10 and ISO would consider this bias a measure of the trueness of the method. In this book, we will use bias to describe both the intermediate case and that for which the three organizations use that term. Bias, Δ, is then determined as

$$\Delta = \mu - \mu_0 \tag{13.3}$$

where μ is the population mean of the experimental results and μ_0 the true value. Since μ is not known, but estimated from an observed mean $\bar{x}$, it would in fact be better to define an estimated bias $D = \bar{x} - \mu_0$ (IUPAC uses $\hat{\Delta}$ or $\tilde{\Delta}$ [3]).

There are two components of bias. The first is *method bias*, the error inherent to the method, the second is *laboratory bias*. The latter is often viewed as the bias introduced by the way a specific laboratory applies an otherwise unbiased method. In certain definitions, it is however considered to be the total bias in a given laboratory. IUPAC [14] states: "The laboratory bias should be defined as the difference of the long-term average value from the true, formulated, or assigned value. The average of all individual laboratory biases is the estimate of the method bias". This definition was made in the context of inter-laboratory comparisons and a wider definition would be useful. ISO [6] states that the laboratory bias is the difference between the expectation of results (i.e., the mean of a sufficiently large number of results) from a particular laboratory and the accepted reference value. The *bias of a measurement method* is defined as the difference between the expectation of test results obtained from all laboratories using that method and the accepted reference value. The *laboratory component of bias* is the difference between the average of a large number of results in that laboratory and the overall average result for the measurement method obtained by all laboratories. According to these definitions, the laboratory bias thus is the (algebraic) sum of the bias of the measurement method and the laboratory component of the bias.

Depending on the situation the laboratory component of the bias can be considered to be part of the systematic or of the random error (Fig. 13.2). From the point of view of the individual laboratory this component of bias is a systematic error.

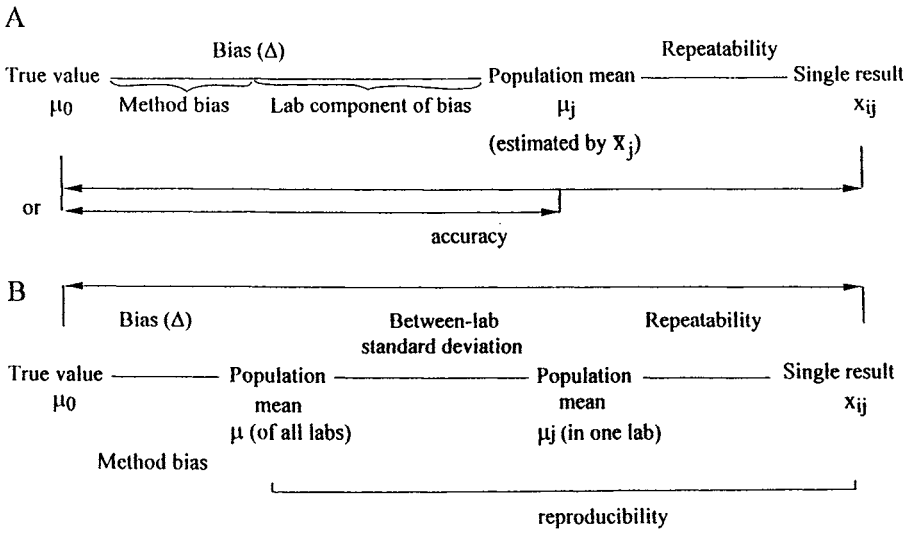


Fig. 13.2. Bias and precision components for (A) a single laboratory *j* working under repeatability conditions; (B) the interlaboratory situation.

However, when carrying out method-performance interlaboratory studies, we can consider that the between-laboratory component of reproducibility includes the laboratory components of the bias of the participating laboratories. When using a standard method, which has been found free from method bias, any bias detected by a laboratory is a lab bias and must be considered to be due exclusively to the laboratory component of bias. To avoid ambiguities concerning the systematic or random nature of errors, ISO [5] also uses the term *uncertainty*. This is defined as "an estimate attached to a test result which characterizes the range of values within which the true value is asserted to lie." This range is determined by several error components, both random and systematic.

Systematic errors may be *constant (absolute)* or *proportional (relative)*. Suppose that the true result of three samples is respectively 100, 200 and 300, that no random error is made and that one finds 110, 210 and 310. This is a constant error. If we were to find 110, 220 and 330, we would call it proportional. A constant error refers to a systematic error independent of the true concentration of the analyte and should be expressed in concentration units. A proportional error depends on the concentration of the analyte and should be expressed in relative units, such as percentage.

The main sources of constant error are insufficient selectivity, which is caused by another component that also yields a response, and inadequate blank corrections. Proportional errors are caused by errors in the calibration, for instance by different slopes of the calibration lines of the standards and in the sample (matrix interference). The incorrect assumption of linearity over the range of analysis will also cause errors related to the concentration to be determined.

13.5.2 Restricted concentration range — reconstitution of sample possible

The experimental design to estimate bias depends on a number of experimental considerations, mainly the extent of the concentration range to be investigated, the availability of blank material (i.e. the matrix to be analyzed without the analyte to be determined) and the possibility to add analyte to such a material in a representative way. We will discuss different typical situations going from the simplest (in this section) to the most difficult (the following sections) from a practical point of view.

In some instances the expected concentration is known within a rather narrow margin, for instance when one needs to verify the content of a manufactured product, such as a drug in a pharmaceutical preparation. When this is technically possible, one validates the method by reconstituting the sample e.g. by preparing synthetic preparations (i.e. preparations with exactly the same composition as the one that has to be assayed) or by spiking (addition of the analyte to blank material). Since one needs to validate over the whole range of concentrations that can be encountered, one prepares a synthetic preparation with a content of 100% of the expected value and contents that are considered to be the extreme limits that can occur in practice, often 80% and 120%.

The validation consists in carrying out the analysis on a number of replicates at each level, often $n = 6$ and comparing the mean obtained with the known content of the placebo. The statistical analysis is therefore carried out with a *t*-test. Let us return to our example from Chapter 4. Suppose that a synthetic preparation containing 100 mg of a substance was prepared. The results obtained are 98.9, 100.3, 99.7, 99.0, 100.6, 98.6 and $\bar{x} = 99.5$.

H_0: $\mu = 100$

H_1: $\mu \neq 100$

Although we observe that $\bar{x} < 100$, the statistical question is two-sided (μ is different from 100) and not one-sided ($\mu < 100$), since no *a priori* reason was given why, in the case of $\mu \neq 100$, it should necessarily be smaller. In method validation, the test is always carried out at the $\alpha = 5\%$ level. Since $n < 30$, the *t*-test is employed. In this case, $s = 0.813$, so that

$$|t| = \frac{|99.5 - 100|}{0.813/\sqrt{6}} = 1.51$$

$t_{0.025,5} = 2.57$

Since $1.51 < 2.57$ the result of the hypothesis test is negative: no significant difference between μ, estimated as 99.5, and 100 can be shown. Let us now suppose that, in fact, the 0.5 mg difference obtained is real and let us also suppose that this difference in the context of the analyzing laboratory is important, so that we would

really want to have detected the difference. Why then was it not detected? The reason is that for the value of the parameter λ (= difference/standard deviation — see Section 4.8) = 0.5/0.813 = 0.61 the number of replicates, $n = 6$, is not large enough to achieve a reasonable β-error, i.e. the probability of not detecting a difference of 0.5 when that difference is real. The ISO norm [11] shows that, in fact, one needs $n = 36$ for $\beta = 0.05$ and $n = 30$ for $\beta = 0.10$! Unfortunately this type of reasoning is seldom applied in method validation, although the definition of Section 13.1 really requires it: one should state how much bias is admitted, and, taking into account the experimentally determined repeatability, compute how many replicates are required to rule out that the bias is larger with a stated probability of making an error. Part 4 of the recent ISO-norm [6] gives a complete description of how to determine laboratory bias by analyzing a reference material with known concentration for the special case that the method applied is a standard method. This norm includes an equation for determining how large n should be.

We should note here that there is a statistical problem. Indeed, the same t-test at the $\alpha = 5\%$ level is carried out three times (at 80, 100 and 120% of the nominal level). Therefore, α for the whole experiment approaches 15% (see Section 5.2). One might take two attitudes:

- the joint level of confidence for the three tests should be $\alpha = 5\%$; therefore each test should really be applied at $\alpha = 5/3 = 1.66\%$
- if something is wrong with the method, then, from a chemical point of view, it is more probable that this occurs at one of the extreme levels. Therefore, experiments at these levels should be considered as separate experiments. In practice, this means that each of them should be judged at the 5% level.

Apparently, the latter approach is always taken. Nevertheless, it would be preferable that the problem be investigated by regulatory agencies and that an explicit decision be taken about which of the two attitudes should be preferred.

13.5.3 Restricted concentration range — reference material available

Reference materials are employed very often. They are of limited value in a full validation of a method, in the sense that they include a certain amount of analyte in a certain matrix and that therefore one will generally not be able to validate with them the whole range of concentrations and matrices required. However, when the range of concentrations and matrices is covered, analyzing reference materials is the validation method to be employed. Moreover, even when the reference materials do not cover the full range, they should be analyzed when available. Obtaining good results on a reference material indicates that the method is at least acceptable for that composition of the matrix and that further full validation over a wider range of compositions has a chance of being successful. A bad result means that further full validation is not useful. Thus analyzing reference materials is often part of the

exploratory validation process (see Section 13.3). Reference materials are also invaluable to detecting the laboratory component of bias and measure the proficiency of individual laboratories in using standard methods (which should be free of method bias).

The statistical analysis of the measurements obtained by a user on a reference material poses a problem. Basically, what must be done is to compare the mean obtained by the user with the mean obtained during the certification process, i.e. the comparison of two means. At first sight this should be done using the *t*-test to compare two means as described in Section 5.1. However, there is a problem, namely that n_2 of eq. (5.6) is not defined. Indeed, the user laboratory compares the mean of the replicate analyses it has carried out (the number of which in this context is called n_u, the n_1 of eq. (5.6)) with the mean of an unknown number of replicate determinations in an equally unknown number of laboratories participating in the certification, so that not all data are available that are required to carry out the comparison of the two means by a *t*-test.

To avoid this difficulty, one should then carry out the computations as described by the certifying organization. For instance the BCR [21] recommends to proceed as follows:

– check that the repeatability of the method is compatible with the repeatabilities of the certifying laboratories. BCR proposes that this be done by verifying that the standard error of the mean of the user laboratory results, $s_u/\sqrt{n_u}$, is less than the standard deviation s of the distribution of certifying laboratory mean values, as stated on the certification document

– if the repeatability standard deviation of the user laboratory is acceptable, then it can verify whether the mean obtained, y_u, falls within the confidence limit of the certified value, which is considered to be $\pm$ 2s and is given by the certifying organization: (certified value – 2s) < y_u < (certified value + 2s) where 2 is the approximate value of t for a sufficiently large number of degrees of freedom or z at the $\alpha = 5\%$ level of confidence.

13.5.4 Large concentration range — blank material available

Since the concentration range is large, one must validate over that whole concentration range. When material to be analyzed without the analyte can be obtained, one can *spike* it. For instance, when one must determine a drug in blood, one can obtain blood without the drug and then add the drug in known concentrations. Spiked samples are also called *fortified* samples [7]. The result is often given as *% recovery* (or recovery rate), i.e the amount found compared to that added expressed as a percentage. The situation is rather similar to that described in Section 13.5.2. However, because the range is large, regression methods can be used: in a narrow range, this would not be recommended (see Section 8.2.4)

because the estimates of the parameters of the regression line would not be optimal. Moreover, the question arises of how many concentration levels to test.

From the chemical point of view, spiking is sometimes less evident than one would think at first sight: for instance, when one analyzes an inorganic species it can be in a different form in the material than in the standard added. This can have a profound influence on the analytical behaviour. In particular, the first steps in some procedures such as dissolution or digestion are difficult to test in this way. We should warn that the chemical problems should not be forgotten by worrying exclusively over the statistics.

One usually adds three to eight different concentration levels, covering the range to be determined. In the same way as for the repeatability and the within-laboratory reproducibility, one should at least determine the recovery at the upper and lower limits of the concentration range, with particular emphasis on any standard or specification value. The Washington consensus document [16] states that one should carry out this type of experiment at at least three concentrations, one near the lower quantification limit (L_Q), one near the centre, and one near the upper boundary of the standard curve.

A first approach is similar to that described in Section 13.5.2 and consists of analyzing enough replicates at each concentration level to be able to carry out a *t*-test. At each level, one compares the mean obtained with the known amount added. At each level, this then is the same situation as in Section 13.5.2. Let us consider an example. The example comes from a study about the analysis of a pharmaceutical drug in urine [22]. The author studied two chromatographic methods A and B. The only difference between the two is that in method A an internal standard is added, while such a standard was not added in method B. The author stated that when a method is under control, addition of an internal standard is not useful and that it will merely increase imprecision, as one adds the variation on the measurement of the internal standard to that of the analyte. The data are given in Table 13.6. We will use the data here to investigate, as an example, whether method B is unbiased. For this purpose we have computed $|t_B| = \frac{|\bar{x}_B - \mu_0|}{s/\sqrt{n}}$ where $\bar{x}_B$ is the mean of the values obtained with method B at a certain level and μ_0 is the concentration level obtained by spiking. This must be compared with $t_{0.025,5} = 2.57$. One concludes that the differences are significant at all levels except at the level 10. There is a (positive) bias in method B. It is small and probably the user would conclude that although the bias exists, it is too small to be of chemical consequence. Indeed, in this area one seems to consider biases up to 15% as acceptable. The bias increases with increasing quantity of drug, so that one would conclude that there is a proportional systematic error and that probably the calibration procedure is not optimal.

TABLE 13.6

Analysis of a drug in urine (adapted from Ref. [22]). t_B is the observed t value when comparing the amount obtained with method B with the known amount.

Amount added (μg/ml)	Method A (μg/ml ± s) 6 replicates	Method B (μg/ml ± s) (6 replicates)	$\lvert t_B \rvert$
0	0.5 ± 0.3	0.5 ± 0.3	4.08
1	1.3 ± 0.2	1.3 ± 0.2	3.67
3	3.6 ± 0.4	3.5 ± 0.4	3.06
10	10.6 ± 0.6	10.3 ± 0.5	1.47
30	34.9 ± 1.5	33.9 ± 1.8	5.31
100	107.2 ± 2.6	104.1 ± 2.9	3.46
300	318.9 ± 8.1	316.1 ± 9.4	4.20
1000	1051.0 ± 29.2	1037 ± 4.4	20.60

The fact that at the level 10 no bias could be shown, will not be considered in this context as an indication that there is no bias, but rather that it could not be detected: a β-error has occurred. The inverse also occurs: one finds no bias at all levels, except one. Suppose that we have carried out tests at the levels 5, 10, 20, 50, 100, 200, 500, 1000 ng/ml and that the test at the 50 ng level shows a difference, while all others do not. What interpretation should be given? It does not make chemical sense to declare that the method is free from bias in the ranges 5–20 ng/ml and 100–500 ng/ml, but is biased at 50 ng/ml. We would probably be tempted to disregard the result at 50 ng/ml, but then we must ask the question whether statistical tests of which we disregard the conclusions should be taken seriously. Carrying out several t-tests in the same validation experiment and interpreting each of them separately, carries with it the philosophical problem we already discussed in Section 13.5.2. For each test separately, one accepts implicitly a possibility of making the decision that there is a difference, while in fact there is not (type I or α error), usually of 5%. In Section 13.5.2 only 3 levels were tested, but it quite often happens that one tests up to 8 levels. Performing eight tests at the 5% level of confidence means that one has about 34% probability that one of the eight tests will lead to a type I error.

If the discrepancy described above occurs at the extremes of the concentration range, one should investigate whether the analytical range was well chosen, for instance by looking at the linearity of the calibration line and as a result probably shorten the range. If the concentration level at which a bias was found is situated in the middle of the range as described higher, one should remember that when several t-tests are carried out in a single experiment (here a method validation experiment), one should really interpret them in a simultaneous fashion and apply, e.g., Bonferroni's method (see Chapter 5).

One might reason that this simultaneous interpretation should include also the extremes of the range. In practice, there is a higher probability of something being wrong at these extreme ranges and one would not like to run the risk to miss that. Bonferroni's method has the disadvantage of making detection of bias at each separate level less sensitive and therefore less adapted to finding bias at the extremes of the analytical range.

Purely from the point of view of detecting bias, the experiment described above is not really optimal. Instead of analyzing 8 levels in 6-replicate, one would have been better inspired to focus on two or three extreme levels (the extremes plus one in the middle). In the latter case, and with the same amount of work, one would have been able to analyze 16 replicates at 3 levels, thereby decreasing the β-error (the possibility of not detecting a bias, when there is one). This is also what we recommend. It is in this case better to concentrate on 3 levels of concentration. However, when one is interested at the same time in determining repeatability in function of concentration, there may be some justification in carrying out the experiment as described.

Instead of summarizing the experiments by using Bonferroni's principle, it is possible to apply regression techniques. One or only a few replicate determinations are then carried out at each level and a graph of the amount found (y) against the amount added (x) is made — see also Fig. 13.3. If no measurement error were made and there were no bias, this would yield the relationship $y = x$, which can be written as:

$$y = 0 + 1\,x \tag{13.4}$$

This ideal situation is depicted in Fig. 13.3a. Because at least random errors are made, one determines the coefficients by regression

$$\hat{y} = b_0 + b_1\,x$$

and one has to show that β_0, estimated by b_0, and β_1, estimated by b_1, are not significantly different from 0 and 1, respectively:

H_0 (intercept): $\beta_0 = 0$; H_1: $\beta_0 \neq 0$

H_0 (slope): $\beta_1 = 1$; H_1: $\beta_1 \neq 1$

Example 5 of Section 8.2.4.1 is an example of method validation using this approach. Let us also apply the same calculations to the data of method B in Table 13.6. Since individual results were not given, we used the mean values to obtain the regression equation. We obtain $b_0 = 1.17 \pm 1.54$ and $b_1 = 1.037 \pm 0.006$, where $\pm$ gives the 95% confidence limits. It follows that $\beta_0 = 0$ and $\beta_1 \neq 1$. If $\beta_1 \neq 1$ (Fig. 13.3b), then the slope of the line differs from what is expected. The difference between the actual line and the ideal one increases with concentration. This is indicative of a proportional systematic error. If it had been found that $\beta_0 \neq 0$, then

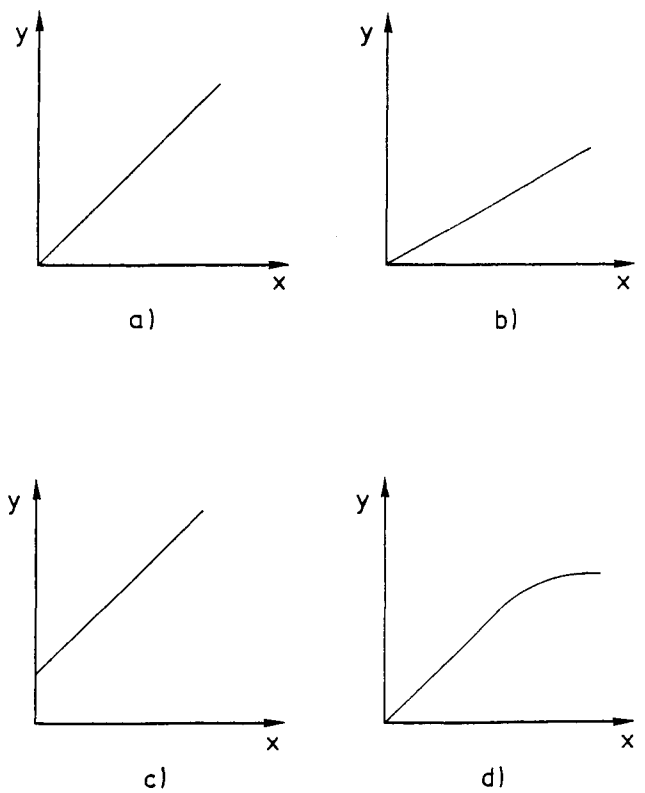


Fig. 13.3. Possible relationships between x (known amount added) and y (amount found); (a) $b_0 = 0$, $b_1 = 1$: no systematic error; (b) $b_0 = 0$, $b_1 \neq 1$: proportional systematic error; (c) $b_0 \neq 0$, $b_1 = 1$: absolute systematic error; (d) non-linearity.

the regression line would be shifted away from what it should be by an amount equal to the intercept b_0 (Fig. 13.3c). This amount does not depend on concentration and one has therefore detected an absolute systematic error.

In the example of Table 13.6, since $\beta_0 = 0$ and $\beta_1 \neq 1$, one concludes that there is a proportional systematic error. Since $b_1 = 1.037$, the best estimate of this error is +3.7%. As described in Section 8.2.4.2, it would be possible, and even recommended, to apply joint confidence intervals. However, in practice this is rarely applied.

Before carrying out the regression analysis, one should first investigate whether the line is indeed a straight line. This can be carried out with the ANOVA procedure described in Section 8.2.2.2. If non-linearity occurs, then this may be due to a problem at one of the extreme levels such as in Fig. 13.3d. One should carry out the validation over a shorter concentration range (if this is still compatible with the original aim!).

An important question is how many levels and how many replicates should be analyzed for a regression experiment. Some guidelines [16] exist and intuitively experimenters seem to favour a relatively large number of levels (up to 8) and a relatively restricted number of replicates. It can be shown that we should test only 3 levels and carry out as many replicates at each level as is considered economically feasible. In fact, in the chapters on experimental design (Chapters 21–25), we will stress that the best estimates of a straight regression line are obtained by concentrating on two (extreme) levels (see also Section 8.2.4.3). The confidence intervals are narrower for, e.g., 6 replicates at 2 levels than for 2 replicates at 6 levels. The third level is added here because it allows us to investigate linearity in the same experiment.

The regression model not only allows us to investigate whether bias occurs, but also when it occurs it helps to diagnose the problem. Indeed, absolute and relative systematic errors are due to different causes and therefore identifying one of these errors points to its source and this is a first step towards finding a remedy.

Requirements for % recovery have been published in a few fields. For instance, the pharmacokinetics consensus document [16] requires a recovery of 85–115% and an EEC guideline for control of residues in food [23] allows the following deviations: ≤ 1 µg/kg: –50% to +20%, > 1 µg/kg to 10 µg/kg: –30% to +10%, > 10 µg/kg: –20% to +10%.

13.5.5 Large concentration range — blank material not available

When no sample can be obtained that does not contain the analyte, the techniques described in the preceding section cannot be applied. This is often the case. For instance, since it is not possible to obtain blood without iron, we cannot apply the techniques of Section 13.5.4 to evaluate the bias in the determination of iron in blood. In this section, we will consider the situation that analyte can be added to the sample in a representative way. As already stated in the preceding section, this is not evident and the analyst should consider carefully whether the chemical composition of the spiking solution ensures that the addition can be considered representative.

As in the preceding section, two approaches are possible. In both one adds m different known amounts and analyzes n replicates of the different concentration levels thus obtained. In the first approach the results at each level are interpreted separately by comparing the known amount added with the difference between the results obtained with and without the addition, and in the second all results are interpreted as one experiment. In the first approach, t-tests are therefore applied, and in the second regression.

Let us first consider interpretation at each level and let us, for simplicity, assume that only one known amount is added ($m = 1$). If the unknown concentration

originally present in the sample is μ_1 and the amount added is such that the concentration increases with the known quantity μ_0, then the total concentration is $\mu_2 = \mu_1 + \mu_0$ and μ_2 too is unknown. One can now carry out the analysis with the method to be validated and obtain the estimates $\bar{x}_1$ and $\bar{x}_2$ of μ_1 and μ_2, respectively. It follows that

$$\bar{x}_2 - \bar{x}_1 = \bar{x} \tag{13.5}$$

Ideally, $\bar{x}$ should be equal to μ_0. As it is obtained from two estimates, $\bar{x}$ itself is an estimate. We call the quantity it estimates μ. No bias is detected if $\mu_0 = \mu$. If the analyses are replicated to a sufficient extent, one can use a t-test as described in Chapter 4 to verify whether

H_0: $\mu = \mu_0$ (H_1: $\mu \neq \mu_0$)

is true. Suppose the sample is analyzed 6 times and the following results are obtained (in mg/ml).

90.00 – 90.80 – 89.70 – 89.20 – 88.60 – 91.20 ($\bar{x}_1 = 89.92$, $s_1 = 0.972$)

and after addition of 60.00 mg/ml

156.00 – 154.20 – 155.30 – 155.60 – 153.80 – 154.70 ($\bar{x}_2 = 154.93$, $s_2 = 0.848$).

The difference between the two series of measurements is $\bar{x} = 65.01$. We should now test whether this differs from 60. We should apply the following t-test procedure. The 95% confidence interval around the difference $\bar{x}_1 - \bar{x}_2$ is given by (Section 5.1.1.2):

$$(154.93 - 89.92) \pm 2.23 \sqrt{[(5 \times 0.972^2 + 5 \times 0.848^2)/10]\left(\frac{1}{6} + \frac{1}{6}\right)} = 65.01 \pm 1.17$$

where $t_{0.025;10} = 2.23$.

Since 60 is not included in the confidence interval, we conclude that 65.01 is significantly different from 60. We could also carry out the test, by subtracting 60 from the second mean (154.93 – 60 = 94.93) and compare the mean 89.92 (s_1 = 0.972) with 94.93 ($s_2 = 0.848$) using the independent t-test as described in Section 5.1.1.2.

What do we validate in this way? Let us suppose we make an absolute systematic error, such as a blank error leading to an overestimation of the concentration by x_b. Since the error is absolute it occurs equally at all concentrations and will affect equally $\bar{x}_2$ and $\bar{x}_1$. When performing the subtraction of eq. (13.5), x_b will be eliminated, so that it will not be revealed by the t-test. Proportional errors on the other hand would be noted. Suppose that there is such an error, so that a result fx ($f \neq 1$) is obtained when one should find x. Then the subtraction of eq. (13.5) leads to

$$f\bar{x}_2 - f\bar{x}_1 = f(\bar{x}_2 - \bar{x}_1) = f\bar{x} \neq \bar{x}$$

and if f is different enough from 1, the difference between μ, which is now estimated by $f\bar{x}$, and μ_0 will be declared significant by the t-test.

The conclusion is that the experiment described here does detect proportional, but *not* absolute, systematic errors. These proportional errors are due to different slopes of the calibration line and the line relating response to concentration in the sample. As a slope is best studied through regression methods, the second approach, which we will now describe, is often preferred. This approach is called the method of *standard additions.* Standard addition was described in Chapter 8. As applied in method validation, standard addition requires the comparison of two lines. The first is the calibration line obtained with aqueous standards, i.e. the calibration line that would be used to analyze unknown samples. The other line is the standard addition line. This is obtained by adding m known amounts to aliquots of the material to be analyzed, often without replication (n = 1), and plotting amount added (x) against signal measured (y). In such a graph, y_0, the value of y measured for $x = 0$ will probably be positive because of the (unknown) amount of substance present at the start of the experiment. It is therefore not possible to do a test on the intercept as in the preceding section, so that one cannot detect absolute systematic errors.

One expects both lines to have the same slope. In Fig. 13.4, line a is a standard addition line. If b were the aqueous calibration line, then one would declare that no bias can be detected; with calibration line c the conclusion would be that there is a proportional systematic error. If b_{1c} is the slope obtained for the calibration line and b_{1s} the slope for the standard addition line, then one tests

H_0: $\beta_{1c} = \beta_{1s}$

How to do this was described in Section 8.2.8. If the slopes are found to differ this means that a relative systematic error is present. This was the case in Example 9 of Section 8.2.8. For an analysis procedure of Al in serum a standard addition experiment was carried out. The slope of the calibration line was found to be 8.63 and that of the

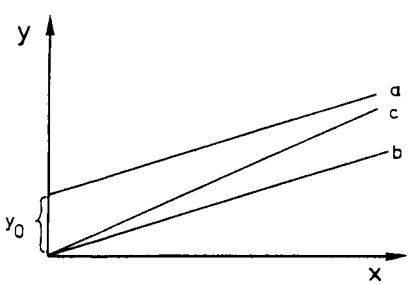


Fig. 13.4. The standard addition method; y = signal measured; x = amount added. Line a = standard addition line, b and c = aqueous calibration lines.

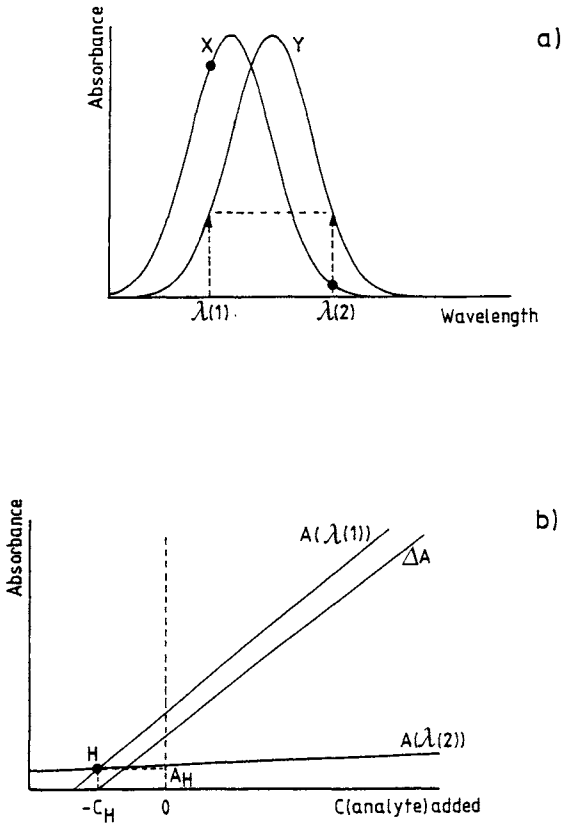


Fig. 13.5. (a) Spectra for analyte X and interferent Y. (b) The H-point.

standard addition line 8.03. The difference between the two slopes was found to be significant. It was concluded that the method is subject to proportional error. The best estimate is that the results are $[1 - (8.03/8.63)] \times 100\% = 7.5\%$ too low.

As the method of standard additions detects only proportional systematic errors other approaches (using reference materials, comparing different methods — see next section) are required to make sure that there is no absolute systematic error. It should be noted that absolute systematic error does not depend on concentration. It is therefore sufficient to show for one or two concentration levels, for instance with a reference material, that there is no bias. Combined with standard addition over a sufficient range of concentration, this validates the absence of bias for the method.

The standard addition method can be extended in certain cases. The *H-point standard addition* [24,25] is a modification of the standard addition method that permits correcting for both absolute and proportional systematic errors. In the simplest case (one interferent, the spectrum of which is known; see Fig. 13.5a) the

method requires measurement of a standard addition line at two wavelengths, λ (1) and λ (2), where the interferent shows the same absorbance. The two standard addition lines intersect at the so-called H-point with coordinates (C_H; A_H), where C_H is the concentration of the analyte in the sample and A_H the analytical signal due to the interferent at both λ_1 and λ_2 (Fig. 13.5b). Modifications of this method can be applied when the spectrum of the interferent is not known.

13.5.6 Comparison of two methods or laboratories

When none of the methods described in the preceding sections can be applied, the last resort is to develop two independent methods and to compare the results. If both methods yield the same results, then both are considered unbiased. It should be noted that both methods should be completely independent, i.e. different. If, for example, method A consists of an extraction, followed by a spectrophotometric determination and method B uses the same extraction and HPLC, then one can validate the spectrophotometric and HPLC steps, but not the extraction. The interpretation can be difficult when the results of both methods are not found to be the same. In that case, one knows that one of the methods, or both, are subject to bias but not more. To know which method is wrong, additional experimentation and chemical reasoning is required.

Comparison of two methods is recommended in the following situations:

– none of the experimental situations described in the preceding sections can be applied to the material to be analyzed. This often occurs. Suppose, for instance, one wants to determine moisture in cheese: it is then not possible to add in a homogeneous and representative way a known amount of water to the cheese.

– one would like to replace an old method, the accuracy of which is considered to be proven, with a new more convenient one. Part 6 of the ISO-standard [6] describes how to do this when one of the two methods is a standard method.

– when it is not sufficient to detect proportional systematic errors with a standard addition experiment, but absolute systematic errors must also be excluded.

A rather similar experimental set-up is encountered when the comparison occurs between two laboratories. Such a comparison is carried out when one lab transfers a method to another lab. This is then called a *transfer suitability check*. The laboratory component of the bias due to the receiving laboratory can be detected by analyzing a set of samples covering the range of application of the method in both the developing and receiving laboratory. Suppose laboratory A has developed and fully validated a method to analyze a certain drug in blood and wants to transfer the method to laboratory B. The best way of studying bias in the receiving laboratory is to analyze the same set of m real blood samples in both laboratories and compare the result according to one of the experimental set ups to be described later. If a difference is found it must be due to lab bias in laboratory B.

The following experimental set-ups are most common.

(a) One analyzes replicates at a restricted set of concentration levels covering the whole concentration range with the two methods. For instance, one might carry out the comparison at the quantification limit, the highest level to be determined and some in-between value.

(b) One analyzes many samples over the whole concentration range. Each sample is analyzed with both methods and few replicate measurements are carried out.

The former method is preferred when a single type of well-defined matrix is analyzed. When many different types of matrix occur, then one prefers the latter method.

Let us first consider the method in which only a few concentration levels are analyzed. At each level several replicate analyses are carried out. Their number is preferably determined using β-error considerations, but is often in practice situated between 5 and 8. As in Section 13.5.4 the number of levels is best restricted to three. The statistical procedure required is an unpaired (independent) *t*-test at each level.

Let us consider one such level. The results obtained are:

$\bar{x}_1 = 32.6 \qquad s_1 = 2.56 \qquad n_1 = 11$

$\bar{x}_2 = 31.6 \qquad s_2 = 2.01 \qquad n_2 = 13$

It should be remembered (Chapter 5) that the independent *t*-test requires that the two series of measurements have the same variance. Therefore, the *F*-test is first carried out.

$F = s_1^2 / s_2^2 = 6.55/4.04 = 1.62$

This is compared with the critical *F*-value for $\alpha = 0.05$ at 10 and 12 degrees of freedom for a two sided-test, $F_{crit} = 3.37$. Since $F < F_{crit}$, $\sigma_1^2 = \sigma_2^2$. Therefore one can pool the variances:

$$s^2 = \frac{10 \times 6.55 + 12 \times 4.04}{22} = 5.18$$

and compute

$$t = (32.6 - 31.6)/\sqrt{5.18(1/11 + 1/13)} = 1.072$$

Because of the low *t*-value we conclude that the two methods are equivalent: there is no bias at the level considered.

When many samples over the whole concentration range are analyzed with both methods, we can apply two types of statistical analysis. A first possibility is to use a paired *t*-test and the second is to compare the results of both methods by regression. In this case preliminary visual analysis of the results is particularly useful and we will first discuss these visual methods.

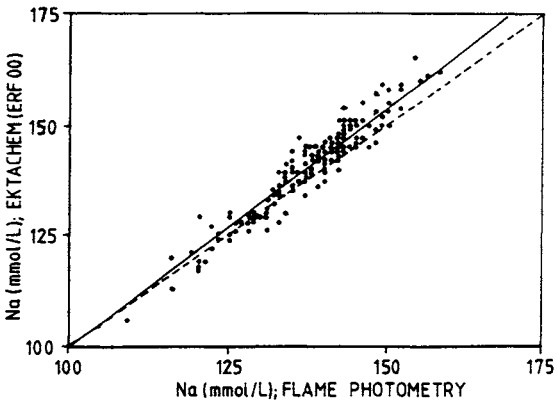


Fig. 13.6. Comparison of two clinical methods. The dashed line is the $y = x$ line, the solid line is the regression line. Adapted from Ref. [26].

In some cases, particularly with automatic measurement techniques, it is possible to analyze many samples (often up to a few hundred). One plots the result for each sample of method 1 against that of method 2. If the two methods yield the same results then the points on the plot should be spread out evenly around the line $y = x$. Including the line for $y = x$ in the plot helps the visual assessment. An example is given in Fig. 13.6. This consists of a comparison of a potentiometric and a flame photometric method [26]. One observes easily that the results in the middle range coincide, but that at higher levels the result of the potentiometric method is higher. There probably is a proportional error (calibration problem) in one of the methods.

Another visual interpretation method will be introduced with a data set from Ref. [27]. This concerns two methods to determine fat in pork products; 19 materials with different fat content were selected. Each was analyzed in duplicate with the two methods. The data are shown in Table 13.7.

One can plot $|w_1|$ and $|w_2|$ against $\bar{x}_1$ and $\bar{x}_2$, respectively (see Fig. 13.7). $|w_1|$ and $|w_2|$ are the absolute differences between the two replicate results (which is equivalent with the range). These plots do not show a trend in the differences between the replicates and therefore indicate that s is constant over the range studied, so that we can expect the standard deviation of the differences between $\bar{x}_1$ and $\bar{x}_2$ to be constant too. This is a necessary condition for carrying out the t-test, because the test assumes that there is one single standard deviation of the differences between the two methods.

One then plots $d = \bar{x}_1 - \bar{x}_2$ in function of $\bar{x} = (\bar{x}_1 + \bar{x}_2)/2$ (see Fig. 13.8). This too does not show a trend, either in magnitude or in range, so that we can safely assume that the conditions are fulfilled for applying the paired t-test. Indeed, we test whether $\bar{d}$ is different from 0. This means that one assumes that all d spring from

TABLE 13.7

Determination of fat in pork products (adapted from Ref. [27])

Product	Method 1		Method 2		
	$\bar{x}_1$	$\|w_1\|$	$\bar{x}_2$	$\|w_2\|$	$\bar{x}_1 - \bar{x}_2$
1	4.83	0.00	4.63	0.00	0.20
2	6.74	0.14	7.39	0.80	–0.65
3	8.46	0.09	8.76	0.24	–0.30
4	12.60	0.33	11.85	1.30	0.75
5	13.52	0.14	13.67	0.49	–0.15
6	15.72	0.62	15.80	1.31	–0.08
7	15.83	0.24	16.05	0.12	–0.22
8	18.37	0.54	18.22	0.74	0.15
9	18.90	0.34	18.79	0.90	0.11
10	23.10	1.03	22.80	0.15	0.30
11	32.45	0.05	32.53	0.39	–0.08
12	40.89	0.64	41.03	0.82	–0.14
13	41.42	0.77	41.52	0.26	–0.10
14	43.36	0.49	43.70	0.16	–0.34
15	45.96	0.59	45.89	0.03	0.07
16	47.70	0.04	47.81	1.16	–0.11
17	50.02	0.07	50.11	0.56	–0.09
18	57.20	0.39	57.51	0.65	–0.31
19	78.48	0.40	79.38	0.09	–0.90

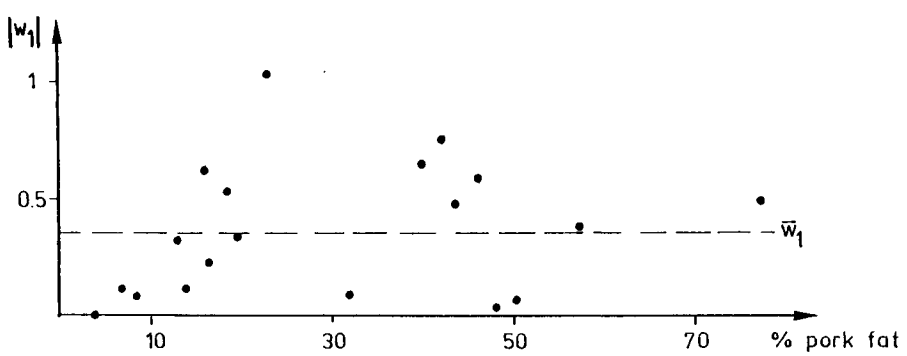


Fig. 13.7. Visual representation of the pork fat data of Table 13.7 for method 1.

the same distribution with true difference δ. If d depends on the concentration this assumption does not hold. In our example, the mean difference, $\bar{d} = -0.099$ (s = 0.351), so that we can write that the differences between the two methods come from a population with mean and confidence interval

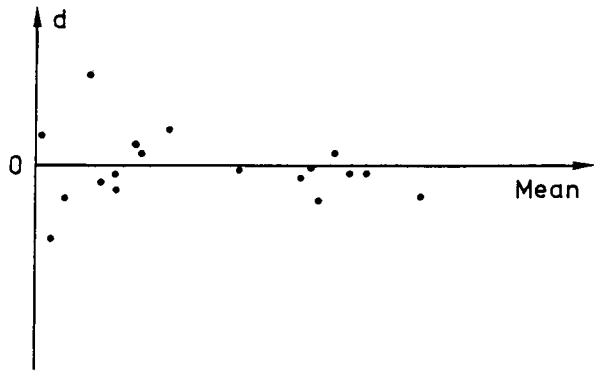


Fig. 13.8. Bland and Altman plot for the data of Table 13.7.

$$-0.099 \pm t_{0.025;18}\,(0.351/\sqrt{19}) = -0.099 \pm 0.169$$

This includes 0 so that the difference is not significantly different from 0: the two methods are equivalent and no bias is detected.

This type of plot, proposed by Bland and Altman [28], is of great diagnostic value by itself. In the present case, without carrying out statistics, it is clear that the two methods are equivalent, except perhaps at the highest concentration and it would be recommended to study this concentration in more detail. This way of plotting permits us to observe certain points that would have escaped attention otherwise. Some typical situations are shown in Figs. 13.9a,b and c. Figure 13.9a would be obtained in the case of an absolute systematic error, b for a proportional error and c is obtained when the variance of the methods depends strongly on concentration.

A problem in applying this method is that, by the selection of real samples of which the concentration is not known *a priori*, one will tend to analyze most samples in the medium concentration range and only a few at the lowest and the highest concentration levels. The data of Table 13.7 are illustrative of the problem: the highest concentration levels are not well represented. Therefore, if at all possible, one should carry out a preselection of the samples, so that the extreme levels are equally well represented as those in the middle.

The same remarks apply, of course, when one carries out the interpretation by regression. The latter is to be preferred in this case to using the *t*-test, because it gives more information. It is very similar to that of the preceding sections (Sections 13.5.4 and 13.5.5). Ideally, the slope should be 1 and the intercept 0. A deviation of the former indicates a proportional discrepancy between the two methods, and if one of the two has been validated earlier (reference method), then the other (test method) is subject to a proportional systematic error. A non-zero intercept is diagnosed as an absolute discrepancy or an absolute systematic error.

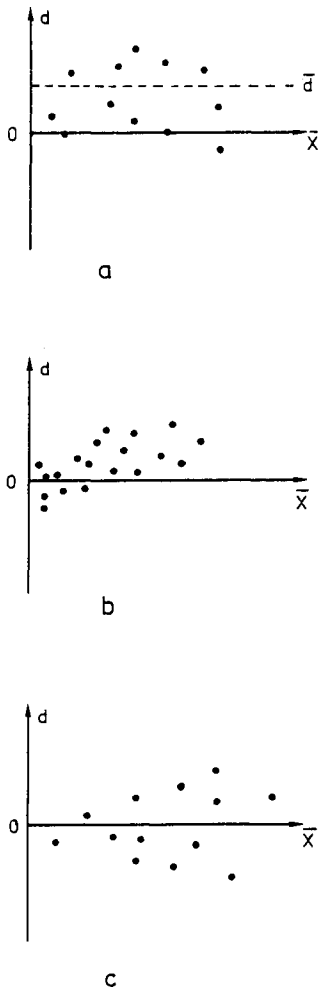


Fig. 13.9. Bland and Altman plots for situation with (a) absolute systematic error; (b) proportional error; (c) heteroscedasticity.

One important problem is the following. In the preceding sections, we plotted on the x-axis a known amount or concentration and on the y-axis the experimental result. In other words, we considered the x-values to be free from random error and the y-values not. Ordinary least squares regression can then be applied. However, when comparing two laboratories or two methods both the x and the y are experimental results. We should use the model of Section 8.2.10 with residuals orthogonal to the regression line, instead of parallel to the y-axis. Using orthogonal residuals means that similar precisions are assumed for both methods. If necessary,

different precisions can also be taken into account but this is very rarely, if ever, done. It should be noted that in practice, the classical regression model is often applied instead. This is acceptable only when the measurement error variance in x (i.e. reproducibility or repeatability — according to experimental conditions — squared), which we will call here σ^2_{mx}, is small compared with the spread of the x values over the concentration range as measured by the variance of the x-values, σ^2_x. From simulations by Hartmann et al. [29], it follows that errors are made quite easily in practice and it therefore seems advisable *not* to apply ordinary least squares at all in such a comparison.

A method which is often applied, but should in fact *never* be used in this type of study is to compute the correlation coefficient between the results obtained with the two methods. The correlation coefficient is a measure of association and all it can be used for is to decide whether two methods give indeed related results. As this is the least one expects, measuring the correlation coefficient is of no use in this context (see also Section 13.6.1).

13.5.7 An alternative approach to hypothesis testing in method validation

All the methods described so far test whether there is no bias e.g. H_0: $\mu_0 = \mu$ as in Section 13.5.2. However, fundamentally it is improbable that there should be no difference at all. We should state rather that the difference should not be larger than a given bias. Such a situation is similar to that described in Chapter 4.10. In this section, it was explained that instead of the use of point hypothesis tests, we may prefer to apply interval hypothesis tests. This was also proposed for method validation by Hartmann et al. [30]. They showed that this also allows a better way of taking β-error into account, which, as stated in preceding sections is often not done at all in method validation.

13.5.8 Comparison of more than two methods or laboratories

From time to time, more than two methods will have to be compared. This will usually not be the case in the full validation step, but it may happen during the exploratory stage. As an example, we refer to Table 6.1a. In this case we were interested in the analysis of a specific mineral–vitamin formulation. The experimental set-up consisted of analyzing 6 replicates of the material by 6 different methods. The statistical analysis is carried out by one-way ANOVA. As explained in Section 13.2, when the concentration levels cover a larger range, or, if different types of matrices can occur, we should analyze several materials. This then constitutes a two-way ANOVA (materials, methods). We are interested in the factor methods and the interaction between materials and methods. The variance due to materials is not of interest, but has to be taken into account in the ANOVA.

TABLE 13.8

Comparison of wet oxidation methods for the analysis of Se (results in μg Se/100 ml sample). Adapted from Ref. [27].

Material	Procedure				
	A	B	C	D	E
1	6.0	7.8	6.7	4.0	8.6
2	16.8	21.6	19.5	18.5	19.6
3	11.9	17.5	13.7	10.5	16.6
4	45.8	48.9	44.6	44.1	49.3
5	75.7	76.6	74.4	74.9	78.6
6	54.4	56.8	54.6	51.4	56.3
7	86.1	90.4	89.0	84.1	89.4
8	19.7	23.9	18.9	18.3	21.5
9	125.9	130.8	127.1	124.2	128.6

ANOVA analysis

Source	df	Sum of squares	Mean square	*F*
Materials	8	65573.71		
Procedures	4	150.81	37.70	34.0
Materials × procedures (= residue)	32	35.63	1.11	

Here, we will discuss another example of the latter type, taken from Ref. [27]. The data concern the analysis of Se in urine using 5 different wet oxidation procedures. B is a fully validated procedure and the other four are possible alternatives. Nine urines were analyzed and the results are given in Table 13.8.

The example is interesting, because, at first sight, it is not possible to estimate the effect of interaction because there is no replication. The materials require 8 degrees of freedom, the procedures 4 and the interaction 32. Since the total number of results is 45, there are 44 degrees of freedom that are used up by the effects, leaving none for the residual. However one should remember that the mean square is a variance. The interaction effect therefore corresponds to a standard deviation of $\sqrt{1.11} = 1.05$. If this is significantly larger than the experimental error, we would conclude that the effect is significant.

In this case, the experimental error is not obtained from the ANOVA experiment, but it was known from earlier experiments that the repeatability standard deviation was of the order of 1.25. As this is certainly not smaller than 1.05, the

interaction effect cannot be significant and its sum of squares and degrees of freedom can be incorporated into the residual error, which can then be estimated (see also Chapter 6.7). It is a good example of how outside information can be put to advantage. In this case, it eliminated the need for replication. Note also that this is what we called in Chapter 6 repeated testing: MS and F for the factor materials are not calculated. We know that the materials are different and must therefore include the effect into the ANOVA to filter out its contribution to the overall variation, but it is not relevant to compute the variance of that effect (MS) nor its significance (through F). As a conclusion, we need only to assess the effect of the factor procedures. Since $F_{0.05;4,32} = 2.67$, there is a clear effect of the procedures. Further statistical analysis shows that only E gives equivalent results with B and that this is the only alternative procedure meriting further consideration.

It is also very useful to carry out intercomparisons with a few laboratories. This type of study is not the complete intercomparison, which will be described in Chapter 14 and which requires much work and planning, but rather an additional step in the intralaboratory validation of a few laboratories with common interests or a preliminary step in a true interlaboratory study. The experimental set-up then consists of q laboratories (often only 3 or 4) using p methods (often 1) to analyze m different materials in n-replicate. As an example, we give here an experiment preliminary to the proposal of a new method for the titrimetric analysis of chlorpromazine in the European Pharmacopeia [31]. Three methods were studied, one being the existing one, and another being proposed because it does not require mercury salts and is therefore environmentally more friendly. Four laboratories participated and three materials were analyzed in 10-replicate. The analysis of the resulting data set can be carried out using a three-way ($q \times p \times m$) ANOVA, but it is not necessarily a good idea to apply this without further thought. Instead it is preferable to have a look at the data first with box plots (see Chapter 12). Indeed, because of the preliminary nature of the intercomparison it is possible that the data contain outliers and that the precisions of the laboratories are very different. If this is the case, classical ANOVA becomes a doubtful proposition and we prefer in this case to use randomization tests (see Chapter 12).

Figure 13.10 shows the box plot for the three methods on one of the materials. A figure like this immediately allows some conclusions to be made. There is no evidence for systematic differences in concentration levels found between the three methods, so that there probably is no systematic error in any of the methods. Also laboratory 1 works with consistently better precision than 3 and the correspondence between laboratories is best with method B. Because this is an indication of ruggedness, method B was chosen for further validation.

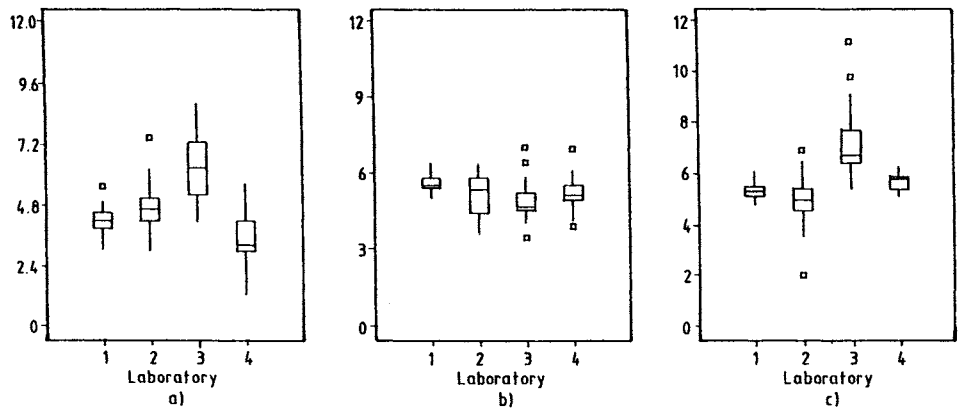


Fig. 13.10. Box plot for the comparison of three methods, (a), (b), (c), carried out by four laboratories (adapted from Ref. [31]).

13.6 Linearity of calibration lines

Most measurement techniques make use of a calibration graph to estimate the analyte concentration in unknown samples. This implies that a decision concerning the nature of the relationship between the concentration and the response has to be taken. Very often a simple straight line relationship is preferred and many measurement techniques are designed to achieve proportionality between response and concentration. In practice, however, deviations from linearity are frequently observed. Therefore, it is essential for a method validation program to include a linearity test. Such tests are discussed in this section.

Before discussing these tests, it should be noted that in pharmaceutical guidelines [17,32] linearity is used in a different context. For instance, the Committee for Proprietary Medicinal Products [32] defines *linearity of a test procedure* as "its ability (within a given range) to obtain test results directly proportional to the concentration (amount) of analyte in the sample". The US Pharmacopeia [17] adds that the mathematical treatment normally is a calculation of a regression line by the method of least squares of test results versus analyte concentrations. In other words, methods described in the context of the determination of bias are used (Section 13.5.4). It also follows that if methods have been shown to be unbiased at the lowest and highest levels of interest and at an intermediate one (for instance, the nominal level when this notion applies), that there is no sense in determining the linearity of the test procedure. As the bias is always investigated at at least three levels including two extreme levels, the determination of the linearity of a test procedure is superfluous. Linearity in this section will therefore mean *linearity of the calibration line.*

13.6.1 The correlation coefficient

As described in Section 8.3 the correlation coefficient, r, between x and y evaluates the degree of linear association between the two variables. It only indicates whether the variables vary together linearly. Therefore, as shown in Fig. 13.11, a correlation coefficient very close to 1 can also be obtained for a clearly curved relationship. Consequently, the correlation coefficient, which is commonly used, in itself is not a useful indicator of linearity.

Nevertheless, calculation of the correlation coefficient is acceptable for a system suitability check if the full method validation has established linearity between the response and the concentration. The check could further consist in a comparison of the correlation coefficient with a default value specified from the method validation results. For instance, one could require that $r > 0.999$. If r is found to be less, this is taken to mean that the calibration line is not good enough. The reason for this can then be ascertained further through visual inspection.

13.6.2 The quality coefficient

Another suitability check is the calculation of the *quality coefficient.* It was defined by Knecht and Stork [33] to characterize the quality of straight line calibration curves and is calculated from the percentage deviations of the calculated x-values from the ones expected:

$$g = \sqrt{\frac{\Sigma(\%\ \text{deviation})^2}{n-1}} \tag{13.6}$$

with

$$\%\ \text{deviation} = \frac{x_{\text{predicted}} - x_{\text{known}}}{x_{\text{known}}}\ 100$$

The better the experimental points fit the line, the smaller the quality coefficient.

A similar expression, based on the percentage deviations of the estimated response, has been used by de Galan et al. [34] in their evaluation of different models to fit curved calibration lines in atomic absorption spectrometry:

$$QC = 100\ \sqrt{\frac{\sum\left(\frac{y_i - \hat{y}_i}{\hat{y}_i}\right)^2}{n-1}} \tag{13.7}$$

with y_i = the measured response, and $\hat{y}_i$ = the response predicted by the model.

Since these expressions are calculated from the percentage deviation of either the calculated x or y values, they assume that the relative standard deviation (RSD)

is constant. For the evaluation of straight line calibration lines with homoscedastic measurements the following adaptation of the quality coefficient has been proposed [35]:

$$QC = 100\sqrt{\frac{\sum\left(\frac{y_i - \hat{y}_i}{\bar{y}}\right)^2}{n-1}} \tag{13.8}$$

Each residual $(y_i - \hat{y}_i)$ being related to the same absorbance value, i.e. $\bar{y}$, which is the mean of all responses measured, implicitly means that the absolute deviation is considered constant.

If a target value for the quality coefficient has been specified, for instance from the full method validation or from previous experience, the suitability of a calibration line can be checked. The line is unacceptable if its QC value exceeds the target value. As an example, consider the calibration line of Fig. 13.11 which is also given in Table 13.9. The straight line equation is $\hat{y} = 0.051 + 0.3225x$. From this fitted line the QC (eq. 13.8) is calculated as follows (see Table 13.10):

$$\text{QC} = 100\sqrt{\frac{2.342\ 10^{-2}}{8}} = 5.4\%$$

The calibration line is from atomic absorption spectrometry for which a target value of 5% has been proposed [35]. Consequently it is concluded that the line is unacceptable, in this case due to non-linearity.

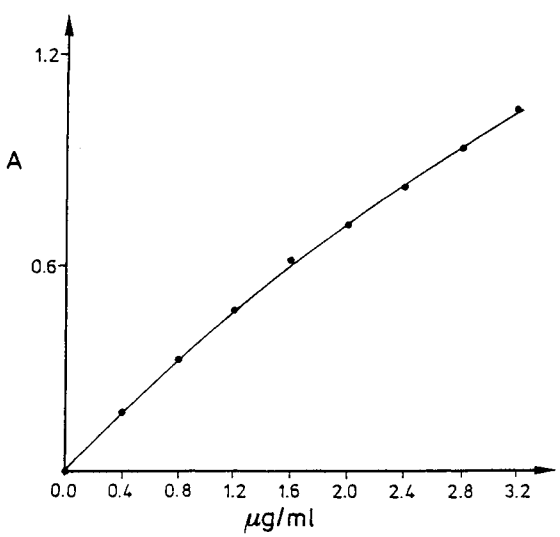


Fig. 13.11. Example of a curved calibration line with a correlation coefficient close to 1 ($r = 0.996$). The data are given in Table 13.9.

TABLE 13.9

Test of linearity: absorption signal y measured as a function of concentration x

x (μg/ml)	y(A)
0.00	0.000
0.40	0.173
0.80	0.324
1.20	0.467
1.60	0.610
2.00	0.718
2.40	0.825
2.80	0.938
3.20	1.048

TABLE 13.10

Calculation of the QC for an atomic absorption calibration line (y = absorption signal, x = concentration)

x_i (μg/ml)	y_i (A)	$\hat{y}_i$ (A)	$((y_i - \hat{y}_i)/\bar{y})^2$
0.0	0.000	0.051	$8.090\ 10^{-3}$
0.4	0.173	0.180	$0.152\ 10^{-3}$
0.8	0.324	0.309	$0.700\ 10^{-3}$
1.2	0.467	0.438	$2.616\ 10^{-3}$
1.6	0.610	0.567	$5.751\ 10^{-3}$
2.0	0.718	0.696	$1.505\ 10^{-3}$
2.4	0.825	0.825	0
2.8	0.938	0.954	$0.796\ 10^{-3}$
3.2	1.048	1.083	$3.810\ 10^{-3}$
	$\bar{y} = 0.567$		$\Sigma = 2.342\ 10^{-2}$

$QC = 100\sqrt{2.342\,10^{-2}/8} = 5.4\%$

Since in eqs. (13.7) and (13.8) the QC corresponds to a relative residual standard deviation it seems in fact more logical to divide by $(n - 2)$ rather than by $(n - 1)$, the former also being used in the expression of the residual standard deviation (see eq. (8.6)).

The QC is to be preferred over the correlation coefficient not only because it gives a better idea of the spread of the data points around the fitted straight line but also because it gives some indication on the percentage error to be expected for the estimated concentrations. Moreover, the QC can also be used in the evaluation of more complex calibration models. Division by $(n - p)$, p being the number of regression coefficients included in the model, might then also be preferred.

13.6.3 The F-test for lack-of-fit

In full method validation we use either the *F*-test for lack-of-fit or the one described in Section 13.6.4. Both are included in draft IUPAC guidelines for calibration in analytical chemistry [36]. The test for lack-of-fit described in Section 8.2.2.2 verifies whether the straight line model adequately fits the calibration data. As pointed out earlier it requires that replicate measurements are available to estimate the pure experimental error. Because of this requirement the *F*-test for lack-of-fit is generally restricted to the full method validation program. For a worked example the reader is referred to Example 2 of Chapter 8.

13.6.4 Test of the significance of b_2

Another possibility to test linearity of a calibration graph is to fit a second degree polynomial to the data:

$y = b_0 + b_1x + b_2x^2$

A straight line relationship is demonstrated if the quadratic regression coefficient, b_2, is not significant. The hypothesis that the quadratic term is zero (H_0: $\beta_2 = 0$; H_1: $\beta_2 \neq 0$) can be tested by means of the confidence interval for β_2 or by means of a *t*-test (see Section 10.4). In the *t*-test the absolute value of

$t = b_2/s_{b_2}$

with s_{b_2} the standard deviation of b_2 as obtained from eq. (10.18), is compared with the tabulated *t* for $n - 3$ degrees of freedom at the chosen confidence level.

Example:

Table 13.9 gives the data shown in Fig. 13.11. The second degree equation is:

$y = 0.0064 + 0.4181\ x - 0.0299\ x^2$

and $s_{b_2} = 0.00324$

From the 95% confidence interval for β_2:

$-0.0299 \pm (t_{0.05;6} \times 0.00324)$

$-0.0299 \pm (2.447 \times 0.00324)$

-0.0299 ± 0.0079

or from the absolute value of:

$t = -0.0299/0.00324 = -9.228$

as compared to $t_{0.05;6} = 2.447$, it is concluded that the quadratic term is not zero. Consequently non-linearity is demonstrated.

ISO [37,38] and IUPAC [36] include this approach, although in a different form, to evaluate the linearity of the calibration line. The test is performed by means of the partial *F*-test discussed in Section 10.3.2. If non-linearity is detected ISO recommends either reducing the working range in order to obtain a straight line calibration function or, if this is not possible, using the quadratic calibration function. However, it should be noticed that the significance of the quadratic term does not imply that the second degree model fits the data correctly. This was also recognized by Penninckx et al. [39] who propose a general strategy for the validation of the calibration procedure that makes use, among others, of the ANOVA lack-of-fit test, the test of the significance of b_2 discussed in this section and a randomization test (see Section 12.4) for lack-of-fit.

13.6.5 Use of robust regression or non-parametric methods

Robust regression methods can be applied to detect non-linearity. Indeed, robust regression methods are not sensitive to outliers. The use of the straight line model when a deviation from linearity occurs, will result in model outliers (i.e. outliers due to the erroneous use of the straight line model). They can be detected with robust regression methods such as the least median of squares (LMS) method as described in Section 12.1.5.3.

In principle, LMS could be used in full validation. However, we prefer for this purpose the methods described in the preceding sections. The method can, however, be applied for a system suitability check to diagnose problems with the calibration line, if *QC* or *r* exceed their threshold value. Its use for this purpose, in conjunction with another test, called the slope ranking method, was proposed by Vankeerberghen et al. [40].

13.7 Detection limit and related quantities

An important characteristic of an analytical method is the smallest concentration of the analyte that can be detected with a specified degree of certainty. In the seventies IUPAC [41] stated that the limit of detection, expressed as the concentration, x_L, or the quantity q_L, is derived from the smallest measure y_L, that can be detected with reasonable certainty for a given analytical procedure, where

$$y_L = \bar{y}_{bl} + ks_{bl} \tag{13.9}$$

with $\bar{y}_{bl}$ the mean of the blank responses, s_{bl}, the standard deviation of the blank responses and k a constant. The detection limit x_L (or q_L) is obtained as:

$$x_L \text{ (or } q_L) = ks_{bl}/S \tag{13.10}$$

with S, the sensitivity of the analytical method (see Section 13.8), corresponding to the slope of the calibration line. IUPAC strongly recommends to use a value of $k = 3$.

In general terms, the detection limit has been defined as that concentration which gives an instrument signal (y) significantly different from the blank signal. The different interpretations of the term "significantly different" have resulted in different definitions for the quantification of the detection limit and this has led to a lot of confusion. The fact that both blank and sample measurements are subject to error requires the problem of chemical detection to be treated in a statistical way. This implies that detection decisions are prone to the two kinds of errors associated with any statistical testing: false positive decisions (type I or α-error) and false negative decisions (type II or β-error). Traditional approaches for determining detection limits (such as the former IUPAC definition [41]) only provide protection against type I errors. They do not take the β-error into account. According to Currie [42, 43] three limiting levels are required to completely describe the detection capabilities of an analytical method: (1) the decision limit at which one may decide *a posteriori* whether or not the result of an analysis indicates detection, (2) the detection limit at which a given analytical procedure may be relied upon to lead *a priori* to detection and (3) the determination limit (or quantification limit) at which a given procedure will be sufficiently precise to yield a satisfactory quantitative estimate. The decision limit is related to the question "Has something been detected?", the detection limit to the question "How little can be detected?". The most recent IUPAC Nomenclature document [3] recognizes the necessity to consider both α and β errors and includes the different limits specified above.

The discussion of the detection limit and related quantities in this section is based on papers by Currie [42], Hubaux and Vos [44], Winefordner and Long [45] and Cheeseman and Wilson [46] and on a textbook edited by Currie [43]. Terminology and symbols for the measurement limits in this discussion are as far as possible as recommended by IUPAC [3]. In the literature and in some specific guidelines several other terms and symbols are used.

13.7.1 Decision limit

Let us first consider the blank measurement. The blank is a sample which is identical to the sample of interest except that the analyte to be measured is not present. The measurement of that blank is of course also subject to error, which we consider to be normally distributed. This means that a sufficiently large number of observations on the blank can be represented by a normal distribution of the responses, with mean μ_{bl}, the true blank value, and standard deviation σ_{bl} as shown in Fig. 13.12. Now consider a response, L_C, which is made k_C standard deviations away from the mean blank value:

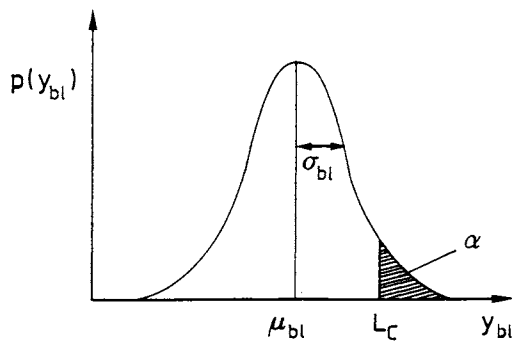


Fig. 13.12. Normal distribution of the blank measurements.

$$L_C = \mu_{bl} + k_C\sigma_{bl} \tag{13.11}$$

The probability to measure a blank signal, y_{bl}, which is larger than L_C is equal to α. The higher L_C is, compared to the mean blank value, the less probable it becomes to obtain a blank signal that is larger than L_C. From this it follows that if L_C is sufficiently larger than the mean blank value a measured signal which is larger than L_C is unlikely to be due to the blank. If signals larger than L_C are interpreted as "component present", then only a fraction α of blanks will be (mis)interpreted as "component present".

The critical value, L_C, thus depends both on the standard deviation of the blank measurements and on the risk one is willing to take of making a wrong decision. For $k_C = 3$, L_C is equal to y_L as formerly defined by IUPAC [41]. A value of $k_C = 3$ corresponds with a probability $\alpha = 0.13\%$ that a signal larger than L_C is due to a blank. Therefore, it can be concluded with a high probability ($1 - \alpha = 99.87\%$) that the component has been detected.

However, if a signal is measured which is lower than L_C it cannot, with the same certainty, be concluded that the component is not present. To explain this, consider a sample with a true concentration corresponding to an average response L_C. The distribution of an infinite number of repeated measurements on this sample is represented, together with the distribution of the blank measurements, in Fig. 13.13. A normal distribution with a standard deviation equal to σ_{bl} is assumed. Note that 50% of the signals observed for the sample will be smaller than the limit L_C. Therefore, the statement that the component is absent if the measurement is smaller than L_C is very unreliable. Indeed, the probability of not detecting the analyte when it is present with a concentration yielding a signal L_C, is 50% (= the β error, the probability of false negative decisions — see Chapter 4). Consequently, with this limit L_C, the probability to decide that the analyte is present when in fact it is absent (= α error) is small whereas the probability to decide that the analyte is absent when in fact it is present (= β error) is very large.

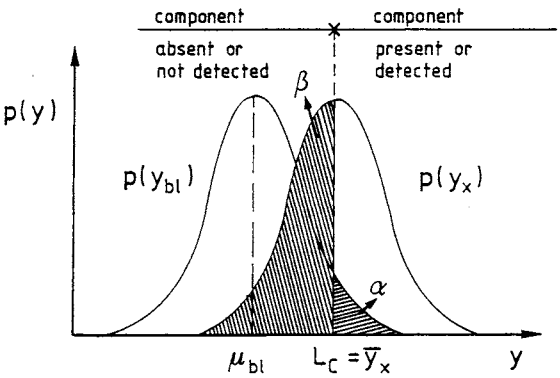


Fig. 13.13. Illustration of the decision limit, L_C.

Due to the large β error it has been proposed to use this limit only for an *a posteriori* decision about the presence of a component, i.e. a decision after the signal is measured. It is then defined as the *critical level* or *decision limit* above which an observed signal may be reliably recognized as detected [3,42,43]. IUPAC [3] proposes a default value for α equal to 0.05. This corresponds with $k_C = 1.645$.

13.7.2 Detection limit

To reduce the β error, so that eventually the α and the β error are better balanced, the two distributions in Fig. 13.13 have to be separated to a larger extent. In Fig. 13.14 the situation is represented where α = β. It is assumed that σ at the detection limit L_D is equal to σ_{bl}.

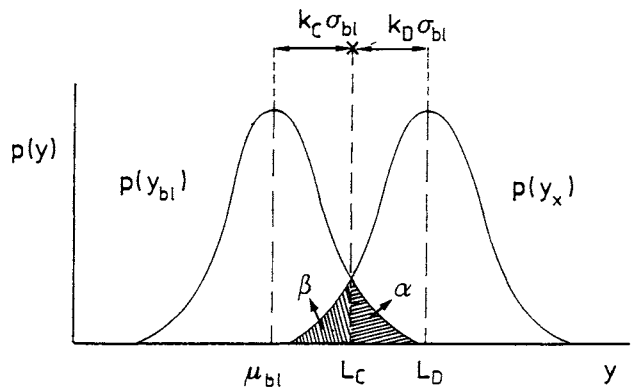


Fig. 13.14. Illustration of the detection limit, L_D.

The detection limit L_D is set k_D standard deviations away from L_C:

$$L_D = L_C + k_D \sigma_{bl} \tag{13.12}$$

$$= \mu_{bl} + k'_D \sigma_{bl} \quad \text{with } k'_D = k_C + k_D$$

From Fig. 13.14, where the relationship between the critical level, L_C, and the detection limit, L_D, is illustrated, it follows that L_D has the following meaning: for a sample that does not contain the analyte (the true concentration corresponds to an average response μ_{bl}) less than α% of the measurements will exceed L_C. For a sample with a true concentration corresponding to a response L_D only β% of the measurements will be below L_C and are indistinguishable from the blank. Therefore, given L_C (or α), L_D protects against false negative decisions.

In Fig. 13.14 $\alpha = \beta$ because $k_C = k_D$. With for example $k_C = k_D = 3$, $\alpha = \beta = 0.13\%$ and

$$L_D = L_C + 3\,\sigma_{bl} = \mu_{bl} + 6\,\sigma_{bl} \tag{13.13}$$

Therefore, with L_D as detection limit, as defined by eq. (13.13), we run a risk of at most 0.13% to conclude that the component is absent when in fact it is present. Consequently, the risk for both false positive (α) or false negative results (β) is very small.

Taking the default values for both α and β equal to 0.05, IUPAC [3] proposes a multiplication factor equal to 3.29 (= 2 × 1.645).

13.7.3 Quantification limit

The *determination* or *quantification limit*, L_Q, is defined as the level at which the measurement precision will be satisfactory for quantitative determination (IUPAC [3] recommends not to use the term determination limit). In other words, the quantification limit is the concentration that can be determined with a fixed maximum relative standard deviation (RSD) and a suitable accuracy.

It is defined as:

$$L_Q = \mu_{bl} + k_Q \sigma_{bl} \tag{13.14}$$

If for quantitative determination an RSD of 5% is required, k_Q should be 20. The relative standard deviation at the level L_Q is thus $1/k_Q$ (= $\sigma_{bl}/(L_Q - \mu_{bl})$). Consequently, the relative standard deviation of the quantitative measurement at the decision level L_C is 33.33% and at L_D 16.67%. IUPAC [3] proposes a default value of 10 for k_Q. The above definition of course assumes that σ at the quantification limit is equal to σ_{bl}. In practice therefore, L_Q is preferably determined from the precision measured at the level thought to be equal to L_Q.

The SFSTP group [13] specifies that the precision and accuracy have to be evaluated at the quantification limit by preparing n independent samples ($n \geq 6$)

containing the component to quantify at the concentration L_Q and performing the test procedure on each sample.

The choice of the k values (or the t values — see next section) determines the risk one is willing to take of making a wrong decision. However, the different values for the constants used have contributed to the existing confusion about detection limits. Therefore, when reporting the lower limits of measurement, the way they are defined and determined should be specified. In this context, the Analytical Methods Committee of the Royal Society of Chemistry [47] recommends the former IUPAC definition for the detection limit which as mentioned earlier, specifies $k = 3$, and discourages the use of the other lower limits of measurement (decision limit and quantification limit). The committee prefers a simple operational definition for the detection limit that is regarded as a rapidly acquired but approximate guide to performance of an analytical system. Because of difficulties in the interpretation of detection limits it is considered that there is no point in trying to estimate them very precisely, or in defining them strictly in terms of confidence intervals. We agree that this makes sense but in many disciplines such as the bio-analysis of drugs [16], one has to state a quantification limit.

13.7.4 Measuring the blank

The expressions for L_C, L_D and L_Q are based on the mean blank and on the variability of the blank. To ensure that realistic estimates of these limits are made it is important to select the appropriate blank. A *solvent* or *reagent blank*, which is the solution that contains the reagents in the same quantity used to prepare the calibration line or to dilute the sample, may give detection limits which are too optimistic. If our main interest, however, is in the comparison of detection limits of different instruments such a solvent blank is perfectly useful. An *analytical blank* contains all reagents and has been analyzed in the same way as the samples. It is a blank solution which has been taken through the whole procedure, from the pretreatment up to the measurement and therefore it is much more appropriate to determine the detection limit of the analytical method that is being validated. Therefore detection limits based on the signal to noise ratio should only be used if they can be obtained from the entire measurement process. The ideal blank is the *matrix blank* which has exactly the same composition as the sample except for the analyte to be analyzed. In some situations a sample in which the analyte is not present can be obtained e.g. if a drug has to be determined in blood, blank blood without the drug can usually be obtained. Alternatively, if a blank sample is not available, the variability of a sample with a very low analyte concentration (concentration near the detection limit) can also be used for the evaluation of the detection and quantification limits.

Good estimates of the mean and the standard deviation of the blank require a reasonable number of blank measurements, n. The fact that the calculated standard deviation is only an estimate can be taken into account by replacing the constants k_C, k_D and k_Q in eqs. (13.11, 13.12 and 13.13), which are derived from the standardized normal distribution, by t values (t_C, t_D and t_Q for $n - 1$ degrees of freedom). However, if a reasonable number of blank measurements have been made, the values obtained for the different limits will be very similar to those given earlier. IUPAC [3] specifies that when σ is estimated as s, a confidence interval must be given for the detection limit L_D to take the uncertainty in s into account. This can be done by considering the confidence interval for σ as derived from eq. (5.16).

Another approach, which is also included in the ICH document [9], is to estimate the standard deviation of the blank from the residual standard deviation of the calibration line (s_e, see eq. (8.6)). It should be obvious that this is only useful provided (i) that the calibration standards have the same composition as the samples to be analyzed, (ii) that the calibration standards have been taken through the whole analytical procedure and (iii) that the calibration data are homoscedastic.

Practices also differ with respect to the blank correction. Indeed eqs. (13.11) to (13.14) are based on the comparison of the measured signal with the blank signal. However, when blank correction is part of the analytical procedure, the measured response should first be corrected for the blank response. The decision that the analyte is present is then based on a comparison of the net signal with zero. If y_N represents the net signal, y_S the gross signal and y_{bl} the blank signal:

$$y_N = y_S - y_{bl}$$

Consequently, unless the blank is well known, the variability of the net signal is:

$$\sigma_N^2 = \sigma_S^2 + \sigma_{bl}^2$$

If the standard deviation is independent of the concentration:

$$\sigma_N^2 = 2\,\sigma_{bl}^2$$

When the sample does not contain the analyte, $y_S = y_{bl}$, and their difference follows a normal distribution with a population mean of zero and a standard deviation $\sigma_0 = \sqrt{2}\,\sigma_{bl}$. Therefore, the decision limit and the detection limit for blank corrected signals are given by

$$L_C = k_C\,\sigma_0 = k_C\,\sqrt{2}\,\sigma_{bl}$$

$$L_D = k'_D\,\sigma_0 = k'_D\,\sqrt{2}\,\sigma_{bl} \quad \text{with } k'_D = k_C + k_D \tag{13.15}$$

Equation (13.15) applies for paired comparisons [43]. This means that with each sample (or each batch of samples) a blank is analyzed and each sample response (or the sample responses within a batch) therefore is individually blank corrected.

If the blank correction is performed by subtracting the mean of n blank determinations the equations given above change into

$$y_N = y_S - \bar{y}_{bl}$$

with

$$\sigma_N^2 = \sigma_S^2 + \sigma_{bl}^2 / n$$

Thus, if the standard deviation is independent of the concentration

$$\sigma_0 = \sqrt{1 + (1/n)}\ \sigma_{bl}$$

and

$$L_C = k_C\ \sigma_0 = \sqrt{1 + (1/n)}\ k_C\ \sigma_{bl}$$

$$L_D = k'_D\ \sigma_0 = \sqrt{1 + (1/n)}\ k'_D\ \sigma_{bl} \quad (k'_D = k_C + k_D) \tag{13.16}$$

Notice that the recent IUPAC [3] basic definitions consider the mean value of the blank response as well as the variance of the blank to be precisely known since the expressions for L_D and L_C, given $\alpha = \beta = 0.05$, and for L_Q, given $k_Q = 10$, are:

$$L_C = 1.645\ \sigma_0 \tag{13.17}$$

$$L_D = 3.29\ \sigma_0 \tag{13.18}$$

$$L_Q = 10\ \sigma_0 \tag{13.19}$$

However several of the possible complications discussed in this section are also treated in the IUPAC document [3]. It also briefly discusses the effect of heteroscedasticity on the expressions for the measurement limits.

13.7.5 Concentration limits

The detection limits have been described so far in terms of the measurement signal. They can be re-expressed into *concentration* or *analyte detection limits* by making use of the slope of the calibration line, b_1:

$$x_C = \frac{L_C - \mu_{bl}}{b_1} = \frac{k_C\ \sigma_{bl}}{b_1}$$

$$x_D = \frac{L_D - \mu_{bl}}{b_1} = \frac{k'_D\ \sigma_{bl}}{b_1} \tag{13.20}$$

With $k = 3$ these expressions correspond to the detection limit as formerly recommended by IUPAC (eq. 13.10) which should be reported as $x_{L(k=3)}$.

In eq. (13.20) it is assumed that the blank is well known since in the blank correction the variability of the blank is not taken into account. If this variability is considered, concentration limits have to be calculated from eq. (13.16).

$$x_C = \frac{L_C}{b_1} = \frac{\sqrt{1 + (1/n)}\, k_C\, \sigma_{bl}}{b_1}$$

$$x_D = \frac{L_D}{b_1} = \frac{\sqrt{1 + (1/n)}\, k'_D\, \sigma_{bl}}{b_1} \quad (k'_D = k_C + k_D) \tag{13.21}$$

13.7.6 Example

Lead in foodstuffs is analyzed by means of GFAAS after microwave digestion of 0.25 g material and dilution to a final volume of 25 ml. The calibration equation for standard solutions containing between 0 and 60 ng Pb/ml is $y = 0.002 + 0.00291x$. The variability of the blank is obtained from the analysis, including the microwave digestion, of 10 analytical blanks. The following responses, expressed as peak area, are measured for those blanks:

blank	area
1	0.025
2	0.037
3	0.012
4	0.029
5	0.048
6	0.026
7	0.024
8	0.015
9	0.041
10	0.019

$\bar{x} \pm s = 0.028 \pm 0.012$

For the determination of the detection limit the traditional IUPAC definition as well as some other approaches discussed earlier will be calculated for $\alpha = \beta = 0.05$.

From eqs. (13.9) and (13.10) the traditional IUPAC detection limit is obtained as follows:

$$y_{L(k=3)} = 0.028 + (0.012 \times 3)$$

and

$$x_{L(k=3)} = (0.012 \times 3) / 0.00291 = 12 \text{ ng/ml}$$

Taking into account the weight of material (0.25 g) and the final volume (25 ml) used in the analysis, this corresponds to a detection limit in the original foodstuff of 1.2 μg Pb/g.

At the 5% probability level chosen ($\alpha = \beta = 0.05$) for the illustration of some other approaches, the (one-sided) t-values for 9 degrees of freedom are $t_C = t_D = 1.833$ (as

compared to $k_C = k_D = 1.645$ for df = ∞). We will consider two different situations.

(A) In the analysis each sample absorbance is blank corrected with the mean blank absorbance calculated previously. The decision limit L_C becomes (see eq. (13.16)).

$$L_C = 1.833 \sqrt{1 + (1/10)}\ 0.012 = 0.023$$

The detection limit, L_D, is then 0.046. Those limits are converted into the following concentration limits (see eq. (13.21))

$$x_C = \frac{0.023}{0.00291} = 8 \text{ ng Pb/ml}$$

$$x_D = 2\, x_C = 16 \text{ ng Pb/ml}$$

Taking into account the weight of material (0.25 g) and the final volume (25 ml) used in the analysis, this corresponds to a decision and a detection limit in the original foodstuff of respectively 0.8 μg Pb/g and 1.6 μg Pb/g. According to Kirchmer (see Ref. [43]) for a sample yielding a Pb concentration of 0.9 μg/ml, it can be decided that Pb is present and the result can be reported as such. On the other hand, an estimated sample concentration of 0.6 μg Pb/g should then be reported as <1.6 μg Pb/g to take into account the possibility of false negative decisions. However, this way of reporting is generally not used. We advise to follow the IUPAC [3] recommendation that all results less than the detection limit, including negative values, and their uncertainty are always reported.

(B) In the analysis the variability of the blank is obtained from the replicate blank measurements given earlier but for each batch of sample analyses a single blank determination is performed. The decision limit is then (see eq. (13.15)):

$$L_C = 1.833 \sqrt{2}\ 0.012 = 0.031$$

and the detection limit

$$L_D = 0.062$$

This corresponds to a decision and a detection limit in the original foodstuff of respectively 1.1 μg Pb/g and 2.2 μg Pb/g.

13.7.7 Alternatives

For analytical methods that involve the measurement of a peak on a noisy baseline (e.g. chromatography) the method detection limit (MDL) has been introduced [48]. It is defined as "the minimum concentration of a substance that can be identified, measured and reported with 99% confidence that the analyte concentration is greater than zero and is determined from analysis of a sample in a given matrix containing the analyte". The method detection limit is obtained as

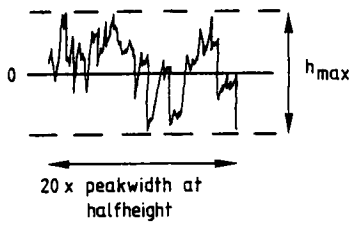


Fig. 13.15. Alternative approach to determine limits of measurement from, e.g., a chromatographic run.

$$\text{MDL} = t_{0.01;n-1}\, s$$

where t is Student's t value at $n - 1$ degrees of freedom and $\alpha = 0.01$ (one-sided) and s is the standard deviation of n replicate analyses of standards or samples with a low concentration of the analyte. The term method detection limit is misleading since the value not only depends on the method but also on the instrument sensitivity, the nature of the samples and the skill of the analyst [43].

For methods such as chromatography the SFSTP [13] advises the following procedure. The complete analysis is performed on a matrix blank and the chromatogram is recorded. The maximum amplitude, h_{max}, over a distance which is equal to twenty times the width at half height of the peak corresponding to the analyte is determined as shown in Fig. 13.15. From this the detection limit is obtained as $3h_{max}R$ and the quantification limit as $10h_{max}R$ where R is the response factor quantity/signal (expressed in peak height). The precision and accuracy at the quantification limit is evaluated by the analysis of samples with a concentration corresponding to the limit of quantification. It should be noted that this procedure applies to signals expressed as peak height. For measurements obtained as peak areas the evaluation of the detection and quantification limits can be based on the variability of a sample with an analyte concentration near the detection limit. The earlier described approaches can then be used.

13.7.8 Determination of the concentration limits from the calibration line

In the following approach for the concentration detection limit, which up to now is not generally practised, allowance is also made for the uncertainty in the calibration line. Indeed, the calibration line is only an estimate of the true regression line. It is possible to take this uncertainty into account by considering the confidence limits of the calibration curve. Therefore, consider Fig. 13.16 in which the lower part of the calibration line with the lower and upper confidence limits are shown.

The possible outcomes for estimations of the response of the blank ($x = 0$) are represented by the distribution drawn at the left of the figure. In this distribution y_C

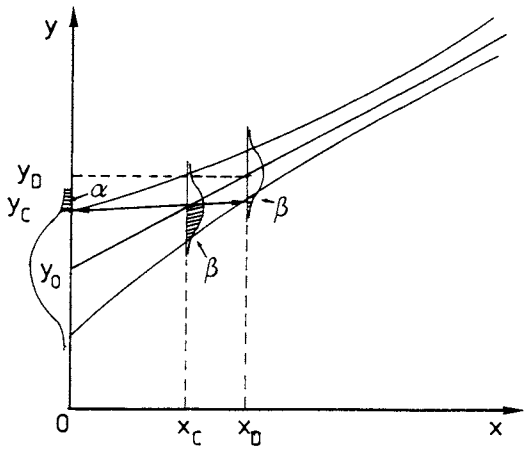


Fig. 13.16. Illustration of the detection limit taking into account the uncertainty in the calibration line.

represents the lowest detectable signal and corresponds to L_C defined earlier: a measured signal larger than y_C is unlikely to be due to the blank, the probability that it is due to the blank being 100α%. From this, the lowest concentration, x_D, which can be distinguished from zero, the blank, can be obtained as the intersection of the horizontal line $y = y_C$ and the curve describing the lower confidence limit. If we measure a sample with an unknown concentration which is smaller than x_D the β error, the risk for false negative decisions, increases. For example x_C cannot be considered as the detection limit because the β error is 50% which is much too high. With x_D as detection limit a better balance between the two types of error, α and β, is obtained. From x_D the corresponding detection limit, expressed in terms of the signal, can be calculated from the calibration function.

The calculation of these limits can be performed as follows:

1. Consider the confidence limits for the mean of m responses at a particular x value, x_0 (see Section 8.2.5.1).

$$b_0 + b_1 x_0 \pm t_{n-2}\, s_e \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

t corresponding to a probability α for the upper limit and β for the lower limit.

2. Compute y_C which is the upper confidence limit (one-tailed) for the mean of m responses when the analyte concentration is zero:

$$y_C = b_0 + t_{\alpha;n-2}\, s_e \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

3. x_D can be obtained in different ways:

– as the intersection of the line $y = y_C$ with the curve describing the lower confidence limit, $y = y_L$, where:

$$y_L = b_0 + b_1 x_D - t_{\beta;n-2}\, s_e \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_D - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

This calculation is rather cumbersome.

– by iteration. x_D is then defined as the lowest value of x which gives a value for y_L exceeding or equal to y_C.

– the following approximation has been proposed by the AOAC [27]:

$$x_D = x_C + t_{\beta;n-2}\, s_e / b_1 \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(2x_C - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

This originates from:

$$x_C = x_D - t_{\beta;n-2}\, s_e / b_1 \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_D - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

x_C being the lower $(1 - \beta)$ % confidence limit for a predicted concentration x_D and the exact value of x_D being near 2 x_C.

4. From the calibration line, if necessary, calculate the corresponding limit in terms of the response

$$y_D = b_0 + b_1 x_D$$

The detection limit determined in this way can be decreased by improving the precision (s_e), increasing the number of standards (n) and the number of measurements made for the sample (m). The calibration design, e.g. the concentration range considered and the distribution of the standards within this range also influences the detection limit through its effect on $\bar{x}$ and $\sum(x_i - \bar{x})^2$. It has been recognized, for example, that limits which do not reflect the real performance capability of the analytical method (because they are too large) can result from a calibration line in which the lowest standard(s) are considerably removed from the origin [43]. It should also be realized that it is assumed that the residual variance, estimated as s_e^2, is equal to the sample measurement precision. When this assumption does not hold, which might, e.g., be the case when simple calibration standards are used, the detection limit will be underestimated. The calibration should then be planned to include the complete measurement process from the pretreatment up to the actual measurement.

IUPAC [3] proposes a propagation of error approach, using the Taylor expansion for the variance of $\hat{x}$ at the detection limit x_D, to take the uncertainty in the calibration line into account. Without an explicit derivation the following expression for the detection limit (homoscedasticity and $\alpha = \beta$) is considered:

$x_D = (2t_{\alpha;df}\,\sigma_0/b_1)\,(K/I)$

where $K = 1 + r(b_0,b_1)\,(\sigma_{b_0}/\sigma_0)\,[t_{\alpha;df}(\sigma_{b_1}/b_1)]$

$I = 1 - [t_{\alpha;df}(\sigma_{b_1}/b_1)]^2$

In this expression $r(b_0,b_1)$ is the correlation coefficient between slope and intercept which is obtained as $-\bar{x}/\sqrt{(\sum x_i^2)/n}$. However it is not indicated how the degrees of freedom are obtained, which are necessary in the selection of the t-value. Moreover it is remarkable that in another IUPAC document [2] this same expression is given with each σ replaced by s.

As a conclusion to the discussion of measurement limits it should be re-emphasized again that, due to the many different approaches that are possible, in reporting these limits it is of paramount importance to specify how they were obtained.

13.8 Sensitivity

13.8.1 Sensitivity in quantitative analysis

IUPAC [3] and ISO [37] define the sensitivity as the slope of the calibration line because a method with a large slope is better able to discriminate between small differences in analyte content. In metrology and in analytical chemistry, the sensitivity is defined as the slope of the calibration line. The reader should note that because one says colloquially that a method is sensitive when it has a low detection limit, sensitivity is sometimes used erroneously in lieu of detection limit.

In fact, there is little sense in including sensitivity as a performance characteristic when it is defined as the slope. It is not sufficient to know the slope of the calibration line to determine whether two concentrations can be discriminated: one also needs the standard deviation on that slope. The smallest difference d that can be distinguished between two signals depends on the standard deviation s of the two signals (which we can consider to be the same for the two) and the risks α and β one takes respectively to conclude there is a difference when there is none and to conclude that there is no difference when it exists (see Chapter 4). To determine the smallest difference d one can distinguish in concentration units one must relate signal to concentration using b_1, the slope. The following equation has been proposed for the sensitivity [13]:

$$d = (t_{1-\alpha/2} + t_{1-\beta})\,s\sqrt{2}\,(1/b_1) \qquad (13.22)$$

where the t-values are determined for $\alpha = 0.05$ (two-sided) and $\beta = 0.05$ (one-sided) for the number of degrees of freedom with which s was determined. Suppose that the relevant precision measure (repeatability, intermediate precision, ...) s was determined with 10 determinations, then $t_{1-\alpha/2} = 2.26$ and $t_{1-\beta} = 1.83$. It then follows that $d = 5.76\,s/b_1$.

13.8.2 Sensitivity and specificity in qualitative analysis

In Chapter 16, we will learn that sensitivity and specificity are used to characterize the quality of assays with a binary output, yes or no (e.g., see Section 16.1.3: are there HIV antibodies in the urine or not?). Recently, these terms have been introduced into analytical chemistry. AOAC [7] proposes the following definitions:

Specificity rate = qualitative — the probability, for a given concentration, that the method will classify the test samples as negative given that the test sample is a "known" negative. Specificity is calculated as number of found negatives/total number of "known" negatives.

Sensitivity rate = qualitative — the probability, for a given concentration, that the method will classify the test sample as positive given that the test sample is a "known" positive. Sensitivity is calculated as number of found positives/total number of "known" positives.

AOAC also proposes the following related definitions:

False positive rate = Number of false positives/total number of "known" negatives.

False negative rate = Number of false negatives/total number of "known" positives.

The proposal states for all four definitions that the term is applicable to immunological assays, microbiological assays, clinical studies and clinical chemistry. The terms false positives and false negatives are now also being used in food analysis [49,50].

13.9 Selectivity and interference

When another substance, a set of substances or the matrix as a whole have an effect on the signal of the analyte measured and this is not accounted for in the method developed, then systematic errors can affect the result and cause bias. This situation is described by stating that there is a lack of selectivity or that interferences occur. There do not seem to be generally accepted and clear definitions for these terms. For instance, several guidelines use "specific" instead of "selective" and some make a distinction between those two terms. In view of the use of the term specific in another context, described in the preceding section, it seems preferable not to use it for the characterization of quantitative analysis procedures.

The term *interference* is often a general term. Van der Linden [51] states "An interfering substance in analytical procedures is one that, at the given concentration, causes a systematic error in the analytical result".

Selectivity and matrix effect or matrix interference have a more restricted meaning. The literature does not make the distinction clearly. In our opinion, the difference lies in the type of systematic error they cause.

Matrix interferences lead to relative systematic errors. The factors yielding such interferences may be physical or chemical and do not lead to a response as such. They affect the slope of the calibration line. The effect can be due to one specific substance (for instance in AAS, the presence of phosphate decreases the slope of the calibration line of Ca, because it forms a compound with it) or to many (for instance, the potential of an ion-selective electrode is affected by the ionic strength and therefore by all the ions that are present). Matrix interferences can be detected by comparing the slope of the calibration line with the relationship between signal and concentration in the matrix, using methods such as standard addition (see Sections 8.2.8 and 13.5.5).

A method is considered *selective* when no concomitant species has a response of its own that adds to that of the analyte. Lack of selectivity would affect the blank: a sample containing all substances in the sample, including the concomitant species, but not the analyte would yield a positive value. If not corrected for, this would lead to a constant systematic error.

Unfortunately, statistics does not help very much in detecting problems with selectivity due to blanks. One must use chemical reasoning, make a list of possible interferents and show experimentally that the substance in question does not influence the result. Often, the interpretation is simple. For instance, in chromatography one can often conclude that the peak of the candidate interferent is completely separated from that of the analyte. When an analyte-free matrix can be obtained, one can analyze this and, if the blank is sufficiently low, conclude that the substances in such a sample do not contribute to the signal with which the analyte will be quantified. This is not a guarantee, since it is always possible that another matrix may contain other concomitant substances, that will be measured. However, Shah et al. [16] write in their guidelines for bioanalysis that analyzing six samples of different origin in this way may be considered as proof of sufficient selectivity. Sometimes, one will do determinations on a number of samples with and without the possible interferent(s). This means one compares two series of measurements sample by sample. This is the same type of comparison as discussed in Section 13.5.5 and one can therefore use the same type of experimental design and statistical interpretation.

In some other cases, bivariate approaches may help. By this we mean that instead of measuring the signal according to one single variable, one can add a second dimension. A bivariate approach is often applied in hyphenated chromatographic techniques such as high performance liquid chromatography (HPLC) with a diode array detector (DAD). If the spectrum measured with the DAD stays constant over the whole length of the peak, this is taken to mean that the peak is due to a single analyte. The interpretation of the data tables obtained is not always straightforward. Sometimes factor analytical techniques (Chapter 40) are required.

References

1. J.K. Taylor and H.V. Oppermann, Handbook for the Quality Assurance of Metrological Measurements. Lewis Publ., 1988.
2. L.A. Currie and G. Svehla, Nomenclature for the presentation of results of chemical analysis. Pure Appl. Chem., 66 (1994) 595–608.
3. L.A. Currie, Nomenclature in evaluation of analytical methods including detection and quantification capabilities. Pure Appl. Chem., 67 (1995) 1699–1723.
4. W.D. Pocklington, Harmonized protocols for the adoption of standardized analytical methods and for the presentation of their performance characteristics. Pure Appl. Chem., 62 (1990) 149–162.
5. ISO standard 3534 (E/F) (1993) Statistics — Vocabulary and Symbols Part 1.
6. ISO standard 5725. 1 to 6 (1994), Accuracy (trueness and precision) of measurement methods and results.
7. Association of Official Analytical Chemists, Definitions and calculations proposed for method performance parameters. The Referee, 6–12, March 1995.
8. W. Horwitz, Protocol for the design, conduct and interpretation of collaborative studies. Pure Appl. Chem., 60 (1988) 855–864.
9. International Conference on Harmonisation (ICH) of Technical Requirements for the Registration of Pharmaceuticals for Human Use, Validation of analytical procedures, Stability of new drug substances and products and Impurities in new drug substances. Draft 1993.
10. Health Protection Branch, Drugs Directorate Guidelines, Acceptable Methods, Draft 31 July 1992.
11. ISO standard 3494 (1976) Statistical interpretation of data — Power of tests relating to means and variances.
12. M. Thompson, Variation of precision with concentration in an analytical system. Analyst, 113 (1988) 1579–1587.
13. Commission SFSTP, Guide de Validation Analytique: Rapport d'une Commission SFSTP I. Méthodologie. STP Pharma Pratiques 2 (4) 205–226, 1992.
14. National Committee for Clinical Laboratory Standards. Evaluation of Precision Performance of Clinical Chemistry Devices, Second Edition; tentative guideline. NCCLS Document EP5-T2, Villanova, PA:NCCLS 1992.
15. W. Horwitz , L.R. Kamps and K.W. Boyer, Quality assurance in the analysis of foods for trace constituents. J. Assoc. Off. Anal. Chem., 63 (1980) 1344–1354.
16. V.P. Shah, K.K. Midha, S. Dighe, I.J. McGilveray, J.P. Skelly, A. Yacobi, T. Layloff, C.T. Viswanathan, C.E. Cook, R.D. McDowall, K.A. Pittman and S. Spector, Analytical Methods Validation: bioavailability, bioequivalence and pharmacokinetic studies. Pharm. Res., 9 (1992) 588–592.
17. US Pharmacopeia, Chapter 1225, Validation of Compendial Assays-Guidelines, Pharmacopeial Forum, pp. 4129–4134, 1988.
18. W.J. Youden and E.H. Steiner, Statistical Manual of the Association of Official Analytical Chemists. The Association of Official Analytical Chemists, Arlington, 1975.
19. Y. Vander Heyden, K. Luypaert, C. Hartmann, D.L. Massart, J. Hoogmartens and J. De Beer, Ruggedness tests on the HPLC assay of the United States Pharmacopeia XXII for tetracycline hydrochloride. A comparison of experimental designs and statistical interpretations. Anal. Chim. Acta, 312 (1995) 245–262.
20. J.A. Van Leeuwen, L.M.C. Buydens, B.G.M. Vandeginste, G. Kateman, P.J. Schoenmakers, M. Mulholland, RES, an expert system for the set-up and interpretation of a ruggedness test in HPLC method validation. Part 2: The ruggedness expert system. Chemom. Intell. Lab. Systems, 11 (1991) 37–55.

21. BCR Information, Report EUR10618EN (1986).
22. I.D. Wilson, Observations on the usefulness of internal standards in the analysis of drugs in biological fluids, in: E. Reid and I.D. Wilson (Eds.), Methodological Surveys in Biochemistry and Analysis, Volume 20. Royal Society of Chemistry, Cambridge, 1990, pp. 79–82.
23. Rules governing Medicinal Products in the European Community, EEC regulation 2377/90, Community procedure for the establishment of maximum residue limits for residues of veterinary medicinal products in foodstuffs of animal origin.
24. F. Bosch-Reig and P. Campins Falcó, H-Point standard additions method. Part 1. Fundamentals and application to analytical spectroscopy. Analyst, 113 (1988) 1011–1016.
25. P. Campíns Falcó, J. Verdú Andrés and F. Bosch-Reig, Development of the H-Point standard additions method for the use of spectrofluorimetry and synchronous spectrofluorimetry. Analyst, 119 (1994) 2123–2127.
26. A. Boeyckens, J. Schodts, H. Vandenplas, F. Sennesael, W. Goedhuys and F. Gorus, Ektachem slides for direct potentiometric determination of sodium in plasma: Effect of natremia, blood pH and type of electrolyte reference fluid on concordance with flame photometry and other potentiometric methods. Clin. Chem., 38 (1992) 114–118.
27. G.T. Wernimont, Use of statistics to develop and evaluate analytical methods. AOAC, Arlington, VA, USA, 1987.
28. J.M. Bland and D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, Feb. 8 (1986) 307–310.
29. C. Hartmann, J. Smeyers-Verbeke and D.L. Massart, Problems in method-comparison studies. Analusis, 21 (1993) 125–132.
30. C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, Y. Vander Heyden, P. Vankeerberghen and D.L. Massart, Reappraisal of hypothesis testing for method validation: Detection of systematic error by comparing the means of two methods or of two laboratories. Anal. Chem., 67 (1995) 4491–4499.
31. J.O. De Beer, B.M.J. De Spiegeleer, J. Hoogmartens, I. Samson, D.L. Massart and M. Moors. Relationship between content limits and assay methods: an interlaboraty statistical evaluation. Analyst, 117 (1992) 933–940.
32. CPMP working party on quality of medicinal products, III/844/87-EN.
33. J. Knecht and G. Stork, Prozentuales und logarithmisches Verfahren zur Berechnung von Eichkurven. Fresenius' Z. Anal. Chemie, 270 (1974) 97–99.
34. L. de Galan, H.P.J. van Dalen and G.R. Kornblum, Determination of strongly curved calibration graphs in flame atomic-absorption spectrometry: Comparison of manually drawn and computer-calculated graphs. Analyst, 110 (1985) 323–329.
35. P. Vankeerberghen and J. Smeyers-Verbeke, The quality coefficient as a tool in decisions about the quality of calibration in graphite furnace atomic absorption spectrometry. Chemom. Intell. Lab. Syst., 15 (1992) 195–202.
36. IUPAC, Guidelines for calibration in analytical chemistry. Part 1. Fundamentals and single component calibration. V.1. Dokument 25/91, final draft 1996.
37. ISO standard 8466-1 (1994) Water quality — Calibration and evaluation of analytical methods and estimation of performance characteristics, Part 1: Statistical evaluation of the linear calibration function.
38. ISO standard 8466-2 (1994) Water quality — Calibration and evaluation of analytical methods and estimation of performance characteristics, Part 2: Calibration strategy for non-linear second order calibration functions.
39. W. Penninckx, C. Hartmann, D.L. Massart and J. Smeyers-Verbeke. Validation of the calibration procedure in atomic absorption spectrometric methods. J. Anal. Atomic Spectrom., 11 (1996) 237–246.

40. P. Vankeerberghen, J. Smeyers-Verbeke and D.L. Massart, Decision support systems for run suitability checking and explorative method validation in electrothermal atomic absorption spectrometry. J. Anal. Atomic Spectrom., 11 (1996) 149–158.
41. Nomenclature, Symbols, Units and Their Usage in Spectrochemical Analysis — II. Spectrochim. Acta, Part B, 33 (1978) 242.
42. L.A. Currie, Limits for qualitative detection and quantitative determination. Anal. Chem., 40 (1968) 586–593.
43. L.A. Currie (Ed.), Detection in Analytical Chemistry. American Chemical Society, Washington DC, 1988.
44. A. Hubaux and G. Vos, Decision and detection limits for linear calibration lines. Anal. Chem., 42 (1970) 849–855.
45. J.D. Winefordner and G.L. Long, Limit of detection — A closer look at the IUPAC definition. Anal. Chem., 55 (1983) 712A–724A.
46. R.V. Cheeseman and A.L. Wilson, Manual on Analytical Quality-Control for the Water Industry. Water Research Center Medmenham, UK, 1978.
47. Analytical Methods Committee, Recommendations for the definition, estimation and use of the detection limit. Analyst, 112 (1987) 199–204.
48. Methods for Organic Chemical Analysis of Municipal and Industrial Wastewater, Environmental Protection Agency Publication EPA-600 14-82-057, July 1982.
49. Codex Alimentarius Commission, CXIMAS 92/15, Rome 1992, Criteria to limit the number of false positive and false negative results for analytes near the limit of detection.
50. W.G. de Ruig and H. van der Voet, Is there a tower in Ransdorp? Harmonization and Optimization of the Quality of Analytical Methods and Inspection Procedures, pp. 20–43 in: D. Littlejohn, D. Thorburn Burns, Reviews on Analytical Chemistry — Euroanalysis VIII. Royal Society of Chemistry, Cambridge, 1994.
51. W.E. van der Linden, Pure Appl. Chem., 61 (1989) 91–95.

Chapter 14

Method Validation by Interlaboratory Studies

14.1 Types of interlaboratory studies

Interlaboratory studies are studies in which several laboratories analyze the same material. Three types can be distinguished.

– *Method-performance* or *collaborative studies* in which the performance characteristics of a specific method are assessed. These performance characteristics usually have to do with precision. How to proceed in this case has been described in an ISO guideline [1], in which this type of study is called a *precision* study. The AOAC/IUPAC protocol [2] can be seen as amending the ISO guideline. ISO itself [3] recently has amended its guideline and has also published a guideline to estimate the bias of a measurement method.

– *Laboratory-performance* or *proficiency studies*, in which a material is analyzed of which the true concentrations are known or have been assigned in some way, often from the interlaboratory experiment itself. The participants apply whatever method is in use in their laboratory. The results of the laboratories are compared to evaluate the proficiency of individual laboratories and to improve their performance. IUPAC [4] describes a protocol for the proficiency testing of analytical laboratories. Recommendations are also given by the Analytical Methods Committee [5].

– *Material-certification studies* in which a group of selected laboratories analyzes, usually with different methods, a material to determine the most probable value of the concentration of a certain analyte with the smallest uncertainty possible. The objective of such a study is to provide reference materials. This latter type of study is very specialized and reserved to institutions created for that purpose, so that we will not discuss it here.

14.2 Method-performance studies

14.2.1 Definition of reproducibility and repeatability

Repeatability refers to precision as measured under repeatability conditions. These have been defined as follows [1]:

> "Repeatability conditions are conditions where mutually independent test results are obtained with the same method on identical test material in the same laboratory by the same operator using the same equipment within short intervals of time."

Reproducibility refers to reproducibility conditions and these were defined as follows:

> "Conditions where test results are obtained with the same method on identical test material in different laboratories with different operators using different equipment."

In other words, reproducibility of a method is measured in collaborative precision studies of a (proposed) standard method, in which several laboratories analyze the same material.

Repeatability and reproducibility conditions are extremes. In the former, the operator, the instrument, and the laboratory are the same and the time interval is kept short. In the other, the laboratory is changed, so that operator and instrument are also different and the time interval is greater. In interlaboratory studies only these extremes are of interest, but we should remember (see Chapter 13) that several intermediate measures of precision are possible.

All reproducibility measures lead to variance models with a within-lab and a between-lab component. Before explaining this model we note the convention that everything related to repeatability is represented by r and everything related to reproducibility by R. For instance, the standard deviation obtained experimentally in a repeatability experiment is written down as s_r and that in a reproducibility experiment as s_R.

The basic statistical model is a random effects model (see Chapter 6.1.4)

$$y = \bar{y} + B + e \tag{14.1}$$

where $\bar{y}$ is the general average for the material analyzed, B represents the laboratory component of bias and e is the random error. The latter is estimated by s_r^2 the repeatability variance, while B gives rise to s_L^2, the between-laboratory variance. By definition the mean value of B is zero. We can then write that:

$$s_R^2 = s_r^2 + s_L^2 \tag{14.2}$$

Equations (14.1) and (14.2) state that, as noted in Chapter 13, laboratory components of biases are systematic errors from the point of view of a single laboratory, but random errors from the interlaboratory point of view.

It should be noted here that the Guide to the Expression of Uncertainty in Measurements [6] defines reproducibility as "the closeness of agreement between the results of measurements of the same measurand, where the measurements are

carried out under changed conditions". The Guide continues to specify that the changed conditions can include different principles or methods of measurement, different observers, different measuring instruments, different locations, different conditions of use or different periods of time. Reproducibility is then no longer linked to a specific method and a note specifies that "a valid statement of reproducibility requires specification of the conditions changed". In the context of interlaboratory studies as described in this chapter and in the ISO, IUPAC and AOAC norms, reproducibility is a precision measurement for a specific method.

14.2.2 Method-performance precision experiments

A precision experiment is carried out by $p \geq 8$ laboratories. They analyze the materials with the method, the performance of which is studied, at m levels of concentration. Indeed, precision can (and usually does) depend on concentration and when a method is validated this has to be done over the whole range concerned (see Chapter 13.2). Each level is analyzed n times. The recommended value for n is 2. The number of levels m depends on the concentration range. ISO [1] recommends $m \geq 6$ and IUPAC [2] at least 5 materials, where a material is a specific combination of matrix and level of analyte. However, there are cases where this makes little sense. For instance, when one validates a method for the analysis of a drug in formulations, its concentration is probably situated in the relatively small range at which it is pharmacologically active but not toxic so that it is improbable that drugs will be formulated at vastly different concentrations. A smaller m can then be accepted. The number of laboratories should be at least 8, but when only one single level of concentration is of interest the ISO standard recommends to include at least 15 laboratories. It should be noted that the numbers of laboratories given here, are those cited in the standard. They are needed if one wants to develop an internationally accepted reference method. This does not mean that method-performance studies with smaller numbers of participants are not useful.

The experiment can be run as a *uniform level experiment* or a *split level experiment*. In a uniform level experiment one carries out $n = 2$ replicate determinations on the same material. This can in some cases lead to an underestimation of the precision measure. If operators know they are analyzing replicates, they may be influenced and produce results that are as alike as possible. This may be avoided with split level experiments, in which a material with slightly different concentrations is used. Each of the levels is analyzed once. In Section 14.2.4 we will describe how to compute the results for uniform level experiments. How to carry out similar computations for split level experiments is described in [1] and [3]. It should be stressed also that this chapter describes what is of interest to the chemometrician, namely the statistical analysis of the data. Readers who want to study in detail how to carry out method performance experiments should be aware that the organizational

aspect is very important. This is described in Refs. [1] and [3]. Tutorial articles such as Ref. [7] are also useful to understand better how to set up such experiments.

14.2.3 Repeatability and reproducibility in a method-performance experiment

The ISO standard [1,3] describes two measures for the repeatability and the reproducibility. The repeatability can be described by the *repeatability standard deviation* s_r, which is the standard deviation measured under repeatability conditions (see Section 13.4.1). How to obtain this value is described in Section 14.2.4. It estimates, as usual, the true repeatability standard deviation σ_r. The *relative repeatability standard deviation* is written as RSD_r. Another measure is the *repeatability (limit)*, r. This is the value below which the absolute difference between two single test results obtained under repeatability conditions may be expected to lie with a probability of 95%. The standard states that

$$r = 2.8\ \sigma_r \tag{14.3}$$

In the same way one can define the *reproducibility standard deviation* s_R, the *reproducibility limit* R and the *relative reproducibility standard deviation*, RSD_R, and write

$$R = 2.8\ \sigma_r \tag{14.4}$$

The value of 2.8 can be understood as follows. The variance of the difference between two replicate measurements is $2\sigma^2$ (σ being estimated by s_r or s_R, according to the situation). The confidence interval at 95% level on the difference is $0 \pm 1.96\ \sqrt{2}\sigma$. If s_r or s_R estimate σ well enough, as should be the case in an interlaboratory experiment, then one can write that the confidence interval is $\pm 1.96\ \sqrt{2}\ s_r$ (or s_R) rounded to $\pm 2.8\ s_r$ (or s_R). There is therefore a 95% probability that the difference between two individual determinations will not exceed $2.8\ s_r$ (or s_R).

The repeatability and reproducibility values can be used to include precision clauses in the description of the method. A typical precision clause is: "The absolute difference between two single test results obtained under reproducibility conditions should not be greater than 0.7 mg/l".

An experimental difference between two values larger than r, therefore indicates that the laboratory's repeatability for the method investigated is not up to standard. Part 6 of the ISO standard [3] describes a procedure on how to judge on the acceptability of results when the repeatability limit is exceeded.

14.2.4 Statistical analysis of the data obtained in a method-performance experiment

The analysis of the data starts by investigating the statistical assumptions. The main assumptions are:

1. Normal distribution of the laboratory means at each level.
2. Equality of variance among laboratories at each level.

Since outliers lead to non-normal distributions, ISO recommends to investigate the existence of outliers to test the first hypothesis. A violation would indicate unacceptable laboratory bias in the outlying laboratories. The second assumption is investigated by testing the ranges (if $n = 2$) or variances (if $n > 2$) for outliers. Violation would indicate unacceptable differences in repeatability among laboratories.

The standards describe a procedure for outlier removal. The flowchart of the outlier removal procedure is shown in Fig. 14.1. One first tests the homogeneity of the variance in the laboratories with the use of the Cochran test (see Chapter 6.2).

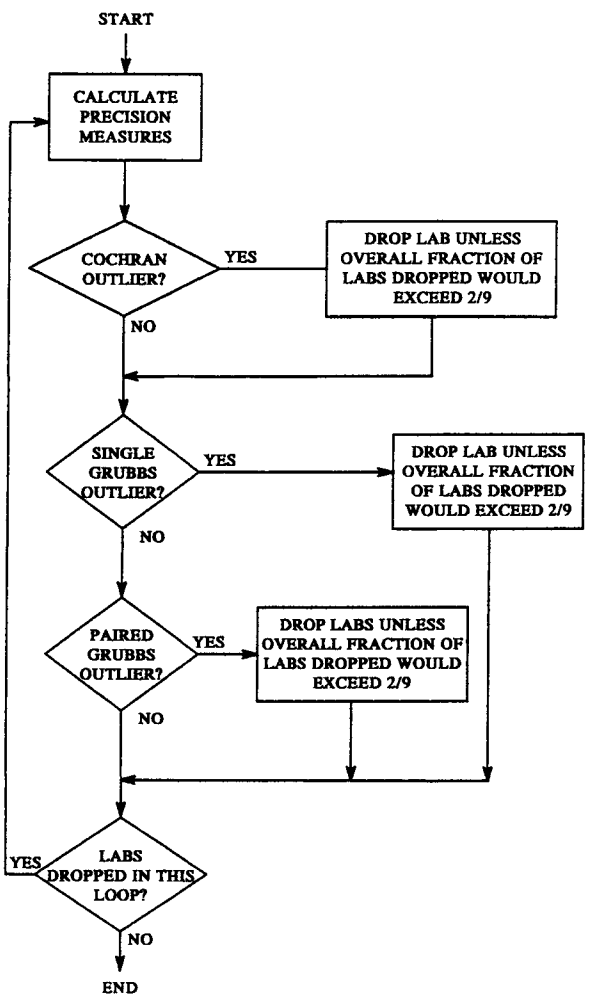


Fig. 14.1. Flowchart for outlier removal from a precision experiment (adapted from Ref. [7]).

The averages for the same level presented by the p laboratories are then tested for outliers by the Grubbs' test (see Chapter 5), first with the 2-sided single outlier test, then with the Grubbs' pair test (double outlier test in the ISO terminology). ISO uses a version of the latter test which detects two simultaneous high or two simultaneous low values, while the AOAC/IUPAC version also detects one high and one low value occurring simultaneously. In all tests, outliers at the 1% level are eliminated (at the 2.5% level in the AOAC/IUPAC protocol [2]), while stragglers at the 5% level are flagged, but are included in the analysis except when an assignable cause for the outlying result is found. A graphical way to evaluate the consistency of results and laboratories, and recommended in [3], is what are called Mandel's h and k statistics. These statistics can also be used to evaluate the quality of laboratories and we describe these methods in Section 14.3 on laboratory-performance studies.

An example of the analysis of a uniform level experiment is given in Table 14.1. The example concerns a standard test method, involving a thermometric titration [8]. Nine laboratories participated, five materials were analyzed in duplicate. The first step is the outlier removal procedure. It is carried out using the data of Tables 14.2 and 14.3. The data of Table 14.2 are subjected to the Cochran test. At level 4, the test value for 1.10 (laboratory 7) is 0.667; at level 5 the test value for 1.98 (laboratory 6) is 0.636. The critical values for $n = 9$ are 0.754 ($\alpha = 1\%$) and 0.638 ($\alpha = 5\%$) (see Tables in Chapter 5). The value 1.10 is a straggler and 1.98 is so close to it that it is also flagged. However, none of these values is eliminated yet. The single Grubbs' test indicates that there are outliers at levels 3 and 4 of laboratory 1 (Table 14.4). The two values are eliminated. The other values obtained by laboratory 1 are not detected as outliers, but they too are high and therefore it was decided to reject all values of laboratory 1. The Cochran test is now applied

TABLE 14.1

A uniform level experiment (from [8]). Data are given in % mass/mass

Lab. i	Level j									
	1		2		3		4		5	
1	4.44	4.39	9.34	9.34	17.40	16.90	19.23	19.23	24.28	24.00
2	4.03	4.23	8.42	8.33	14.42	14.50	16.06	16.22	20.40	19.91
3	3.70	3.70	7.60	7.40	13.60	13.60	14.50	15.10	19.30	19.70
4	4.10	4.10	8.93	8.80	14.60	14.20	15.60	15.50	20.30	20.30
5	3.97	4.04	7.89	8.12	13.73	13.92	15.54	15.78	20.53	20.88
6	3.75	4.03	8.76	9.24	13.90	14.06	16.42	16.58	18.56	16.58
7	3.70	3.80	8.00	8.30	14.10	14.20	14.90	16.00	19.70	20.50
8	3.91	3.90	8.04	8.07	14.84	14.84	15.41	15.22	21.10	20.78
9	4.02	4.07	8.44	8.17	14.24	14.10	15.14	15.44	20.71	21.66

TABLE 14.2

Cell ranges for the data of Table 14.1

Laboratory i	Level j				
	1	2	3	4	5
1	0.05	0.00	0.50	0.00	0.28
2	0.20	0.09	0.08	0.16	0.49
3	0.00	0.20	0.00	0.60	0.40
4	0.00	0.13	0.40	0.10	0.00
5	0.07	0.23	0.19	0.24	0.35
6	0.28	0.48	0.16	0.16	1.98
7	0.10	0.30	0.10	1.10	0.80
8	0.01	0.03	0.00	0.19	0.32
9	0.05	0.27	0.14	0.30	0.95

TABLE 14.3

Cell averages of the data of Table 14.1

Laboratory i	Level j				
	1	2	3	4	5
1	4.415	9.340	17.150	19.230	24.140
2	4.130	8.375	14.460	16.140	20.155
3	3.700	7.500	13.600	14.800	19.500
4	4.100	8.865	14.400	15.550	20.300
5	4.005	8.005	13.825	15.660	20.705
6	3.890	9.000	13.980	16.500	17.570
7	3.750	8.150	14.150	15.450	20.100
8	3.905	8.055	14.840	15.315	20.940
9	4.045	8.305	14.170	15.290	21.185

TABLE 14.4

Application of Grubbs' test to the cell averages of Table 14.3

Level	Single low	Single high	Double low	Double high
1	1.36	1.95	0.502	0.356
2	1.57	1.64	0.540	0.395
3	0.86	2.50	–	–
4	0.91	2.47	–	–
5	1.70	2.10	0.501	0.318
$\alpha = 1\%$	2.215	2.215	0.149	0.149
$\alpha = 5\%$	2.387	2.387	0.085	0.085

again. The critical value for $n = 8$ being 0.680 at $\alpha = 5\%$, none of the two previously identified stragglers can be considered an outlier. However, the value 16.58 of level 5 from laboratory 6 might by mistake have come from level 4 and this result is therefore considered to have an assignable cause and eliminated. The result at level 4 for laboratory 7 is however accepted. It should be noted that outlier rejection is not applied here (and should never be applied) as an automatic procedure, but rather that the statistical conclusions are only one aspect of the whole context leading to the final decision.

After elimination of the outliers, one can determine s_r and s_R by one-way analysis of variance at each concentration level (and, by multiplication with the factor 2.8, r and R can also be obtained). Simple hand calculations are also possible and are carried out as follows for a uniform level experiment. One uses the equation for paired results, eq. (2.8). This yields:

$$s_r^2 = \frac{1}{2p} \sum d_i^2 \quad (i = 1, ..., p)$$

$$s_L^2 = \left[\frac{1}{p-1} \sum 2(\bar{y}_i - \bar{y})^2 - s_r^2 \right] / 2$$

$$= \frac{1}{p-1} [\sum (\bar{y}_i - \bar{y})^2] - s_r^2 / 2$$

where d_i is the difference and $\bar{y}_i$ the mean of the two results obtained by laboratory i and $\bar{y}$ the grand mean, i.e. $\sum \bar{y}_i / p$. When s_r^2 and s_L^2 have been computed, s_R^2 can be obtained from eq. (14.2).

An example of the calculations for one of the m levels is given in Table 14.5. The calculations are given only for level 5. The results for all the levels are summarized in Table 14.6.

A last step is to investigate whether there is a relationship between the s_{rj} (or s_{Rj}) and the concentration $\bar{y}_j$ obtained at each level j. Indeed, it is known that precision measures can depend on concentration (see Section 13.4.2). The ISO document [3] recommends the following models be tested:

$$s_r = b_1 \bar{y}$$

$$s_r = b'_0 + b'_1 \bar{y} \qquad (14.5)$$

$$\log s_r = b''_0 + b''_1 \log \bar{y}$$

Similar models are developed for s_R. The simplest of the models that is found to fit the data sufficiently well is adopted for further use. Weighted regression is applied and a procedure to decide on the weights is described in the norm. For s_r the results are:

TABLE 14.5

Example of calculation of r and R (adapted from ISO [1]) (level 5 of Table 14.1)

Laboratory	$\bar{y}_i$	d_i	d_i^2	$\bar{y}_i - \bar{y}$	$(\bar{y}_i - \bar{y})^2$
2	20.155	0.49	0.240	−0.255	0.0650
3	19.500	0.40	0.160	−0.910	0.8281
4	20.300	0.00	0.000	−0.110	0.0012
5	20.705	0.35	0.122	0.295	0.0870
7	20.100	0.80	0.640	−0.310	0.0961
8	20.940	0.32	0.102	0.530	0.2809
9	21.185	0.95	0.903	0.775	0.6006
Σ	142.885		2.167		1.9583

$\bar{\bar{y}} = 20.41$

$s_r^2 = 2.167/14 = 0.1548$ $\quad s_r = 0.393$ $\quad r = 2.8 \times 0.393 = 1.102$

$s_L^2 = 1.9583/6 - 0.1548/2 = 0.2490$

$s_R^2 = 0.1548 + 0.2490 = 0.4038$ $\quad s_R = 0.6358$ $\quad R = 2.8 \times 0.6358 = 1.783$

TABLE 14.6

Repeatability (r) and reproducibility (R) values for the data of Table 14.1

Level	$\bar{\bar{y}}$	r	R
1	3.94	0.258	0.478
2	8.28	0.501	1.393
3	14.18	0.355	1.121
4	15.59	0.943	1.620
5	20.41	1.102	1.783

$$s_r = 0.019\,\bar{\bar{y}}$$

$$s_r = 0.032 + 0.015\,\bar{\bar{y}}$$

$$s_r = 0.031\,\bar{\bar{y}}^{0.77}$$

No formal procedures have been described to decide which of the equations fits best. In this case, it is decided that the simplest equation is good enough, so that this is the one that is finally adopted. In the present case, the first equation is adopted.

14.2.5 What precision to expect

One of the questions that could be asked in a method-performance experiment is what values of s_r and s_R should be expected. Interesting work in this context has been carried out by Horwitz et al. [9]. They examined results of many interlaboratory

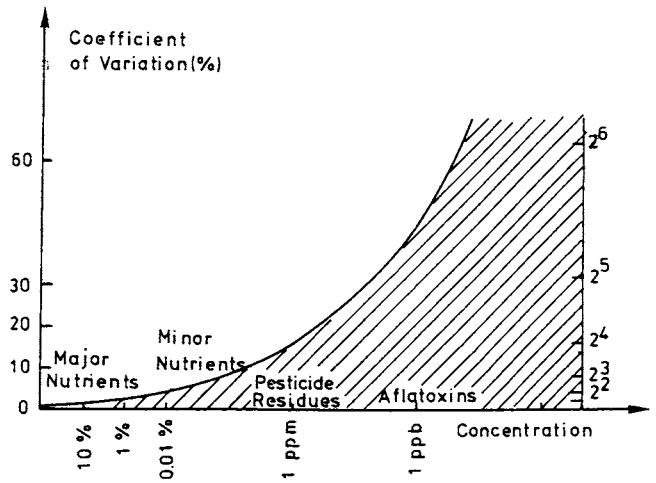


Fig. 14.2. Relative reproducibility standard deviation RSD_R as a function of concentration (adapted from Ref. [9]).

collaborative studies on various commodities ranging in concentration from a few percent (salt in foods) to the ppb (ng/g) level (aflatoxin M1 in foods) but including also studies on, for example, drug formulations, antibiotics in feeds, pesticide residues and trace elements. They concluded that the relative reproducibility standard deviation RSD_R (%) as a function of concentration is approximated by the following relationship (see also Fig. 14.2):

$$RSD_R\% = 2^{(1-0.5 \log_{10} C)} \tag{14.6}$$

where C is the concentration expressed as a decimal fraction (for example for a concentration of 1 μg/g, $C = 10^{-6}$ g/g). This equation states that RSD_R approximately doubles for every 100-fold decrease in concentration, starting at 2% for $C = 1$ (or $\log C = 0$). This means, for instance, that when one carries out a purity check by analysing the main component, one should count with a relative reproducibility standard deviation of 2%. This leads to Table 14.7. It should be noted that these results have been obtained under optimal conditions. Laboratories participating in a between-lab study want to be able to show good results and probably exercise somewhat more care than would be the case in normal routine.

One of the notable conclusions of Horwitz' study is that the RSD_R depends only on the concentration and not on the analyte or method. This is true to such an extent that in a later publication [10] the author states that RSD_R-values are suspect, when they exceed by more than a factor 2 what is expected from eq. (14.6). The ratio between the reproducibility obtained and the one expected from eq. 14.6 is sometimes called the HORRAT (short for Horwitz ratio).

TABLE 14.7

Relative reproducibility standard deviation for some concentrations (in %) (from Ref. [9])

Type (Concentration)	Fractional	RSD_R
Pure substances (100%)	1	2
Drug formulations (1%)	0.01	4
Drug in feeds (0.01%)	0.0001	8
Trace elements (μg/g)	0.000001	16
Aflatoxins (10 ng/g)	10^{-8}	32

Another interesting result is that the corresponding repeatability measure (RSD_r) is generally one-half to two-thirds of the reproducibility. A similar result was obtained in clinical chemistry by Steele et al. [11]. From eq. (14.2), it follows that s_L^2 is about 0.5 to 0.75 of s_R^2.

14.2.6 Method-performance bias experiments

Until recently, method-performance experiments were synonymous with precision experiments. Part 4 of the ISO standard [3] describes methods for estimating the bias of a measurement method and the laboratory bias, when applying that method. It is restricted to the study of standard methods and requires that a reference value can be established, for example by measurement of reference materials. It also considers only situations where the measurement is carried out at a single level and where no interferences are present.

14.3 Laboratory-performance studies

14.3.1 Visual display methods

Table 14.8 gives a summary of the data of an experiment for the proficiency of laboratories in analyzing clenbuterol in urine [12]. This is a β-blocker illegally used as growth promoter for cattle. The set-up was the following. The 10 participating laboratories, identified by a code number in Table 14.8, used their own method to analyze (in 10-replicate) 3 samples, namely a spiked sample containing the known amount of 1.5 ng/ml (mean result $\bar{y}_1$), a real sample with a concentration close enough to 1.5 ng/ml not to be easily distinguishable from the first sample (mean result $\bar{y}_2$), a second real sample with higher content (mean result $\bar{y}_3$). The samples were randomly coded so that the two first samples in particular could not be distinguished by the participants. Three visual display methods are discussed in this section. A fourth, based on principal components, is found in Section 17.5.2.

TABLE 14.8

Performance assessment of laboratories by ranking. The data are for clenbuterol in urine [12]

Lab	Results (in ng/ml)			Ranks			Score
	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	R_1	R_2	R_3	
11	1.15	1.33	3.13	8	8	8	24
12	1.49	1.76	3.32	7	4	7	18
20	1.64	1.70	3.40	4	5.5	5	14.5
30	1.51	1.70	3.38	5	5.5	6	16.5
40	1.67	1.98	4.42	3	2	1	6
50	1.07	0.74	2.39	9	10	9	28
60	1.50	1.82	4.02	6	3	2	11
70	0.69	0.85	1.32	10	9	10	29
81	1.68	1.54	3.67	2	7	4	13
82	1.91	2.01	4.01	1	1	3	5

14.3.1.1 Box plots

In all cases where this is relevant one should first evaluate the results visually. The box plot (see Chapter 12) is a very useful way to do that. Figure 14.3 gives the results for the first two samples. The two figures show, for instance, that laboratory 70 delivers clearly lower results than the others and so, to a lesser extent, do 50 and 11. Laboratory 60 is less repeatable than the others. Comparison with the known content of 1.5 immediately shows that laboratories 11, 50, 70, 82 do not deliver the correct result. Others, such as 40, are probably significantly biased but the difference is not large enough to be considered important.

14.3.1.2 Youden plots

The information in Fig. 14.3 can be looked at in another way, namely by making a *Youden plot* [13]. This consists of plotting the results of two samples with similar concentration, for instance in a split level experiment, against each other for each laboratory. Such sets of samples are sometimes called *Youden pairs*. The two first samples of Table 14.8 have slightly different concentrations and their Youden plot is shown in Fig. 14.4.

In the original publication [13], the origin of the plot was situated at the median values of both samples, in later publications one uses the averages. Through this origin one draws lines parallel to the x- and the y-axis, thereby dividing the graph into four quadrants. If the origin of the graph is accepted as the most probably true value for both samples, then laboratories situated in the upper right corner have a positive bias on both samples. Laboratories in the lower left corner have a negative bias on both samples and the two other quadrants are high for sample 1 and low for

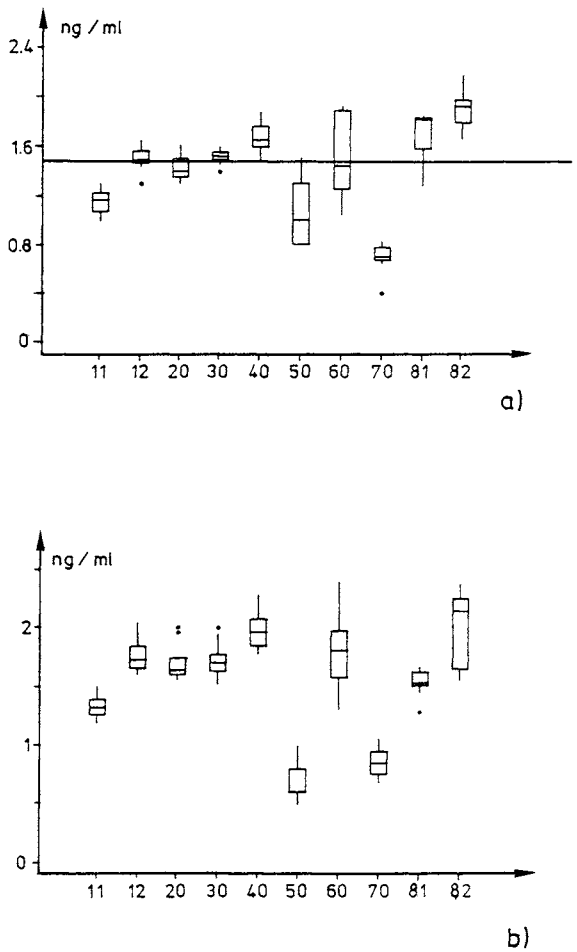


Fig. 14.3. Laboratory-performance study for the analysis of clenbuterol [12]. The numbers 11–82 are the codes of the laboratories. (a) Blank urine spiked with 1.5 ng/ml. (b) Real sample.

sample 2, respectively low for sample 1 and high for sample 2. If only random errors occurred, one would expect the points to be more or less equally distributed over all quadrants. However, the situation of Fig. 14.4 occurs more frequently: there are more points in the upper right and lower left quadrant, indicating that most laboratories have either consistently too high or too low results, in other words, there is a systematic error. When all laboratories use the same method, these systematic errors are laboratory biases. In our example the laboratories all use different methods, so that it is not possible to conclude whether the systematic error is due to the methods or to the laboratories.

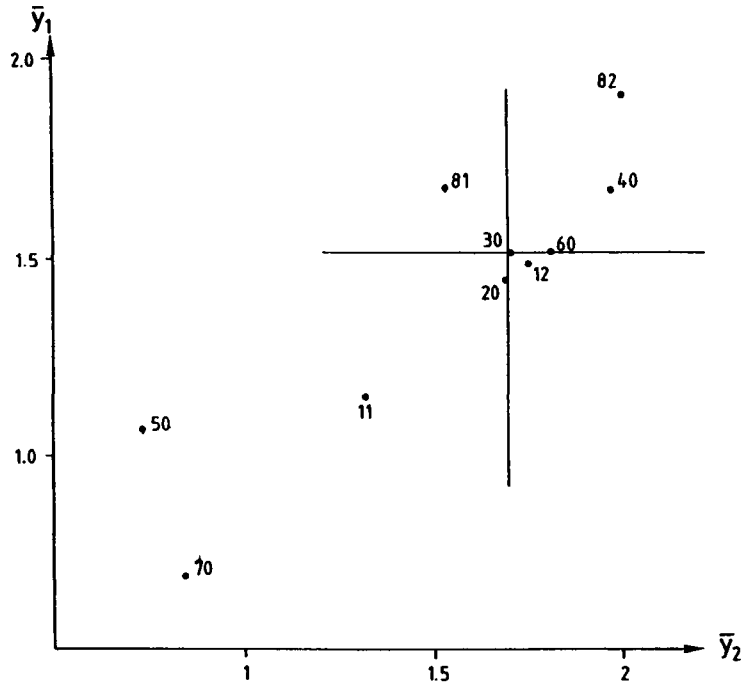


Fig. 14.4. Youden plot. Average concentrations found for sample (a) and (b) by the laboratories of Fig. 14.3 plotted one against the other.

A more general treatment was given by Mandel and Lashof [14]. Youden's article assumed that the concentration of the analyte in the two materials was nearly the same, so that the repeatability as well as the laboratory biases would be the same for two materials. Mandel and Lashof investigate the situation where the two samples do not have a similar concentration so that random and systematic errors are no longer necessarily the same for both methods. They showed that in all cases the points in the plot fall within an elongated ellipse. When Youden's assumptions are obeyed, then the major axis makes a 45° angle, but when these assumptions are found to be incorrect other angles may be obtained. Their paper contains a procedure to decide whether lab bias occurs or not and to estimate all variance components.

14.3.1.3 Mandel's h and k consistency statistics

Mandel's h and k consistency statistics are a graphical way of describing the variability in the set of data and to look for inconsistencies [3]. The h-statistic essentially studies the variability of the mean results obtained at each level and the k-statistic compares standard deviations of the laboratories.

The equations are the following:

$$h_{ij} = \frac{\bar{y}_{ij} - \bar{y}_j}{\sqrt{\frac{1}{p_j - 1} \sum_{i=1}^{p_j} (\bar{y}_{ij} - \bar{y}_j)^2}} \tag{14.7}$$

where j is the jth level and the other symbols have the same meaning as in Section 14.2.4.

$$k_{ij} = s_{ij}/S_j \text{ with} \tag{14.8}$$

$$S_j = \sqrt{\sum_{i=1}^{p_j} s_{ij}^2 / p_j}$$

The ISO standard also gives indicator values at the 1 and 5% levels. $\bar{y}_{ij} - \bar{y}_j$ is the deviation of the cell average $\bar{y}_{ij}$ of laboratory i at level j from $\bar{y}_j$, the general average at level j and the h-ratio therefore compares the deviation for a laboratory from the general average at that level with the standard deviation of the mean values obtained by all p_j laboratories that have reported results at level j. A plot for the data of Table 14.1 is shown in Fig. 14.5 and the clenbuterol data are found in Fig. 14.6. In Fig. 14.5, one observes that four of the five values of laboratory 1 are larger than the 5% indicator values. Moreover, all values for that laboratory are rather high. This confirms that the results of laboratory 1 are inconsistent with those of the others. No other inconsistencies are observed. For the clenbuterol data of Fig. 14.6, it is observed that laboratory 70 yields too low results. One of the results of laboratory 50 also exceeds the 5% indicator value, but we probably would not decide to eliminate 50 from further consideration.

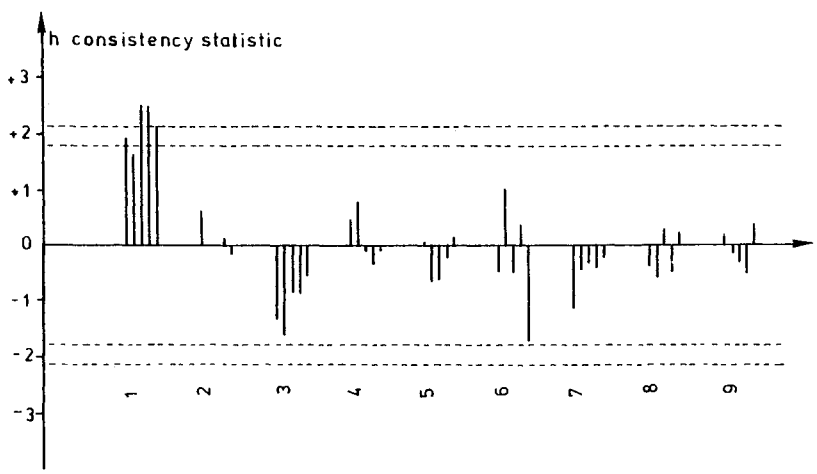


Fig. 14.5. Mandel's h-statistic for the data of Table 14.1.

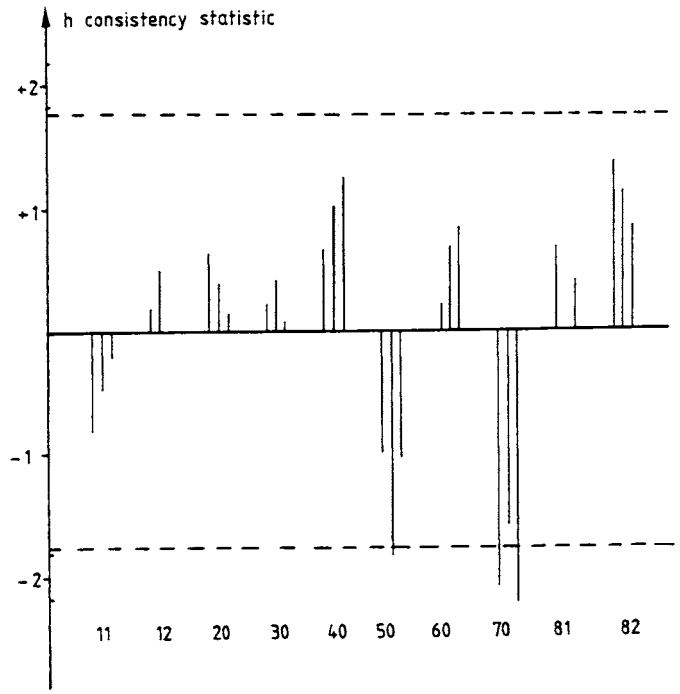


Fig. 14.6. Mandel's *h*-statistic for the data of Table 14.8.

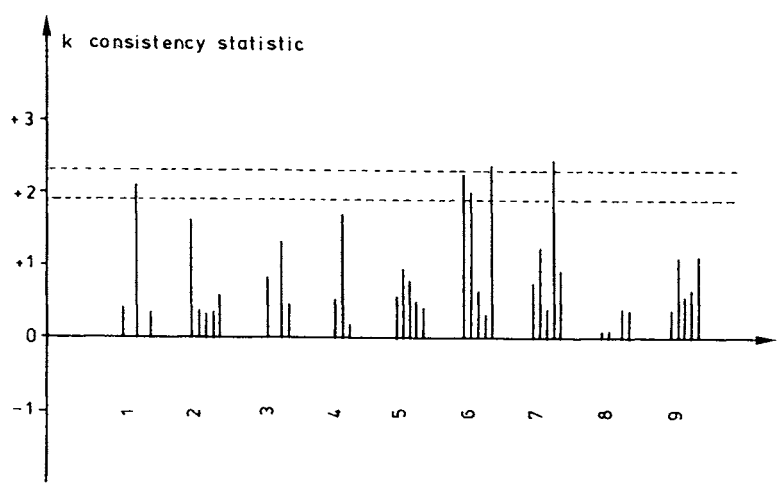


Fig. 14.7. Mandel's *k*-statistic for the data of Table 14.1.

The k-ratio compares the standard deviation s_{ij} of laboratory i at level j with that of all laboratories at that level, S_j. The plot for the data of Table 14.1 is given in Fig. 14.7. It shows that two values exceed the 1% level and also that, in general, laboratory 6 seems to deliver less repeatable results. As explained in Section 14.2.4, it was found that the result at level 5 should be eliminated.

14.3.2 Ranking method

Some methods for evaluating laboratory performance make use of scores calculated for the laboratories. In the ranking method [15] these scores are based on a ranking of the results obtained by the laboratories. Table 14.8 illustrates the procedure for the clenbuterol data introduced in Section 14.3.1. For each of the three urine samples the highest result is given rank 1, the one but highest is given rank 2 and so on. When ties are present the mean of the ranks is computed. The ranked results for the three urine samples are given in the table under R_1, R_2 and R_3. The laboratory score is obtained as the sum of the ranks the laboratory received. For our example the lowest possible score is 3 while the highest possible score is 30. The former will be obtained by a laboratory that systematically reports the highest results for all materials while the latter will be the score of a laboratory that systematically reports the lowest results for all materials. Therefore extreme scores are an indication for the presence of systematical errors. Table 14.9 gives, for different combinations of m (number of materials analyzed) and p (number of laboratories involved in the study), upper and lower limits for the scores. For each combination of p and m two critical values are listed. A calculated score which is less than or equal to the smaller critical value or greater than or equal to the larger critical value results in the rejection of the hypothesis of a random ranking at the 5% significance level. For our example ($p = 10$ and $m = 3$) random ranking would result in a score not less than 5 and not more than 28. The score 29 for laboratory 70 therefore is considered extreme and points to a systematic error resulting here in low results. The performance of laboratories 50 and 82, respectively with scores 28 and 5 at the border of the critical region, might also be questioned.

14.3.3 The z-score method

Scores have the advantage that they constitute a simple way to compare laboratories among each other or the present performance of a laboratory with its previous performance and that they can be used as a formal way of eliminating laboratories, that do not perform sufficiently well, from accreditation. Such a score, known as the z-score, was described by the Analytical Methods Committee [5] and IUPAC [4] and will be explained here.

The z-score of an individual laboratory is obtained as follows:

TABLE 14.9

Critical values of ranking scores [15]

Number of labs. (p)	Number of materials (m)												
	3	4	5	6	7	8	9	10	11	12	13	14	15
3		4	5	7	8	10	12	13	15	17	19	20	22
		12	15	17	20	22	24	27	29	31	33	36	38
4		4	6	8	10	12	14	16	18	20	22	24	26
		16	19	22	25	28	31	34	37	40	43	46	49
5		5	7	9	11	13	16	18	21	23	26	28	31
		19	23	27	31	35	38	42	45	49	52	56	59
6	3	5	7	10	12	15	18	21	23	26	29	32	35
	18	23	28	32	37	41	45	49	54	58	62	66	70
7	3	5	8	11	14	17	20	23	26	29	32	36	39
	21	27	32	37	42	47	52	57	62	67	72	76	81
8	3	6	9	12	15	18	22	25	29	32	36	39	43
	24	30	36	42	48	54	59	65	70	76	81	87	92
9	3	6	9	13	16	20	24	27	31	35	39	43	47
	27	34	41	47	54	60	66	73	79	85	91	97	103
10	4	7	10	14	17	21	26	30	34	38	43	47	51
	29	37	45	52	60	67	73	80	87	94	100	107	114
11	4	7	11	15	19	23	27	32	36	41	46	51	55
	32	41	49	57	65	73	81	88	96	103	110	117	125
12	4	7	11	15	20	24	29	34	39	44	49	54	59
	35	45	54	63	71	80	88	96	104	112	120	128	136
13	4	8	12	16	21	26	31	36	42	47	52	58	63
	38	48	58	68	77	86	95	104	112	121	130	138	147
14	4	8	12	17	22	27	33	38	44	50	56	61	67
	41	52	63	73	83	93	102	112	121	130	139	149	158
15	4	8	13	18	23	29	35	41	47	53	59	65	71
	44	56	67	78	89	99	109	119	129	139	149	159	169

$$z = (x - X)/S \tag{14.9}$$

where x is the result obtained by a laboratory, X the true concentration of the analyte or its best estimate $\hat{X}$ and S some kind of standard deviation. The difficulty with using eq. (14.9) is the determination of X and S.

The best way of determining X is to add a known amount of analyte to the base material. Unfortunately, there are several chemical reasons why this often is not possible, nor recommended. Among them there is the difficulty of adding homogeneously the analyte to many types of base material, the speciation which may be different and the problem of obtaining such materials, free of analyte. When it is not possible to determine X in this way, one needs consensus values. These can be obtained in two ways, namely by a group of expert laboratories using best possible methods or by the participants themselves. The former is more costly, but has the advantage to provide an external performance standard with which to measure the proficiency of other laboratories. The latter is cheaper, since the consensus is not determined in a special round of experiments, but in the actual proficiency testing round. A further question is how to compute $\hat{X}$. One possibility is to obtain the mean of the participating laboratories, after elimination of outliers. Another is to obtain a robust mean such as the median, the biweight or winsorized mean described in Chapter 12.

S can be obtained in a proficiency testing round as the standard deviation of the laboratories' results after the elimination of outliers. However when possible it is to be preferred that S should be a target value of precision. This could be derived from the precision required to perform a certain task, from method performance studies (Section 14.2.3) or from the Horwitz curve (Section 14.2.5).

With good estimates of $\hat{X}$ and S, z corresponds to a standardized variable. Consequently one expects that $|z| > 2$ in only 4.55% of the cases and $|z| > 3$ in only 0.27%. The latter indicates unacceptably poor performance in terms of accuracy while for a satisfactory performance $|z| \leq 2$ is required.

When, as is usual, more than one test material is being analyzed a composite score over all the test materials might be required. For this purpose one can use some type of sum of scores. The Analytical Methods Committee and IUPAC prefer the sum of squared scores, SSZ

$$\text{SSZ} = \sum z^2$$

This follows a χ^2 distribution with m degrees of freedom, where m is the number of scores that is combined into SSZ. An interpretation for the SSZ similar to that for the z-scores requires the use of Table 14.10. A SSZ $\leq$ A is satisfactory, if A < SSZ < B the performance is questionable and for SSZ $\geq$ B it is unsatisfactory. A and B are respectively the 4.55% and 0.27% points of the χ^2 distribution which corresponds to the two-sided z-values 2 and 3 used in the interpretation of the z-score. The above-mentioned organizations do not recommend the general use of combination scores because a significant single score can systematically be masked in this way.

TABLE 14.10

Classification of SSZ scores [5]

m	A	B	m	A	B
2	6.18	11.83	11	19.99	28.51
3	8.02	14.16	12	21.35	30.10
4	9.72	16.25	13	22.70	31.66
5	11.31	18.21	14	24.03	33.20
6	12.85	20.06	15	25.35	34.71
7	14.34	21.85	16	26.66	36.22
8	15.79	23.57	17	27.96	37.70
9	17.21	25.26	18	29.25	39.17
10	18.61	26.90	19	30.53	40.63

References

1. ISO Standard 5725, 1986 (E), Precision of test methods — Determination of repeatability and reproducibility for a standard test method by inter-laboratory tests.
2. AOAC/IUPAC revised protocol for the design, conduct and interpretation of method-performance studies. JAOAC Int., 78 (1995) 143A–160A.
3. ISO Standard 5725, 1994. Accuracy (trueness and precision) of measurement methods and results.
4. IUPAC, The international harmonized protocol for the proficiency testing of (chemical) analytical laboratories. Pure Appl. Chem., 65 (1983) 2123–2144.
5. Analytical Methods Committee, Proficiency testing of analytical laboratories: organization and statistical assessment. Analyst, 117 (1992) 97–117.
6. BIPM/IEC/IFCC/ISO/IUPAC/IUPAP/OIML, Guide to the Expression of Uncertainty in Measurements and Symbols, 1993.
7. W. Horwitz, Harmonized protocol for the design and interpretation of collaborative studies. Trends Anal. Chem., 7 (1988) 118–120.
8. Standard methods for testing tar and its products, as cited in ref. [3].
9. W. Horwitz, L.R. Kamps and K.W. Boyer, J. Assoc. Off. Anal. Chem., 63 (1980) 1344–1354.
10. K.W. Boyer, W. Horwitz and R. Albert, Anal. Chem., 57 (1985) 454–459.
11. B.W. Steele, M.K. Schauble, J.M. Becktel and J.E. Bearman, Am. J. Clin. Pathol., 67 (1977) 594-602.
12. H. Beernaert, IHE-report, Ringtest Clenbuterol in Urine. September 1992.
13. W.J. Youden, Graphical diagnosis of interlaboratory test results. Indust. Qual. Control, May (1959), 24–27.
14. J. Mandel and T.W. Lashof, Interpretation and generalization of Youden's two-sample diagram. J. Qual. Technol., 6 (1974) 22–36.
15. G.T. Wernimont, Use of statistics to develop and evaluate analytical methods. AOAC, Arlington, VA, 1987.

Additional literature

M. Feinberg, Basics of interlaboratory studies: the trends in the new ISO 5725 standard edition. TrAC, 14 (1995) 450–457.

Chapter 15

Other Distributions

15.1 Introduction — Probabilities

In Chapter 3 we described the normal distribution and stated that it is the best known probability distribution. There are several other probability distributions. In certain cases the data are not continuous, but discrete and we need other distributions such as the binomial, the hypergeometric and Poisson distributions to describe these data. Other probability functions have been defined for specific purposes. Examples of some of the more important special distributions are described in Sections 15.5 and 15.6. Hypothesis tests based on these distributions have been developed, but will not be described here.

Since we will have to apply probability calculus in this and several of the next chapters, some concepts and axioms must be defined here.

(1) The *probability* of an event X is a non-negative number smaller than or equal to 1

$$0 \leq P[X] \leq 1 \tag{15.1}$$

(2) The sum of the probabilities of all possible events X_i for a given situation is equal to 1

$$\sum P[X_i] = 1 \tag{15.2}$$

When one carries out quality control, some objects will be found to have no defects, some a single defect and some two defects. Supposing that objects with three or more defects do not occur, one can write:

$$P[\text{no defect}] + P[1 \text{ defect}] + P[2 \text{ defects}] = 1$$

(3) X and Y are said to be *mutually exclusive* if no event belongs to both X and Y. In this case

$$P[X \cup Y] = P[X] + P[Y] \tag{15.3}$$

$X \cup Y$ (X or Y) is called the *union* of X and Y. It is the event which consists of all the simple events belonging to X or Y or both X and Y. It can also be written as event X + Y.

Suppose that an item can have either defect A or B, but not both, then defects A and B are mutually exclusive. If, of 100 items investigated, 10 are found to have defect A and 10 defect B, then

$P[A \cup B] = P[\text{defect}] = P[A] + P[B] = 0.2$

(4) When an event can belong to both X and Y, then X and Y are not mutually exclusive and the probability of the union is then given by

$$P[X \cup Y] = P[X] + P[Y] - P[X \cap Y] \tag{15.4}$$

$X \cap Y$ (X and Y) is the *intersection* of two events X and Y. It is the event consisting of all events belonging to both X and Y and is sometimes written as X.Y or XY.

Suppose now that A and B are not mutually exclusive i.e. an object can have only defect A, only defect B or both defects. Then $P[\text{both defects}] = P[A \cap B]$.

If $P[\text{only A}] = 0.1$, $P[\text{only B}] = 0.1$, $P[\text{both defects}] = 0.05$, then $P[A] = 0.15$, $P[B] = 0.15$ and $P[\text{defect}] = P[A] + P[B] - P[A \cap B] = 0.15 + 0.15 - 0.05 = 0.25$. Equation (15.4) is known as the *addition law*.

(5) Two events X and Y are said to be *independent* if and only if

$$P[X \cap Y] = P[X]\,P[Y] \tag{15.5}$$

A company produces two types of items, blue-coloured ones and red-coloured ones. Defects are not related to colour. The amount of blue-coloured objects produced is one-quarter of the total production ($P[\text{blue}] = 0.25$). Suppose $P[\text{defect}] = 0.1$. The occurrence of blue objects with a defect in the total population of objects is

$P\,[\text{blue} \cap \text{defect}] = 0.025$

Two events are *complementary* if

$$P[X] + P[Y] = 1 \tag{15.6}$$

An object can either present no defect or at least one (object defective). Both events are complementary:

$P[\text{defective}] + P[\text{no defect}] = 1$

In that case all events that do not belong to X can be written as not-X ($\tilde{X}$), so that

$P[\tilde{X}] = P[Y]$

(6) $P[X|Y]$ is called the *conditional* or *posterior probability* of event X given the occurrence of event Y. To detect a defect, one carries out a test. This test can have two outcomes (positive, which is supposed to mean there is a defect, or negative, leading to the conclusion there is no defect). The test in question is however not perfect: in a few cases a test will not be positive when there is a defect and now and then a non-defective sample will yield a positive test. $P[\text{defect}|\text{positive}]$ is then the

probability that an object is indeed defective, when a positive test has been obtained. It can be shown that

$$P[X|Y] = P[X \cap Y]/P[Y] \tag{15.7}$$

By contrast, $P[X]$ is then often called the *prior probability*. P[defect] is the (prior) probability that a specific object would be defective if no other information is available. When a test has been carried out, more information is present. P[defect|positive] is then the (posterior) probability that there is a defect, when the test was positive. A test is all the more informative when the difference between posterior and prior probability is larger. This is related to the concepts of sensitivity and selectivity of tests, introduced in Section 13.8.2 and treated in detail in Chapter 16 and to the information theory described in Chapter 18.

15.2 The binomial distribution

15.2.1 An example: the counter-current distribution method

Counter-current distribution (CCD) is a separation method in which one repeatedly partitions an analyte between two liquid phases. One phase is called stationary and the other mobile. In this technique one starts by bringing all the analyte into the first (Fig. 15.1a) of a series of cells. In each cell a stationary phase is available and in cell 1, one has also added mobile phase and one carries out the partition in that cell between the mobile phase and the stationary phase. The partition coefficient is given by $K = q/p$, where q is the fraction of analyte in the mobile phase and p is the fraction in the stationary phase. Since q and p are fractions, it also follows that

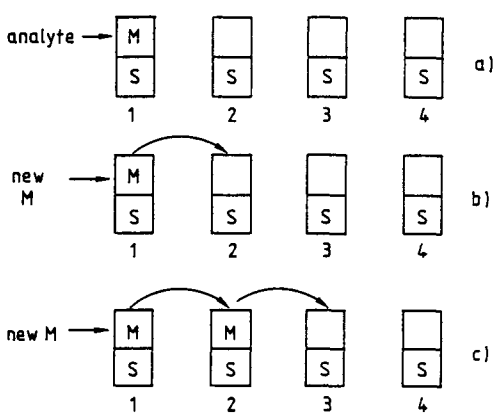


Fig. 15.1. The CCD process: (a) initial situation, (b) first transfer, (c) second transfer.

$$p + q = 1 = (p + q)^1 \tag{15.8}$$

The mobile phase of cell 1 is now transferred to cell 2 (Fig. 15.1b) and new mobile phase is added to the first cell. The analyte present in both cells is partitioned between the two phases in those cells. If Nernst's partition law is followed, the partition coefficient is independent of concentration and therefore the same in all cells throughout the distribution. In cell 1 there is a fraction p present. A fraction p of p (p^2) remains in the stationary phase and a fraction q of p (pq) is transferred to the mobile phase of cell 1. In cell 2 a fraction p of q (pq) remains in the stationary phase and a fraction q of q (q^2) will be found in the mobile phase. When the mobile phases of both cells are now transferred to the next cell, one finds in cell 1 a fraction p^2, in cell 2 a fraction 2 pq (pq remaining in the stationary phase of cell 2, pq from the mobile phase of cell 1) and in cell 3 a fraction q^2. The total amount is of course still equal to 1, so that

$$p^2 + 2\,pq + q^2 = 1 = (p + q)^2 \tag{15.9}$$

We can verify that after the next round of partitioning and transfer (Fig. 15.1c) of the mobile phase, the fractions in successive cells are p^3, 3 p^2q, 3 pq^2 and q^3, so that these fractions are given by the terms of $(p + q)^3$.

$$(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3 = 1 \tag{15.10}$$

In separation chemistry, counter-current distribution is used to separate two or more substances with different partition coefficients K. These will yield different distributions over the cells so that a separation effect occurs.

Let us now rephrase this in a way that is more customary in statistics. The analyte consists of a certain number of molecules. These molecules can assume two states in the separation process: they can go into the mobile phase (M) or stay in the stationary phase (S). Let us suppose that the molecules had three occasions to choose between M and S. Then we sort them as follows:

Situation 1: SSS

Situation 2: MSS

Situation 3: MMS

Situation 4: MMM

The order is not relevant. MSS means that the molecule once was M and twice S. The M-state may have occurred in partition step one, two or three.

The calculations of this section show that the different combinations of two possible states can be computed using the terms of the equation described by $(p + q)^n$. The resulting distribution is called the *binomial distribution* and we will discuss this distribution more formally in the next section.

15.2.2 The distribution

The binomial distribution can be used to model the distribution of objects that can take on two states. In statistics books, the binomial distribution is often explained with an urn in which there are a number of balls, the pth fraction of which are red, and the rest black. When one takes a ball from this urn it has a probability p to be red and a probability $q = 1 - p$ to be black. If R is red and B is black, then the following combinations of red and black balls can be obtained when three balls have been drawn.

Combination 1: RRR

Combination 2: BRR

Combination 3: BBR

Combination 4: BBB

where the order of B and R is not relevant: only the final result is important. It is not difficult to see the connection with the process described in the preceding section. It is shown that, if the drawing operation is repeated n times (i.e. there are n independently selected items) and the selected ball is replaced (i.e. put back into the urn each time), then the probability of having selected x red balls is:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0,1,\ldots,n \qquad 0 \le p \le 1 \tag{15.11}$$

with

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{15.12}$$

An example of a binomial distribution is given in Fig. 15.2. The population parameters for the binomial distribution are

$$\mu = np \qquad \sigma = \sqrt{np(1-p)} \tag{15.13}$$

or

$$\mu = p \qquad \sigma = \sqrt{p(1-p)/n} \tag{15.14}$$

depending on whether one presents the results as counts (number defectives, for instance) or as fractions (fraction defectives, for instance).

It should be noted here that one can also describe multinomial distributions. In this case each object can take on more than two states.

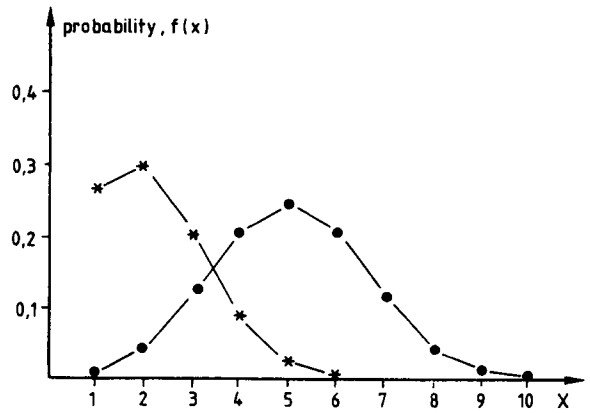


Fig. 15.2. Binomial distributions for $n = 10$, $p = 0.2$ (*) and $p = 0.5$ (●).

15.2.3 Applications in quality control: the np charts

Suppose that in a manufacturing process one selects at random 100 produced items and inspects them for a certain characteristic. It is known that the process, when it is under control, produces 2% defective items. The mean number of defectives will then be $\mu = np = 100 \times 0.02 = 2$. Of course, when one draws 100 items and inspects them this will often lead to the detection of fewer or more than 2 defective items. When the number of defects becomes much higher than 2, this will be considered as an indication that the process is no longer performing correctly. This reasoning is applied in constructing the *np chart* for attributes. At regular times one draws n items from the process and inspects them. One then sets warning and action lines such that (see Chapter 7) certain probabilities are not exceeded. These limits can be obtained from tables of the binomial distribution, but usually one applies approximate equations that are very similar to those applied in Chapter 7. The general equation is:

$$\text{Limit} = \mu + k\sigma$$

For the warning line (probability = 0.05) $k = 2$ and for the action line (probability = 0.01) $k = 3$, so that

$$\text{Upper action limit} = np + 3\sqrt{np(1-p)}$$

$$\text{Upper warning limit} = np + 2\sqrt{np(1-p)}$$

For our example (see Fig. 15.3) this becomes:

$$\text{Upper action limit} = 2 + 3\sqrt{2(0.98)} = 6.2$$

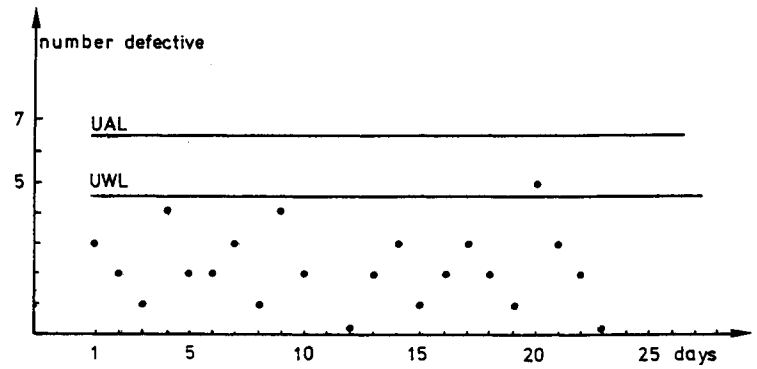


Fig. 15.3. An *np* chart. The number of defectives per 100 items is charted and the upper warning limit (UWL) and upper action limit (UAL) are given for a mean number of defectives equal to 2 ($p = 0.02$). Item 20 exceeds the UWL, but item 21 does not, so that the process is considered under control.

which means that when 7 defectives are found, one will consider that action must be taken. The upper warning limit is 4.8, so that 5 defectives is the level at which a warning is issued. The rules applied are the same as in Chapter 7. For instance, when two consecutive runs lead to a warning, then the process is considered to be out of control and action must be taken.

Instead of *np* charts, *p charts*, in which one plots the proportion of defectives, have also been described.

15.3 The hypergeometric distribution

The *hypergeometric distribution* is similar to the binomial distribution but sampled without replacement. Consider again the urn with red and black balls. To obtain the binomial distribution, we took one ball, noted its colour and put it back into the urn before taking the next ball. Suppose now that instead of putting each ball back in the urn, we simply take n balls and count the number of red balls. Suppose that there are m red balls in the urn, then the probability of taking a red ball at the outset is $p = m/N$ where N is the total number of balls. If the first ball to be taken out, turns out to be red, then for the second ball, p has changed to $(m-1)/(N-1)$. The probability of finding x individuals characterized by probability p in a sample of n items from a population of size N is then given by the hypergeometric distribution:

$$f(x) = \frac{\binom{N-m}{n-x}\binom{m}{x}}{\binom{N}{m}} \qquad (x = 0,1 \ldots, n) \tag{15.15}$$

In quality control terms, f(x) would be the probability of finding x defectives in a sample of size n taken from a population of size N in which the probability p of finding a defective sample is given by $p = m/N$.

At first sight, the QC application described in Section 15.2.3 might be considered a situation that should be described better by the hypergeometric than by the binomial distribution. However, it can be shown that when N is large compared with n, the hypergeometric distribution reduces to the binomial one. This follows for instance from a comparison of the population parameters. For the situation where the results are presented as proportion defectives, these population parameters are given by Duncan [1, p. 103]:

$$\mu = p \qquad \sigma = \sqrt{[p(1-p)/n] \cdot [(N-n)/(N-1)]} \tag{15.16}$$

The variance of the hypergeometrical distribution is equal to that of the binomial one multiplied by $(N-n)/(N-1)$ which becomes very close to 1 when $N >> n$.

15.4 The Poisson distribution

15.4.1 Rare events and the Poisson distribution

In the two preceding sections, we studied situations in which the total number of objects could be counted. We selected a sample of n objects, each of which was in one of two states. For instance, we took 10 objects from a production line and counted the number of defectives. The rest is not-defective. This is not always possible. Suppose the production consists of spray-painting a metal surface. Small defects may then occur on the painted surface. The number of defects can be counted, but how do we define the number of non-defects? In principle, the defects will occur only here and there and will therefore be relatively rare events. If we want to do quality control of the spray-painting shop, we will have to make conclusions based on the probability of 1, 2, 3,... defects occurring on a certain area: we have to study the distribution of a rare event. Other such situations are found in epidemiology or in measurements based on counting (microbiology, radioactivity). If x is the number of defects or counts observed in a given unit (area in m^2 for the spray-painting, time for radioactivity counting, ...) then f(x), the probability of observing x defects or counts in such a unit is given by

$$f(x) = e^{-\lambda}(\lambda^x/x!) \tag{15.17}$$

where λ is the average number of defects or counts observed in the given unit. Examples are shown in Fig. 15.4.

The population parameters are

$$\mu = \lambda \qquad \sigma = \sqrt{\lambda} \tag{15.18}$$

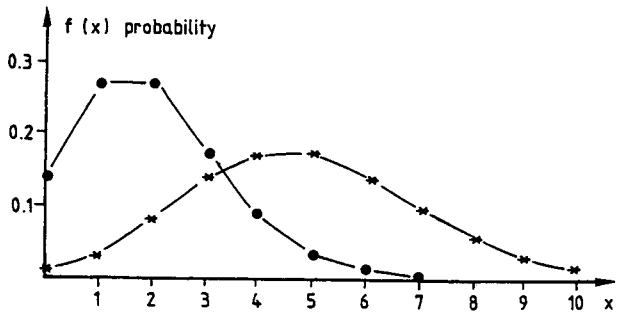


Fig. 15.4. Poisson distributions for $\lambda = 2$ (●) and $\lambda = 5$ (*).

Suppose, for instance, that a radioactive source is measured repeatedly. Then, if the mean counting rate is 10000 counts/min, the standard deviation is 100 counts/min. One observes that the higher the number of counts, the smaller is the relative standard deviation: counting precision is better when there are more counts.

It is possible to determine confidence limits for rarely occurring events. Suppose 3 new cases of juvenile diabetes are observed in a population of 30000 people over a certain period. The observed value 3 is an estimate of the true incidence in that

TABLE 15.1

95% confidence limits for Poisson-distributed values. n = observed value, LL = lower limit factor, UL = upper limit factor.

n	LL	UL	n	LL	UL
0	0	3.00			
1	0.0253	5.57	22	0.627	1.51
2	0.121	3.61	24	0.641	1.49
3	0.206	2.92	26	0.653	1.47
4	0.272	2.56	28	0.665	1.45
5	0.324	2.33	30	0.675	1.43
6	0.367	2.18	35	0.697	1.39
7	0.401	2.06	40	0.714	1.36
8	0.431	1.97	50	0.742	1.32
9	0.458	1.90	60	0.770	1.30
10	0.480	1.84	80	0.798	1.25
11	0.499	1.79	100	0.818	1.22
12	0.517	1.75	150	0.849	1.178
13	0.532	1.71	200	0.868	1.151
14	0.546	1.68	250	0.882	1.134
15	0.560	1.65	300	0.892	1.121
16	0.572	1.62	400	0.906	1.104
17	0.583	1.60	500	0.915	1.093
18	0.593	1.58	600	0.922	1.084
19	0.602	1.56	800	0.932	1.072
20	0.611	1.54	1000	0.939	1.064

population over such a period. The 95% confidence limits on that observation can be obtained by multiplying it by the lower and upper limit factors from Table 15.1 [2]. This yields the confidence interval 3×0.206 to 3×2.92 or 0.62 to 8.8. The confidence interval for the incidence rate per 100000 people is therefore 2.1 to 29.3. The uncertainty will be smaller when the sample size is larger. Suppose one had observed the same incidence rate on a population of 300000. We would then have observed 30 cases. The confidence intervals for the incidence rate per 100000 people would then be $30 \times 0.675 \times (100000/300000)$ to $30 \times 1.43 \times (100000/300000)$ or 6.75 to 14.3. Since we can determine confidence intervals, we can also carry out hypothesis tests. The test will be more powerful when the sample size is larger, since the confidence intervals are then relatively smaller. The same conclusion was reached in Section 4.8.

15.4.2 Application in quality control: the c and u-charts.

The *c-chart* is the traditional name for a chart which monitors the number of defectives in situations such as the spray-painting example. $\bar{c}$ is then the average number of defects per unit (i.e. is equal to λ, the symbol more generally used in statistical texts). The limits are then given by:

Upper action limit ($p = 0.01$) $= \bar{c} + 3\sigma = \bar{c} + 3\sqrt{\bar{c}}$

Upper warning limit ($p = 0.05$) $= \bar{c} + 2\sigma = \bar{c} + 2\sqrt{\bar{c}}$

Suppose the mean number of defects $\bar{c}$ per unit is 4.0, then from eq. (15.18), it follows that $\sigma = \sqrt{4.0} = 2.0$. The upper warning limit is then given by $4.0 + 2 \times 2.0 = 8.0$ and the upper action limit by $4.0 + 3 \times 2.0 = 10.0$. An example of a chart is shown in Fig. 15.5.

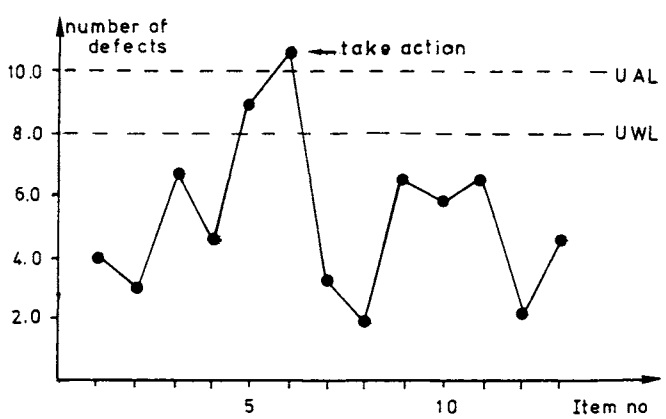


Fig. 15.5. A *c*-chart. Item no. 5 exceeds the UWL, the inspection is repeated and yields no. 6 which exceeds the UAL. Action must be taken and is successful (item no. 7). Adapted from Ref. [3].

We use *u-charts* when the unit investigated is not always the same. Suppose that instead of investigating always the same area of spray-painted metal, we decide to investigate metal plates of varying area; u would then be the number of defects per m^2 and follows again a Poisson distribution.

15.4.3 Interrelationships between the binomial, Poisson and normal distributions

One approach to introducing the Poisson distribution is to consider it as a limiting case of the binomial distribution for which n tends to infinity, p becomes very small but np remains constant and equal to λ. The Poisson distribution tends to normality when λ is sufficiently large ($\lambda > 10$). The binomial distribution can be approximated with a normal distribution when $np > 5$ and $n(1-p) > 5$.

15.5 The negative exponential distribution and the Weibull distribution

The *negative exponential distribution* is of interest in the SPC for describing lifetime (time-to-failure), degradation or reliability of a product. It is given by

$$f(x) = \frac{1}{\theta} e^{-x/\theta} \tag{15.19}$$

or

$$f(x) = \frac{1}{\theta} e^{-(x-t)/\theta} \tag{15.20}$$

where t is a threshold. The mean is equal to θ in eq. (15.19) and to $\theta + t$ in eq. (15.20). In both cases the standard deviation is equal to θ. An example of the distribution with and without threshold is given in Fig. 15.6.

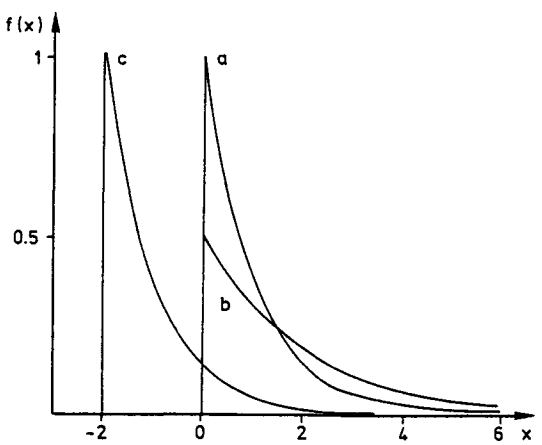


Fig. 15.6. The negative exponential distribution: (a) $t = 0$, $\theta = 1$; (b) $t = 0$, $\theta = 2$; (c) $t = -2$, $\theta = 1$.

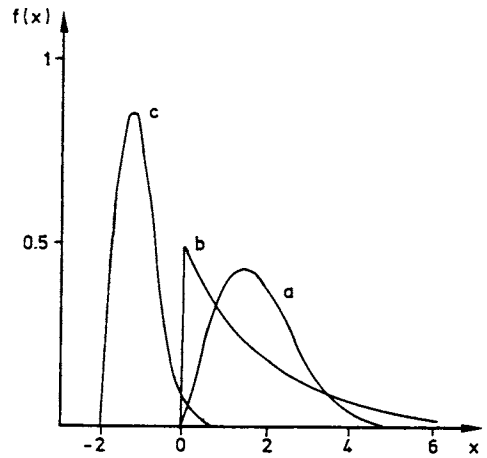


Fig. 15.7. The Weibull distribution: (a) $t = 0, \theta = 2, \beta = 2$; (b) $t = 0, \theta = 2, \beta = 1$; (c) $t = -2, \theta = 1, \beta = 2$.

The negative exponential distribution is a special case of a more general distribution, called the *Weibull distribution* (Fig. 15.7). It is given by

$$f(x) = (\beta/\theta)\left(\frac{x-t}{\theta}\right)^{\beta-1} \exp\left(-\left(\frac{x-t}{\theta}\right)^{\beta}\right) \tag{15.21}$$

Again, t is a threshold, often equal to 0, and $\beta, \theta > 0$. Examples of the distribution are shown in Fig. 15.7. It can be verified that for $\beta = 1$ the negative exponential distribution is obtained.

15.6 Extreme value distributions

Extreme values are of interest when describing or predicting catastrophic situations, e.g. the occurrence of floods or for safety considerations. In this section we will follow the description of extreme-value techniques given by Natrella [4].

Figure 15.8 is a typical curve for the distribution of largest observations. This curve is the derivative of the cumulative probability function

$$F(x) = \exp[-\exp(-x)] \tag{15.22}$$

This distribution is skewed and describes largest values, for instance the largest values of atmospheric pressure obtained in a year in a certain location. The distribution can be used for extreme-value plots similar in approach to normal-probability plots. In one axis we plot the value of x and in the other the probability, according to eq. (15.22), of the observations ranked from smallest to largest.

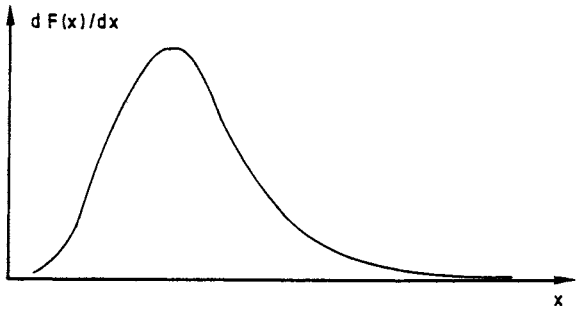


Fig. 15.8. An extreme-value distribution (adapted from Ref. [4]).

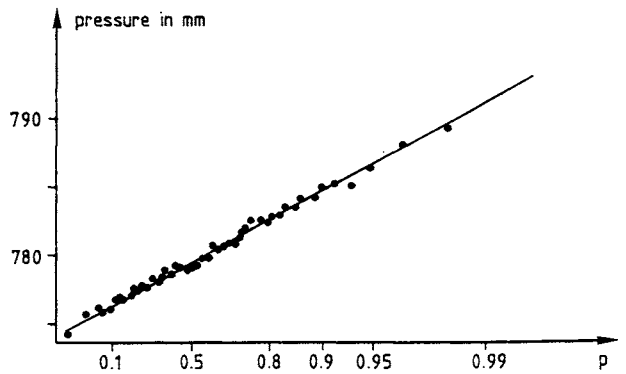


Fig. 15.9. Extreme-value plot. The plot describes the largest annual atmospheric pressure in Bergen (Norway) in the period 1857–1926 (adapted from Ref. [5]).

Suppose there are 50 observations, then the smallest one will be plotted at the probability value 1/(50+1). An example is shown in Fig. 15.9. This graph allows to conclude that the probability that the largest value will exceed 793 mm in any year is 0.01. Smallest values can also be plotted. They are of interest, for instance, in fracturing or fatigue situations.

Other distributions can also be applied, for example, the Pareto distribution of Fig. 2.8. One of the main tasks in studying extreme value distributions is then to decide which of the many possible distributions fits the data best. More details can be found in a book by Beirlant et al. [6].

References

1. A.J. Duncan, Quality control and industrial statistics. Irwin, Homewood, 5th edn. 1986.
2. W. Haenszel, D. Loveland and M.G. Sirken, Lung-cancer mortality as related to residence and smoking histories. J. Nat. Cancer Inst., 28 (1962) 947–1001.

3. J.S. Oakland, Statistical Process Control, Wiley, New York, 1992.
4. M.G. Natrella, Experimental Statistics, National Bureau of Standards Handbook 91, Chapter 19, 1963.
5. E.J. Gumbel and J. Lieblein, Some applications of extreme-value methods. Am. Statistician, 8 (5) (1954), as cited by Natrella, Ref. 4.
6. J. Beirlant, L. Teugels and P. Vynckier, Practical Analysis of Extreme Values. Leuven University Press, Leuven, Belgium, 1996.

Chapter 16

The 2×2 Contingency Table

16.1 Statistical descriptors

16.1.1 Variables, categories, frequencies and marginal totals

A *contingency table* arises from the classification of a sample according to two qualitative variables [1]. Hence, the cells of a contingency table contain counts or frequencies. Each cell of a contingency table, say in row i and column j, represents the number of elements in the sample that have been observed to belong simultaneously to category i of the first variable and to category j of the second variable. If the two variables have n and p categories, respectively, we speak of an $n \times p$ contingency table. We assume that the categories of any given variable are exhaustive and mutually exclusive, which means that each element in the sample can be classified according to that variable into one and only one category.

In this chapter we deal exclusively with 2×2 contingency tables. The general $n \times p$ case will be treated extensively in Chapter 32. Here we assume that the two variables are dichotomous. A dichotomous variable provides for only two categories. This is the case with the outcome of a screening assay when it is reported as either positive or negative, and with a diagnosis of a patient when it is stated as either diseased or healthy. In dichotomous variables, such as outcome and diagnosis above, there is no provision for categories in between. For example, one may have studied a cohort of twenty patients that became HIV positive at about the same time. Ten of these received antiviral monotherapy (e.g. AZT), the remaining ten received combination therapy (e.g. AZT + ddI). The aim of the study is to know whether the expected proportion of patients that developed AIDS symptoms after 5 years of therapy is smaller under combination therapy than under monotherapy. There are four contingencies in this study, i.e.: AIDS (mono), no AIDS (mono), AIDS (combination) and no AIDS (combination), and the corresponding frequencies of occurrence can be arranged in a 2×2 contingency table. Clearly, if combination therapy is superior to monotherapy we would expect to obtain more patients without AIDS in the former category. The difference between observed and expected frequencies, however, may be due to chance, especially as we are dealing with a small number of patients. In this chapter we will develop the necessary statistical concepts and tests which will lead to correct conclusions from 2×2 contingency tables.

TABLE 16.1

Outcome of assay A for diagnosis of HIV infection

Outcome	Diagnosis		Total
	Diseased	Healthy	
Positive	276	15	291
Negative	24	285	309
Total	300	300	600

By way of illustration we consider the outcomes of an assay for the presence of HIV antibodies in urine which we have adapted from Science Briefings [2]. We assume here that this screening test, which we call assay A, has been applied to 600 subjects. Specimens of urine were collected from 300 persons with confirmed HIV infection and from 300 uninfected persons. It was found that 276 of the 300 diseased persons tested positively in this assay, while only 15 of the 300 healthy persons obtained a positive outcome. The two variables in this example are diagnosis (diseased or healthy) and outcome (positive or negative). The results are summarized in Table 16.1.

The row and column totals of a contingency table are called the marginal totals. They indicate the cell frequencies observed in each category of the two variables. The grand total represents the sample size. In Table 16.1 we find that the sample of size 600 has been evenly divided between the two diagnostic categories (diseased and healthy). Note that the column totals in this example are fixed by the design of the assay. We also observe that the same sample is divided into 291 persons with positive outcome and 309 with negative outcome. In practice, these row totals are subject to random sampling errors. Replication of the assay, using identical sample size and marginal column totals, will probably produce different row totals. Later on we will discuss the case where only the grand total is fixed. This situation occurs for example in epidemiological studies where the number of persons found with or without a disease is also subject to sampling error. On rare occasions a design is obtained in which both sets of marginal totals are fixed.

16.1.2 Probability and conditional probability

We consider a 2×2 contingency table with the dichotomous variables of diagnosis and outcome. In Table 16.2 we define a general layout for 2×2 contingency tables in the context of diagnostic assays [3]. This layout can be adapted to chemical and other applications by replacing the names of the variables and their dichotomous categories.

TABLE 16.2

2×2 contingency table for diagnostic assay

Outcome	Diagnosis		Total
	Diseased	Healthy	
Positive	tp	fp	tp + fp
Negative	fn	tn	fn + tn
Total	tp + fn	fp + tn	N

TABLE 16.3

2×2 contingency table for diagnosis and outcome, in terms of probabilities

Outcome	Diagnosis		Total
	Diseased	Healthy	
Positive	$P[d \cap p]$	$P[h \cap p]$	$P[p]$
Negative	$P[d \cap n]$	$P[h \cap n]$	$P[n]$
Total	$P[d]$	$P[h]$	1

In Table 16.2 N represents the sample size or grand total. The variable diagnosis has the categories diseased (d) and healthy (h). The variable outcome consists of the categories positive (p) and negative (n). The four cells of the table correspond to the four possible contingencies, the probabilities of which are shown in Table 16.3:

true positive (tp) = diseased and positive $(d \cap p)$
false negative (fn) = diseased and negative $(d \cap n)$
false positive (fp) = healthy and positive $(h \cap p)$ (16.1)
true negative (tn) = healthy and negative $(h \cap n)$

We denote by $P[X]$ the probability of the occurrence of event X. In particular, $P[d \cap p]$ is the joint probability of observing simultaneously a positive outcome and a diseased person. Likewise, $P[d]$ is the *prevalence* of the disease and $P[p]$ is the probability of a person obtaining a positive outcome whatever his diagnosis. The probabilities in the table are estimated by dividing observed frequencies by the sample size N (Table 16.4):

$P[d \cap p]$ is estimated by tp/N
$P[d]$ is estimated by $(tp + fn)/N$
$P[p]$ is estimated by $(tp + fp)/N$, etc.

TABLE 16.4

2×2 contingency table for diagnosis and outcome, showing observed proportions

Outcome	Diagnosis		Total
	Diseased	Healthy	
Positive	tp/*N*	fp/*N*	(tp + fp)/*N*
Negative	fn/*N*	tn/*N*	(fn + tn)/*N*
Total	(tp + fn)/*N*	(fp + tn)/*N*	1

From the axioms of probability, which have been stated in Chapter 15, we can derive the following propositions in the context of our example from clinical assays:

$$P[\mathrm{d}] \geq 0$$

$$P[\mathrm{d} \cup \mathrm{h}] = P[\mathrm{d}] + P[\mathrm{h}] = 1 \tag{16.2}$$

$$P[\mathrm{d} \mid \mathrm{p}] = P[\mathrm{d} \cap \mathrm{p}]/P[\mathrm{p}]$$

We must also have that:

$$P[\mathrm{p}] = P[\mathrm{p} \cap \mathrm{d}] + P[\mathrm{p} \cap \mathrm{h}] \tag{16.3}$$

since d and h are mutually exclusive and exhaustive events: a positive person is either diseased or healthy.

After substitution of eq. (16.3) in eq. (16.2) we obtain:

$$P[\mathrm{d} \mid \mathrm{p}] = \frac{P[\mathrm{d} \cap \mathrm{p}]}{P[\mathrm{p} \cap \mathrm{d}] + P[\mathrm{p} \cap \mathrm{h}]} \tag{16.4}$$

(Note that $\mathrm{p} \cap \mathrm{d}$ equals $\mathrm{d} \cap \mathrm{p}$ because of the commutative property of the intersection.)

If we replace probabilities by observed relative cell frequencies in eq. (16.4) we develop an expression for the conditional probability:

$$P[\mathrm{d} \mid \mathrm{p}] \approx \frac{\mathrm{tp}/N}{(\mathrm{tp} + \mathrm{fp})/N} = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}} \tag{16.5}$$

The symbol ≈ denotes that two expressions are approximately equal. It is used here to indicate that theoretical probabilities are substituted by observed proportions.

16.1.3 Sensitivity and specificity

In order to compare assays with one another we need to measure their performance. Let us consider, for example, a competitive assay for the detection of HIV antibodies in urine. This assay, which we refer to as assay B, has been applied to

TABLE 16.5

Outcome of assay B for diagnosis of HIV infection

Outcome	Diagnosis		Total
	Diseased	Healthy	
Positive	364	12	376
Negative	36	388	424
Total	400	400	800

400 persons who have been diagnosed previously to be carriers of the infection and to 400 persons who were known to be free of infection. The observed outcomes are found in Table 16.5. If we compare assays A and B, which of the two is to be preferred? The ideal assay is, of course, the one which produces a positive outcome for all persons in which the disease is present. Such an assay possesses a *sensitivity* of 100%. Sensitivity is defined here as the proportion of true positives with respect to the total diseased:

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Total number of diseased}} = \frac{\text{tp}}{\text{tp} + \text{fn}} \approx P[\text{p} \mid \text{d}] \tag{16.6}$$

where tp and fn represent the number of true positives and the number of false negatives, respectively. The sensitivity of an essay is also called power. In the context of testing of hypotheses (Section 4.7) one also defines power as $1 - \beta$, where β represents the probability of obtaining a false negative, or $P[\text{n} \mid \text{d}]$ in our notation.The ideal assay also produces a negative outcome for all persons in which the disease is absent. Such an assay has a *specificity* of 100%. Specificity is defined here as the proportion of true negatives with respect to total healthy.

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Total number of healthy}} = \frac{\text{tn}}{\text{tn} + \text{fp}} \approx P[\text{n} \mid \text{h}] \tag{16.7}$$

where tn and fp represent the number of true negatives and the number of false positives, respectively. In a statistical context one relates (1 – specificity) to the level of significance of a test which is denoted by α and which represents the probability of obtaining a false positive, or $P[\text{p} \mid \text{h}]$ in our notation.

In the case of assay A for screening against HIV infection we estimate the sensitivity and specificity to be as follows:

Sensitivity = 100(276/300) = 92%

Specificity = 100(285/300) = 95%

In the case of assay B we obtain the following estimates:

Sensitivity = 100(364/400) = 91%

Specificity = 100(388/400) = 97%

On the basis of these estimates alone we cannot decide which of the two assays performs best. On the one hand, assay A shows greater sensitivity than assay B. On the other hand, assay B possesses greater specificity. But are these differences significant in a statistical sense? Perhaps replication of the assays using different samples of infected and non-infected persons may produce different results. It is important to realize that the above sample estimates are not the population values of sensitivity and specificity of assays A and B. In order to make a sound statistical analysis we also need to know the variances of these sample estimates. This idea will be pursued in Section 16.2 on hypothesis testing. Let us suppose, however, that the above estimates reflect statistically significant differences which are highly unlikely to be due to the effect of random sampling. The question then arises whether these differences have practical relevance.

Assay A has the greatest sensitivity and is therefore expected to produce more true positives and, hence, a smaller number of false negatives. The cost to society of declaring a diseased person to be healthy may be enormous. False negatives may unknowingly spread the disease, contaminate blood banks and so on. Assay B has the greatest specificity and is therefore expected to yield more true negatives and, consequently, fewer false positives. In the case of mass screening it is mandatory to retest all positives in order to protect false positives from unnecessary treatment and discomfort. However, second-line assays are usually more expensive and time-consuming than primary screening assays. Additionally, as we will see in the next section, rare diseases may produce large numbers of false positives, even with highly specific assays. Hence, the cost to society of declaring healthy persons to be diseased may also be considerable. As will be discussed in more detail later, an increase in sensitivity is usually at the expense of a decrease in specificity, and *vice versa*. The balancing of costs and risks associated with the introduction of new diagnostic assays and therapeutic treatments is a delicate task which also involves the competence of health economists.

In the context of analytical chemistry definitions for specificity and sensitivity similar to those described here have been proposed by the Association of Official Analytical Chemists (AOAC). These definitions are described in Section 13.8 and 13.9 and are applicable to immunological assays, microbiological assays, clinical studies and clinical chemistry. In other areas of analytical chemistry these terms have very different meanings. Sensitivity is defined as the slope of a calibration line relating the strength of the output signal to the concentration of a component in a material to be analyzed [4]. Specificity is described as the ability of an analytical method to respond to only one specific component in a mixture [5]. It is

clear that the latter definitions are unrelated to the statistical definitions which we use here in our discussion of the 2×2 contingency table.

16.1.4 Predictive values

In the laboratory, sensitivity and specificity of a diagnostic assay can be estimated from samples of diseased and healthy persons, the sizes of which can be fixed in the design of an experiment. But, in screening for a disease, the sizes of the samples of diseased and healthy persons depend on the prevalence of the disease, i.e. the proportion of diseased persons in the population at the time of observation. Hence, in a 2×2 contingency table relating diagnosis of a disease to outcome of a screening test in a large sample we find that all the marginal totals are subject to sampling error.

In Tables 16.6 and 16.7 we have constructed the presumed outcomes from screening 1 million persons for infection by HIV, using the sensitivities and specificities of assays A and B which have been estimated above. It is assumed that the prevalence of HIV infection in the general population is 1 in 2000. Consequently, in a large sample of 1 million persons we expect to find 500 infected persons. With assay A we expect to find 40 false negatives against 45 with assay B. This is in accordance with a difference of 1% in sensitivity in favour of assay A.

TABLE 16.6

2×2 contingency table for screening with assay A

Outcome	Diagnosis		Total
	Diseased	Healthy	
Positive	460	49975	50435
Negative	40	949525	949565
Total	500	999500	1000000

TABLE 16.7

2×2 contingency table for screening with assay B

Outcome	Diagnosis		Total
	Diseased	Healthy	
Positive	455	29985	30440
Negative	45	969515	969560
Total	500	999500	1000000

With assay A we expect to obtain 49975 false positives compared to 29985 with assay B. This is in agreement with a difference of 2% in specificity in favour of assay B. The large number of false positives is typical for screening of phenomena with low prevalence in a large sample, such as for HIV infection, tuberculosis, drug abuse, doping, etc. Even highly specific assays may still produce large numbers of false positives.

The performance of a screening assay can be measured by means of its *positive predictive value* (PPV) which is an estimate of the proportion of true positives (tp) with respect to the total number of positives (tp + fp):

$$\text{PPV} = \frac{\text{Number of true positives}}{\text{Total number of positives}} = \frac{\text{tp}}{\text{tp} + \text{fp}} \approx P[\text{d} \mid \text{p}] \tag{16.8}$$

the right side of which is equal to that of eq. (16.5).

It thus follows that the positive predictive value of an assay is the conditional probability for the presence of a disease when the outcome is positive, or $P[\text{d} \mid \text{p}]$ in our notation.

In screening for HIV infection we obtain:

PPV = 100 (460/50435) = 0.91% using assay A

PPV = 100 (455/30440) = 1.50% using assay B.

The positive predictive values of both assays for HIV antibodies in urine are expected to be quite low when these assays will be applied to mass screening. The positive predictive value depends on both the sensitivity and specificity of the assay and on the prevalence of the phenomenon. This relationship follows from Bayes' theorem which will be demonstrated below.

In a similar way we define the *negative predictive value* (NPV) as:

$$\text{NPV} = \frac{\text{Number of true negatives}}{\text{Total number of negatives}} = \frac{\text{tn}}{\text{tn} + \text{fn}} \approx P[\text{h} \mid \text{n}] \tag{16.9}$$

16.1.5 Posterior and prior probabilities, Bayes' theorem and likelihood ratio

Using the axioms of probability (eq. (16.2)) we can express the conditional probability in eq. (16.4) in the following form:

$$P[\text{d} \mid \text{p}] = \frac{P[\text{p} \mid \text{d}] \times P[\text{d}]}{P[\text{p}]} = \frac{P[\text{p} \mid \text{d}] \times P[\text{d}]}{P[\text{p} \mid \text{d}] \times P[\text{d}] + P[\text{p} \mid \text{h}] \times P[\text{h}]}$$

which can be rearranged into:

$$P[\text{d} \mid \text{p}] = \frac{(P[\text{p} \mid \text{d}]/P[\text{p} \mid \text{h}]) \times P[\text{d}]}{(P[\text{p} \mid \text{d}]/P[\text{p} \mid \text{h}]) \times P[\text{d}] + (1 - P[\text{d}])} \tag{16.10}$$

since $P[\text{h}] = 1 - P[\text{d}]$.

The conditional probability $P[d \mid p]$ is also called the *posterior probability*. It is the probability of finding a person with a disease, knowing that his outcome in a diagnostic assay is positive. It also follows from eq. (16.8) that the posterior probability can be estimated by means of the positive predictive value. The probability $P[d]$ represents the prevalence of the disease which is also called the *prior probability*. It is the probability of finding a person with the disease without having any prior knowledge.

Equation (16.10) is known as *Bayes' theorem*. In this context it relates the positive predictive value of an assay $P[d \mid p]$ to the prevalence of the disease $P[d]$ by means of the ratio of two conditional probabilities which is called the *likelihood ratio LR*:

$$\text{Likelihood Ratio} = LR = P[p \mid d]/P[p \mid h] \tag{16.11}$$

After substitution of probabilities by relative cell frequencies we derive that:

$$LR = \frac{P[p \cap d]}{P[d]} \Big/ \frac{P[p \cap h]}{P[h]} \approx \frac{tp/N}{(tp + fn)/N} \Big/ \frac{fp/N}{(tn + fp)/N} \tag{16.12}$$
$$= \frac{tp}{tp + fn} \Big/ \frac{fp}{tn + fp}$$

and after rearrangement we obtain:

$$LR \approx \frac{tp}{tp + fn} \Big/ \left(1 - \frac{tn}{tn + fp}\right) = \frac{\text{Sensitivity}}{1 - \text{Specificity}} \tag{16.13}$$

Finally, after substitution of eq. (16.11) into the expression of Bayes' theorem by eq. (16.10), we obtain the relationship between posterior probability and prior probability (prevalence):

$$\text{Posterior Probability} = \frac{LR \times \text{Prevalence}}{LR \times \text{Prevalence} + 1 - \text{Prevalence}} \tag{16.14}$$

16.1.6 Posterior and prior odds

The concept of *odds* stems from the study of betting. How is a stake to be divided fairly between two betters which are betting for the occurrence of an event with known probability P? A bet can be regarded to be fair when the expected gain is zero for each of the two parties. In such a fair game (also called Dutch book) there is no advantage of betting one way or another, i.e. for the occurrence or for the non-occurrence of an event with probability P. In this case it can be shown [6] that the stake is to be divided between two betters according to the ratio:

$$\text{Odds}(P) = P/(1 - P) \tag{16.15}$$

which is called the odds for an event with probability P. The odds associated with an event range between 0 for certain non-occurrence ($P = 0$) to infinity in the case of certain occurrence ($P = 1$). Odds larger than unity are associated with events that

possess probabilities larger than 0.5. Conversely, odds smaller than unity are related to events with probabilities smaller than 0.5. The relation between odds and probabilities (eq. (16.15)) can be inverted yielding:

$$P = \text{Odds}/(1 + \text{Odds}) \tag{16.16}$$

The use of odds instead of probabilities is favoured in anglo-saxon countries. It also presents an advantage in the calculation of pay-offs in horse races and other gambling games. Here, we introduce the concept of odds because it simplifies formulas that involve prior and posterior probabilities.

In the previous subsection we have established that the positive predictive value (PPV) is an estimate of posterior probability (eq. (16.10)). In terms of odds we can now write that:

$$\text{Posterior Odds} \approx \frac{\text{PPV}}{1 - \text{PPV}} = \frac{\text{tp}}{\text{fp}} \tag{16.17}$$

The *posterior odds* reflect our belief in the occurrence of a disease in the light of the outcome of an assay.

Similarly, we can define *prior odds*:

$$\text{Prior Odds} \approx \frac{\text{Prevalence}}{1 - \text{Prevalence}} = \frac{\text{tp} + \text{fn}}{\text{tn} + \text{fp}} \tag{16.18}$$

The prior odds reflect our belief in the occurrence of a disease before an assay has been performed.

Finally, if we combine eqs. (16.12), (16.17) and (16.18), we obtain a very elegant relationship between prior and posterior odds:

$$\text{Posterior Odds} = LR \times \text{Prior Odds} \tag{16.19}$$

The likelihood ratio LR thus appears as a factor which, when multiplied by the prior odds, returns the posterior odds. It is the ratio of posterior odds to prior odds. Hence it tells by how much we are inclined to modify our initial belief in the occurrence of a disease when we are informed about a positive outcome of an assay. This is a *Bayesian approach* to statistical decision-making. It differs from the widely practised *Neyman–Pearson approach* of hypothesis testing, which will be discussed in Section 16.2. One of the principal aspects of the Bayesian approach is that often one starts with a subjective guess about the prior probability. In the light of evidence that becomes available, the latter is transformed into a posterior probability, which in turn becomes a revised prior probability in a subsequent analysis. It is claimed that in the end, after many revisions, the posterior probability becomes independent of the initial subjective prior probability.

From a philosophical point of view, the Bayesian approach of posterior probabilities is deemed to be more scientific than the testing of a single hypothesis, as

the former accumulates knowledge from several assays, whereas the latter relies on the outcome of a single experiment [6].

Bayesian statistics can be regarded as an extension of Neymann–Pearson hypothesis testing, in which the sample size is artificially increased as a result of prior information. The stronger the *a priori* evidence, the larger will be the number of *pseudosamples* that can be added to the observed sample size. Hence, the power of the test is effectively increased, and stronger assertions can be derived from the Bayesian test than would have been the case with classical statistics.

The Bayesian approach finds important applications in medical diagnosis [7,8] and in risk assessment. It is of importance in all disciplines where decisions have to be made under conditions of uncertainty in the light of experimental outcomes and where maximal use must be made of prior knowledge.

16.1.7 Decision limit

The elevation of serum creatine kinase (SCK) is a diagnostic indicator for the destruction of heart tissue during myocardial infarction (MI). Radack et al. [7] have presented the outcomes of an experiment in which SCK (in IU/ml) was determined in 773 persons who complained of chest pain. Of these, 51 were confirmed to suffer from an attack of myocardial infarction, while the other 722 did not. The categorized data are shown in Table 16.8 and the corresponding hand-fitted distributions are presented in Fig. 16.1.

The mean SCK in patients with myocardial infarction is 234 with standard deviation of 190 IU/ml (coefficient of variation of 81%). The mean SCK in patients without myocardial infarction is 117 with standard deviation of 97 IU/ml (coefficient of variation of 83%). Although patients with the disease can be expected to

TABLE 16.8

Serum creatine kinase (SCK in IU/ml) and myocardial infarction in persons complaining of chest pain [7]. Number of persons (N) and proportion (f) of total number in each group of persons

SCK(IU/ml)	Myocardial infarction			
	Present		Absent	
	N	f	N	f
0–120	23	0.451	471	0.652
121–240	6	0.118	201	0.278
241–360	7	0.137	24	0.033
361–480	6	0.118	12	0.017
>480	9	0.176	14	0.019
Total	51	1.000	722	1.000

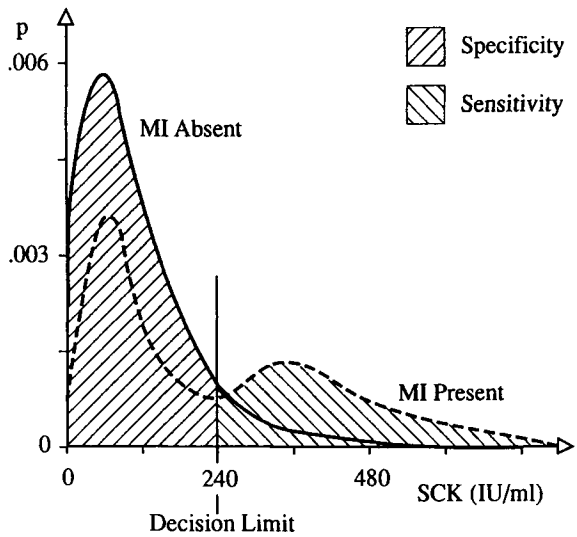


Fig. 16.1. Distributions of outcomes of serum creatine kinase (SCK in IU/ml) in subjects with myocardial infarction present and absent. The decision limit of SCK is fixed at 240 IU/ml. (Adapted from data by Radack et al. [7]).

produce larger values of SCK, the overlap between the distributions of persons with and without the disease is considerable, as is apparent from Fig. 16.1. It also appears that the distributions are far from being normal. Both show positive skewness towards large values of SCK. There is also an indication that one of the distributions may be bimodal. This situation is not unexceptional with clinical chemistry data, especially with enzyme assays [9].

From the data in Table 16.8 we can estimate the prevalence of myocardial infarction in patients with chest pain to be $51/(51 + 722) = 0.066$. Hence, the prior odds of the disease is estimated at $0.066/(1 - 0.066) = 0.071$.

From these data we construct four 2×2 contingency tables, corresponding with four different decision limits, namely 120, 240, 360 and 480 IU/ml (Table 16.9). The *decision limit* defines the value above which the outcome of an assay is declared to be positive. The decision limit also defines the sensitivity and specificity of the assay. The shaded areas on Fig. 16.1 are proportional to the specificity and sensitivity values of the SCK assay at the decision limit of 240 IU/ml. At each of the four decision limits we computed the different measures of performance which we have discussed so far: sensitivity, specificity, predictive value and likelihood ratio. These results are presented in Table 16.10.

From this analysis it follows that the positive predictive value of the assay is rather low. At the highest decision level of 480 IU/ml we expect only 39.1% of all positive outcomes to be true positives. This is due in this case to the low prevalence of the disease (0.066). The likelihood ratio indicates, however, that the assay may

TABLE 16.9

SCK and myocardial infarction (MI) in patients with chest pain using different decision limits [7]

SCK	MI		Total	SCK	MI		Total
	Present	Absent			Present	Absent	
>480	9	14	23	>360	15	26	41
≤480	42	708	750	≤360	36	696	732
Total	51	722	773	Total	51	722	773
>240	22	50	72	>120	28	251	279
≤240	29	672	701	≤120	23	471	494
Total	51	722	773	Total	51	722	773

TABLE 16.10

Measures of performance of the SCK assay for myocardial infarction at different decision limits

Decision limit	Sensitivity	Specificity	Positive predictive value	Likelihood ratio
120	0.549	0.652	0.100	1.58
240	0.431	0.931	0.306	6.25
360	0.294	0.964	0.366	8.17
480	0.176	0.981	0.391	9.26

be of considerable value. At the decision level of 480 IU/ml it modifies our prior odds of the disease (0.071) into posterior odds by a factor of 9.26. Even at lower decision limits this assay appears to perform well. For example, an outcome of 240 IU/ml would increase the odds for the disease by a factor of 6.25.

From Table 16.10 we observe an inverse relationship between sensitivity and specificity. An increase in sensitivity (proportion of positive outcomes from total with disease) is at the expense of specificity (proportion of negative outcomes from total without disease). Conversely, we have a direct relationship between sensitivity and (1 – specificity), which is called the *receiver operating characteristic* (ROC). The ROC is also used extensively in statistical process control (Chapter 20). It will be developed in more detail below.

16.1.8 Receiver operating characteristic

The predictive value of an assay can be displayed in the form of a plot of sensitivity against (1 – specificity) at various settings of the decision limit. The

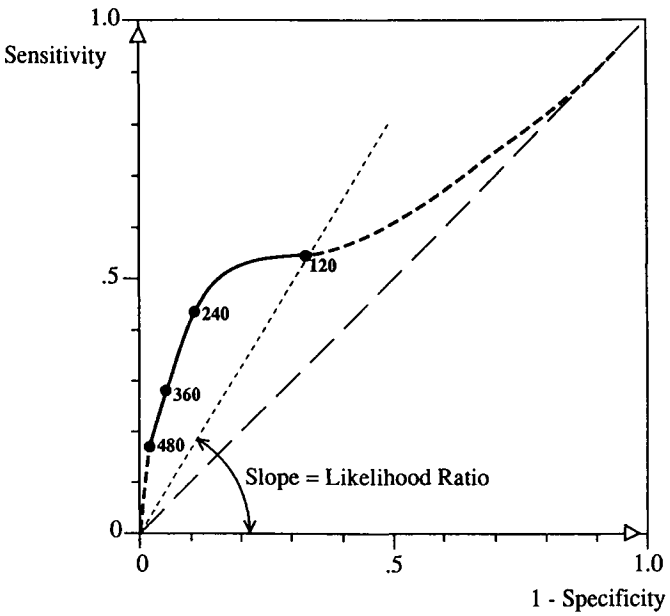


Fig. 16.2. ROC of SCK assay for myocardial infarction, from data by Radack et al. [7].

resulting curve is called the receiver operating characteristic or ROC for short. The use of ROC curves stems from the study of signal detection (initially in the development of radar) and has been applied extensively in psychophysiology [10]. The ROC curve for the serum creatine kinase (SCK) assay for myocardial infarction in persons complaining of chest pain [7] is presented in Fig. 16.2.

The diagonal line of this diagram represents the case of an assay with zero predictive value. The greater the distance of the ROC curve from this diagonal, the more performing is the assay. ROC curves allow to estimate the performance of an assay over a wide range of decision limits and independently from the prevalence of the phenomenon under investigation. It is well-suited for the comparison of assays and methods.

An alternative ROC diagram represents sensitivity and (1 – specificity) along axes of normal deviates (z) [11]. If the distributions of the outcomes in subjects with and without the disease are normal, then the ROC curve is transformed into a straight line. The distance of this line from the diagonal line is again a measure of the performance of the test. Figure 16.3 represents the transformed ROC curve of Fig. 16.2. The deviation from linearity of the transformed ROC curve of Fig. 16.3 is the result of the apparent lack of normality of the distributions of SCK in populations with the disease present and absent (Fig. 16.1).

In statistical terms, the ROC defines the relationship between the previously defined α-error (of obtaining a false positive), the β-error (of obtaining a false

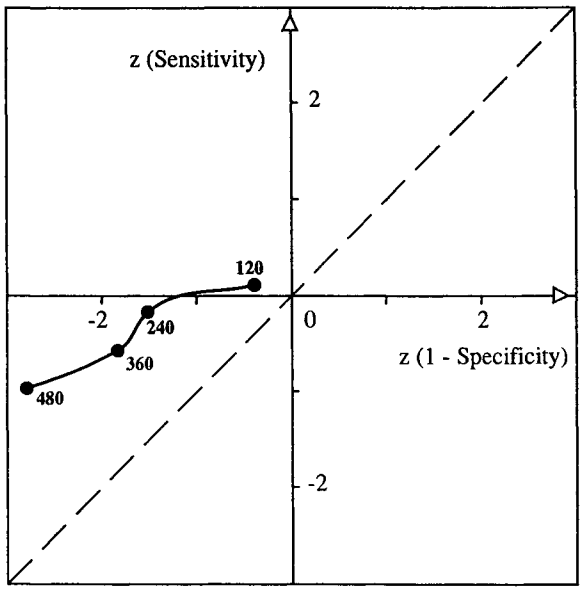


Fig. 16.3. Transformed ROC of SCK assay for myocardial infarction on double normal deviate (z) axes, from data by Radack et al. [7].

negative) and the smallest detectable difference δ of the test. In the present terminology we need to replace specificity by $1 - \alpha$, sensitivity by $1 - \beta$ and the decision limit by δ. For any given α and δ one can derive β, from the ROC. The power of the test is thus defined as $1 - \beta$.

De Ruig and van der Voet [12] have advocated the use of sensitivity and specificity as quality criteria for the analytical chemical laboratory (Chapter 13). They also pointed out that a trade-off is to be made between sensitivity and specificity. Increasing sensitivity (by lowering β) automatically leads to a decrease of specificity (by raising α) for any given smallest difference δ, as defined by the ROC of the test. From their point of view, the ROC defines the intra-laboratory standards for the test and forms part of the contract between the laboratory and its clients. Inter-laboratory quality standards can also be compared more easily by means of ROC.

16.2 Tests of hypothesis

16.2.1 Test of hypotheses for 2×2 contingency tables

We have already briefly outlined in Section 16.1 that observed frequencies in a 2×2 contingency table can be generated in three different ways, depending on the design of the experiment. In order to expand on the differences between these, we

TABLE 16.11

Observed frequencies in a 2×2 contingency table

Variable 1	Variable 2		Total
	Category 1	Category 2	
Category 1	n_{11}	n_{12}	n_{1+}
Category 2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

adopt the notation of Table 16.11. For the purpose of illustration, we assume that variable 1 is the outcome of an assay with categories positive and negative, and that variable 2 is a diagnosed disease with categories present and absent. Furthermore, we assume that all the observations which generated the 2×2 contingency table have been made independently from one another. The double subscripts of the symbols refer to the variables in the order given above. For example, n_{ij} is the element at the intersection of row i and column j of the table. A subscripted plus sign indicates that the sum has been taken over the categories of the corresponding variable. For example $n_{1+} = n_{11} + n_{12}$, etc. The size of the sample is the grand total of the table and is denoted in this section by n_{++}.

In a first type of design only the sample size n_{++} is fixed. In the context of clinical assays this occurs when a sample is selected at random from a population. In this case we find that the number of subjects with the disease present n_{+1} and the number of subjects without the disease n_{+2} are variable and subject to sampling error. Of course, the number of positives n_{1+} and the number of negatives n_{2+} are also variable and subject to random fluctuation. This case is called a *double dichotomy* [13].

A second type of design fixes one set of marginal totals. This arises when we select at random a predefined number of persons n_{+1} in which a disease has been diagnosed to be present and another predefined number of persons n_{+2} in which the disease is absent. Here, only the number of positives n_{1+} and the number of negatives n_{2+} is variable. In this design we test the *homogeneity* of the two samples, i.e. whether the proportion n_{11}/n_{+1} is equal to the proportion n_{12}/n_{+2}.

Finally, a third and rather uncommon design fixes all the marginal totals. We may imagine a design of this type in which an additional constraint is imposed. We inform the analyst in charge of the assay that he must produce exactly n_{1+} positive outcomes. The analyst is allowed to vary the decision limit of his assay such as to match the number of positive outcomes with the fixed n_{1+}. In this case we test the *independence* of the two variables, i.e. whether n_{11}/n_{21} equals n_{12}/n_{22}. In the

previously discussed HIV assay the decision limit may be based on the amount of precipitation produced by the antibody–antigen reaction. By lowering the decision limit below that recommended, one could increase the number of positive outcomes in the example to such an extent that the total number would be 300 instead of 291 (Table 16.1). This also decreases the number of negative outcomes from 309 to 300, since the total number of patients was fixed at 600. Although a design with all totals fixed is rather artificial, it is, nevertheless, the only one for which, strictly speaking, *exact probabilities* can be computed. It has been shown that these exact probabilities are the best possible approximations to those that arise in designs in which not all marginals are fixed [13].

16.2.2 Fisher's exact test for two independent samples

Fisher derived the exact probability of obtaining a given 2×2 contingency table from the hypergeometric distribution (Section 15.2) which assumes that all marginal totals are fixed:

$$P = \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{n_{11}!n_{12}!n_{21}!n_{22}!n_{++}!} \tag{16.20}$$

where the factorial function $k!$ denotes the consecutive product $1\,2\,3\,...(k-1)k$, and particularly where 0! equals 1.

A statistical test for the significance of a 2×2 contingency table is set up by considering all similar tables with the same marginal totals that have cell frequencies as extreme or more extreme than that observed [14]. In the so-called *Fisher's exact test* we reject the null-hypothesis of independence between the two variables if the sum of all the resulting probabilities is less than some predefined level of significance α. Fisher's test produces the best possible approximation to the exact probability when the strong assumption of fixed marginal totals is not met. In the laboratory, we most often find that only one set of the marginal totals is fixed. In this case we will test for homogeneity of a variable in two samples.

By way of illustration we consider Table 16.12 which relates the outcome of an assay (positive or negative) to a treatment (medication or control). The sizes of both treatment groups have been fixed to 10 by the design of the experiment. In a randomized design, subjects are assigned at random between the two samples. These two samples are called independent as they are composed of different subjects. After conclusion of the trial we find that 2 out of the 10 patients on medication still produce a positive outcome in the assay, against 5 of the 10 controls. The magnitude of the difference amounts to 30% in favour of treatment. Can we conclude that treatment had a significant effect on outcome? Or is the difference due to random variation?

TABLE 16.12

Observed outcomes of an assay in two treatment groups

Outcome	Treatment		Total (t)
	Medication (m)	Control (c)	
Positive (p)	2	5	7
Negative (n)	8	5	13
Total (t)	10	10	20

TABLE 16.13

Outcomes of an assay in two treatment groups that are more extreme than those observed in Table 16.12

Outcome	Treatment		Total (t)	Outcome	Treatment		Total (t)
	m	c			m	c	
Positive (p)	1	6	7	Positive (p)	0	7	7
Negative (n)	9	4	13	Negative (n)	10	3	13
Total (t)	10	10	20	Total (t)	10	10	20

Here we test the null hypothesis that the variable outcome is homogeneous in the two treatment categories. The alternative hypothesis is that the proportion of positive outcomes in the medication group is smaller than the one in the control group. Note that we test a one-sided hypothesis, as we are only interested in differences in one direction (i.e. less positive outcomes with medication). To this end we compute the exact probability of Table 16.12 (which we call $P1$) and of the ones that represent situations more extreme than that observed. The latter are shown in Table 16.13. Here, there are two cases which are more extreme than the one which we have observed, i.e., when the number of positive outcomes in the medication group is 1 or 0 instead of 2. (The corresponding probabilities are called $P2$ and $P3$).

The three relevant exact probabilities $P1$, $P2$ and $P3$ are computed as follows:

$$K = \frac{10!\ 10!\ 7!\ 13!}{20!}$$

$$P1 = \frac{K}{2!\ 5!\ 8!\ 5!} = 0.1463$$

$$P2 = \frac{K}{1!\ 6!\ 9!\ 4!} = 0.0271$$

$$P3 = \frac{K}{0!\ 7!\ 10!\ 3!} = 0.0015$$

In order to compute the probability of obtaining the observed case and all cases which are more extreme, we have to add $P1 + P2 + P3$ which yields 0.1749. Hence if we have fixed the level of significance in a one-sided test α at 0.05 then clearly we have to conserve the null hypothesis. In this case, we cannot accept the alternative hypothesis that the less frequent positive outcomes are due to the effect of the medication, although there is a tendency for the patients who received medication to be better off than those in the control group. Perhaps failure to detect a real effect of the observed magnitude is due to the limited size of the sample that has been studied. In other words, our test lacks sufficient power to detect a difference of 30% at the 0.05 level of significance. The concept of power has been introduced in Section 16.1.3 in relation to the sensitivity of a test, i.e. the power to detect a real effect. If we had enrolled more patients in the study, then perhaps the result might have turned out to be significant, provided that the observed effect is not due to random events.

As the number of discrete probabilities that are to be calculated increases with the size of the sample n_{++}, Fisher's exact test has been usually reserved for small samples. But this is a practical rather than a theoretical constraint. As we have already pointed out, Fisher's test is the best choice for testing hypotheses about 2×2 contingency tables, even if its strict assumption of fixed marginal totals is rarely met [13]. Although the test is essentially for one-sided hypotheses, it can be extended to handle two-sided hypotheses as well.

16.2.3 Pearson's χ^2 test for two independent samples

If we reconsider the observed outcomes of an assay in the two treatment groups in Table 16.12 we might ask what values we would expect if we knew in advance that there was no difference at all between the two treatments. In other words, is it possible to calculate expected values for the elements of a 2×2 contingency table under the null hypothesis. It can readily be seen that the expected number of positive outcomes in the medication group must be proportional to the corresponding marginal totals, i.e. the total number of patients in the medication group (10), the total number of patients with positive outcomes (7) and to the total number of patients in the study (20). In what follows we will approach this problem in a formal way.

Under the assumption of independence between the variables of a contingency table, we can express the *maximum likelihood estimate* of any cell frequency as the product of its corresponding marginal totals [13]:

$$E(n_{ij}) = \frac{n_{i+}\, n_{+j}}{n_{++}} \qquad \text{with } i = 1, 2 \text{ and } j = 1, 2 \tag{16.21}$$

where $E(n_{ij})$ means *expected value* of n_{ij} and where the marginal totals n_{i+}, n_{+j} and n_{++} have been defined before. The assumption of independence implies that all marginal totals are fixed, a situation that is rarely met in practice. Notwithstanding this limitation, expected values of a 2×2 contingency table are most often calculated by eq. (16.21), as if all marginal totals are fixed. As mentioned already, the practical consequences of the violation of the strict assumption are minimal.

Using the result of eq. (16.21) we can derive *Pearson's χ^2 statistic for goodness of fit* between observed and expected values of a 2×2 contingency table:

$$\chi^2 = \sum_{i}^{2} \sum_{j}^{2} \frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})} \tag{16.22}$$

which possesses one degree of freedom (df). Note that a 2×2 contingency table with fixed marginal totals possesses only one degree of freedom. This means that, given one of the four cell frequencies, we can derive the other three cell frequencies using the fixed marginal totals.

After substitution of expected values from eq. (16.21) into the expression of χ^2 we obtain [14]:

$$\chi^2 = \frac{(n_{11}\, n_{22} - n_{12}\, n_{21})^2 n_{++}}{n_{1+}\, n_{+1}\, n_{2+}\, n_{+2}} \qquad \text{with df} = 1 \tag{16.23}$$

The χ^2 statistic, as a measure of goodness of fit, has also been applied in the test of normality described in Section 5.6.

We apply the above expression of eq. (16.23) to the data in Table 16.12 in order to test the homogeneity of outcomes (positive or negative) in the two treatment groups (medication or control):

$$\chi^2 = \frac{(2 \times 5 - 5 \times 8)^2 20}{7 \times 13 \times 10 \times 10} = 1.978 \qquad \text{with df} = 1$$

From tabulated values of the χ^2 distribution function we can determine the probability of obtaining a value of χ^2 as large or larger than the one observed. More conveniently, we can make use of the property that χ^2 with one degree of freedom is distributed as the square of the standard normal deviate z. Hence, we determine:

$$z = \sqrt{\chi^2} = \sqrt{1.978} = 1.406$$

and we look up the corresponding one-sided probability in a table of the standard normal distribution function, which yields that $p = 0.0798$.

The one-sided probability of Pearson's χ^2 test statistic is at variance with Fisher's exact probability, which produced $p = 0.1749$ (one-sided). The lack of agreement is attributed to the fact that the distribution of Pearson's χ^2 is continuous, whereas Fisher's exact probabilities are derived from the discrete hypergeometric

distribution. Yates has proposed a *correction for continuity* which tends to make probabilities obtained in Pearson's χ^2 test conform more to those derived from Fisher's exact test [14]:

$$\chi^2 = \frac{(|n_{11}\, n_{22} - n_{12}\, n_{21}| - n_{++}/2)^2 n_{++}}{n_{1+}\, n_{+1}\, n_{2+}\, n_{+2}} \qquad \text{with df} = 1 \tag{16.24}$$

The correction proposed by Yates is recommended when the sample size is small. If Yates' correction is applied in our example, we obtain:

$$\chi^2 = \frac{(|2 \times 5 - 5 \times 8| - 10)^2 20}{7 \times 13 \times 10 \times 10} = 0.8791 \qquad \text{with df} = 1$$

and

$$z = \sqrt{0.8791} = 0.9376$$

From the table of the standard normal distribution we find that the probability of Pearson's χ^2 with correction for continuity is 0.1742 (one-sided), whereas Fisher's test produced an exact probability of 0.1749 (one-sided). In this example, the difference is negligible.

Cochran [15] has proposed a set of rules which may help in deciding between Fisher's exact test and Pearson's χ^2 test. Cochran's diagram indicates that Fisher's exact probability test is to be preferred above the corrected χ^2 test in the case of small sample sizes ($n_{++} \leq 20$) and in the case of near-zero cell frequencies ($n_{ij} < 5$). These conditions are presented schematically in Fig. 16.4. It is observed from eqs.

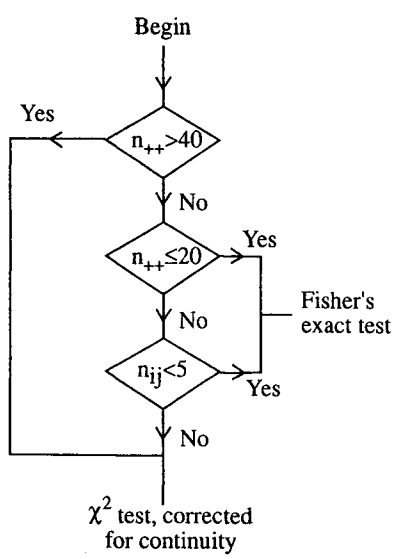


Fig. 16.4. Selection of a test of hypothesis for two independent samples according to Cochran [15].

(16.23) and (16.24) that the corrected χ^2 is always smaller than the uncorrected one. Hence, the corresponding probability of the corrected χ^2 is always larger. For this reason it is stated that the correction is *conservative*, as it tends to preserve the null hypothesis more often than the uncorrected test statistic. More recently, the suitability of Yates' correction has been questioned [16] especially in the case of small sample sizes, although many authors still recommend its use [1]. However, due to increased availability of computational resources, there is a trend towards recommending Fisher's exact test throughout, e.g. for compliance to the guidelines for good laboratory practices [17].

In summary, Fisher's test is only 'exact' in the sense that it correctly derives probabilities for discrete cases, whereas the χ^2 statistic applies to continuous cases. Strictly speaking, Fisher's test requires that all marginal totals be fixed, an assumption which is rarely met in practice. Nevertheless, Fisher's exact test is regarded as the best choice, even when the strict assumption of fixed marginals is not satisfied, and certainly in the case of small sample sizes and near-zero cell frequencies.

16.2.4 Graphical χ^2 test for two independent samples

In practice it often happens that a large battery of assays and observations is performed on the same pair of independent samples, for example in the comparison of various effects of a treatment in a medication and a control group. This results in a large number of tests for homogeneity of outcomes in the two samples. The statistical tests can be performed graphically by means of the so-called '*elevation-contrasts' diagram* [18]. On the vertical and horizontal axes of the diagram in Fig. 16.5 we represented the proportion of positive outcomes of the SCK assay for myocardial infarction, as described in Section 16.1.7 on decision limits. The vertical axis thus represents sensitivity and the horizontal axis is defined by (1 – specificity). Both axes are logarithmic in order to allow for a wide range of variation. Each point in the diagram corresponds to a particular assay or observation.

In the 'elevation-contrasts' diagram one has to focus on the position of a point relative to the diagonal line. This line represents the case of complete homogeneity of outcomes in the two treatment groups. It is the line of zero contrast, where contrast is to be understood in the sense of difference or heterogeneity. Points above the line correspond to more positive outcomes in the medication group than would be expected under the hypothesis of homogeneity. These points reflect positive contrasts. Points below the line correspond to less positive (or more negative) outcomes in the control group than can be expected under the null hypothesis. These points possess negative contrasts. The further away a point is from the diagonal, the stronger is its contrast. The curved contours are the significance

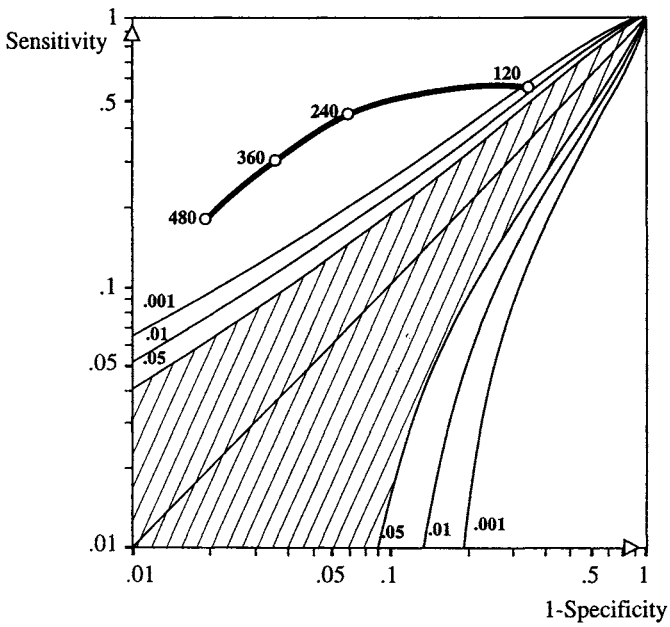


Fig. 16.5. 'Elevation-contrasts' diagram for a battery of assays applied to a medication and control group. Graphical χ^2 test for two independent samples. Data are those of Table 16.9 (adapted from Radack et al. [7]).

bands outside which the hypothesis of homogeneity is rejected at a predefined level of significance using Pearson's χ^2 test for 2 independent samples corrected for continuity (with two-sided $\alpha = 0.05$, 0.01 and 0.001).

Points that project on the high side of the diagonal have a high average rate of positive outcomes. Points that project on the low side of the diagonal possess a low average rate. The larger the average rate of positive outcomes in both treatment groups, the higher is the *elevation* of the corresponding point along the direction of the diagonal. The 'elevation-contrast' diagram allows us to visualize the significance of contrasts together with the magnitudes of the outcomes in a battery of tests. Hence its name 'elevation-contrasts' diagram. From Fig. 16.5 it appears that an optimal choice for the decision level on the scale of SCK may be chosen between 360 and 480 IU/ml. This will ensure high sensitivity and high specificity together with a manageable number of false positives. The positions of the significance bands on the 'elevation-contrasts' diagram depend on the sample size. Hence, the contours must be recomputed for each new experiment, but this is a practical rather than a theoretical difficulty. A problem arises, however, because multiple comparisons are made on the same two samples. The likelihood of obtaining a significant result by chance is thus considerably augmented. One may

account for this effect of multiple comparisons by lowering the level of significance in proportion to the number of comparisons. This procedure is referred to as a Bonferroni test (Section 5.2) [19].

16.2.5 Large-sample χ^2 test statistic for two independent samples

In Section 16.2.3 we have shown that a 2×2 contingency table possesses only one degree of freedom provided that all marginal totals are fixed. This allows us to focus on one particular cell frequency, say n_{11}, since all three others can be derived from it using the marginal totals. For example, n_{12} equals $(n_{1+} - n_{11})$, etc.

We define the difference between the observed and expected value $(O - E)$ of a cell frequency, say n_{11}, using eq. (16.21):

$$O - E = n_{11} - \mathrm{E}(n_{11}) = n_{11} - \frac{n_{1+}\, n_{+1}}{n_{++}} \tag{16.25}$$

Under the assumption of fixed marginal totals we obtain the variance V of any cell frequency from the hypergeometric distribution [13]:

$$V = \frac{n_{1+}\, n_{+1}\, n_{2+}\, n_{+2}}{n_{++}^2 (n_{++} - 1)} \tag{16.26}$$

From eqs. (16.25) and (16.26) we can derive a test statistic for large values of the sample size n_{++}:

$$z^2 \approx \chi^2 = \frac{(O - E)^2}{V} = \frac{(n_{11}\, n_{22} - n_{12}\, n_{21})^2\, (n_{++} - 1)}{n_{1+}\, n_{+1}\, n_{2+}\, n_{+2}} \qquad \text{with df} = 1 \tag{16.27}$$

Note that this result is asymptotically equivalent (when n_{++} becomes large) to Pearson's χ^2 test statistic (eq. (16.23)).

16.2.6 McNemar's χ^2 test statistic for two related samples

A special situation occurs when the same sample is assayed on two occasions. For example, the outcome of an assay is recorded in a sample of subjects at the beginning and at the end of a period of treatment. Note that Pearson's χ^2 test and Fisher's exact test are not applicable here since the observations have not been obtained independently from each other. We assume that n_{pp} subjects remained positive from the beginning to the end of the period. Likewise, n_{nn} subjects remained negative. But n_{pn} subjects which were initially positive changed to negative, and n_{np} subjects changed from negative to positive. This gives rise to the 2×2 contingency Table 16.14.

TABLE 16.14

2×2 contingency table for two related samples

Before	After		Total
	Positive	Negative	
Positive	n_{pp}	n_{pn}	n_{p+}
Negative	n_{np}	n_{nn}	n_{n+}
Total	n_{+p}	n_{+n}	n_{++}

We test the null hypothesis that the changes from positive to negative and *vice versa* are due to chance. Under this hypothesis we expect that the frequencies n_{pn} and n_{np} are equal, apart from random variation. Under the null hypothesis of no change we must have that the expected values for the changes are equal [14]:

$$E(n_{pn}) = E(n_{np}) = \frac{n_{pn} + n_{np}}{2} \tag{16.28}$$

We now set up the χ^2 test statistic for goodness of fit for the two occurrences of interest (eq. (16.22)):

$$\chi^2 = \frac{(n_{pn} - E(n_{pn}))^2}{E(n_{pn})} + \frac{(n_{np} - E(n_{np}))^2}{E(n_{np})} \tag{16.29}$$

which becomes after substitution of expected values by eq. (16.28):

$$\chi^2 = \frac{(n_{pn} - n_{np})^2}{n_{pn} + n_{np}} \qquad \text{with df} = 1 \tag{16.30}$$

Note that in eq. (16.28) we have only considered the *discordances* (n_{pn} and n_{np}) and have deliberately disregarded the *concordances* (n_{pp} and n_{nn}).

McNemar's χ^2 test statistic can be corrected for continuity:

$$\chi^2 = \frac{(|n_{pn} - n_{np}| - 0.5)^2}{n_{pn} + n_{np}} \qquad \text{with df} = 1 \tag{16.31}$$

By way of example, we consider the case of 10 subjects which have been observed at the beginning and at the end of a period of treatment, the outcomes (positive or negative) of which are shown in Table 16.15. Before treatment, 6 patients were found to be positive, of which 2 remained positive and of which 4 changed to negative. Before treatment, 4 patients were originally found to be negative, of which 3 remained negative and 1 changed to positive. McNemar's χ^2 test statistic according to eq. (16.30) for this case is:

TABLE 16.15

Outcomes of an assay before and after treatment

Before	After		Total
	Positive	Negative	
Positive	2	4	6
Negative	1	3	4
Total	3	7	10

$$\chi^2 = \frac{(4-1)^2}{4+1} = \frac{9}{5} = 1.80 \qquad \text{with df} = 1$$

$$z = \sqrt{1.80} = 1.342 \text{ and } p = 0.090 \qquad \text{(one-sided)}$$

and with correction for continuity according to eq. (16.31):

$$\chi^2 = \frac{(|4-1| - 0.5)^2}{4+1} = \frac{6.25}{5} = 1.250 \qquad \text{with df} = 1$$

$$z = \sqrt{1.250} = 1.118 \text{ and } p = 0.154 \qquad \text{(one-sided)}$$

On the basis of the probability we must retain the null hypothesis that there is no change in outcome between the start and the end of treatment at the one-sided level of significance α of 0.05.

16.2.7 Tetrachoric correlation

The *tetrachoric correlation* coefficient r has been proposed as a measure of association between two dichotomous variables, for example outcome of treatment and outcome without treatment:

$$r = \sqrt{\frac{\chi^2}{n_{++}}} \tag{16.32}$$

where χ^2 is Pearson's test statistic for two independent samples and where n_{++} is the total sample size [20]. It may be considered as an estimate of the size of the effect produced by treatment. Numerically, the tetrachoric correlation coefficient is identical to the product–moment correlation coefficient computed from the two dichotomous variables. Using the corrected χ^2 obtained in the previous example we compute from eq. (16.32):

$$r = \sqrt{\frac{1.250}{10}} = 0.354$$

Unfortunately, the sampling distribution of the tetrachoric correlation coefficient cannot be obtained in a simple way. Hence, it is difficult to derive its variance and to produce the corresponding confidence interval. For this reason, the tetrachoric correlation coefficient is now only of historic interest [13]. A more tractable estimate of effect size will be discussed in the section devoted to the odds ratio.

16.2.8 Mantel–Haenszel χ^2 test statistic for multiple 2×2 contingency tables

Often a 2×2 contingency table is the result of pooling of outcomes from several samples. For example, in a study of the effect of a medication one may pool results obtained in subjects from different categories of gender and age. Although simple pooling increases the size of the sample and hence improves the statistical power of tests of hypotheses, it may also lead to biased conclusions. To illustrate this point we consider a hypothetical case where two independent studies (study I and study II in Table 16.16) each involving 110 subjects have been pooled into a single large study (Table 16.17) with a pooled sample size of 220 subjects.

We can also represent this case in a graphical way. The vertical axis of Fig. 16.6 represents the proportion of positive outcomes in the medication group $n_{11}/(n_{11} + n_{21})$ and in the control group $n_{12}/(n_{12} + n_{22})$. For example, in study I the proportions of

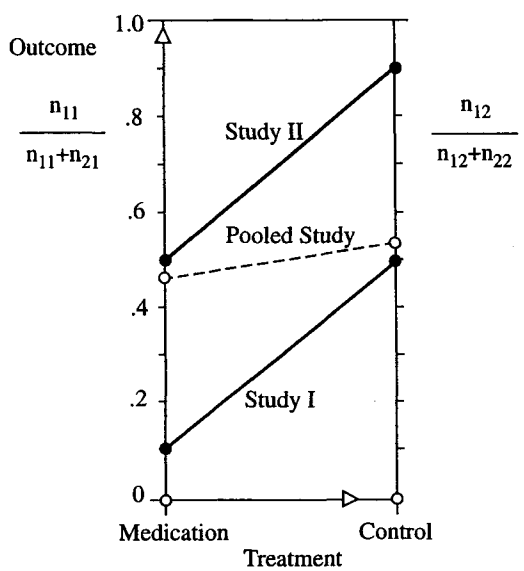


Fig. 16.6. Outcome of an assay (positive or negative) as a function of treatment (medication and control) in two studies labelled I and II. The slope of the response line is a measure for the effect of treatment. The difference in elevation is due to an effect of study. The response line of the pooled study is severely biased.

TABLE 16.16

Two independent studies of the outcome (positive or negative) of an assay as a result of treatment (medication or control)

Outcome	Study I		Total	Outcome	Study II		Total
	Med.	Contr.			Med.	Contr.	
Positive	1	50	51	Positive	50	9	59
Negative	9	50	59	Negative	50	1	51
Total	10	100	110	Total	100	10	110

TABLE 16.17

Pooled study of the outcome (positive or negative) of an assay as a result of treatment (medication or control)

Outcome	Pooled study		Total
	Med.	Contr.	
Positive	51	59	110
Negative	59	51	110
Total	110	110	220

positive outcomes are 1/10 = 0.1 and 50/100 = 0.5 for the medication and control groups, respectively.

From the slopes of the response lines we can judge that the individual studies I and II show a marked *effect of treatment*. The proportion of positive outcomes in both studies is 40% less in the treatment group in comparison to the control group. However, the pooled result indicates hardly any difference between the two forms of treatment, as the slope of the corresponding response line is almost flat. Note that the control group in study I and the medication group in study II are given a 10 times larger weight in the pooling, as appears from Table 16.16.

If we apply the large-sample χ^2 test statistic (eq. (16.27)) to the pooled data in Table 16.17 we obtain:

$$O - E = n_{11} - \frac{n_{1+}\, n_{+1}}{n_{++}} = 51 - \frac{110 \times 110}{220} = -4$$

$$V = \frac{n_{1+}\, n_{+1}\, n_{2+}\, n_{+2}}{n_{++}^2 (n_{++} - 1)} = \frac{110 \times 110 \times 110 \times 110}{220^2 \times (220 - 1)} = 13.813$$

$$\chi^2 = \frac{(O - E)^2}{V} = \frac{(-4)^2}{13.813} = 1.158 \qquad \text{df} = 1, \qquad p = 0.141$$

TABLE 16.18

Statistical results from individual and combined studies from data given in Tables 16.16 and 16.17. Probabilities are one-sided.

Study	N	$O-E$	V	χ^2	p
I	110	–3.636	2.281	5.795	0.008
II	110	–3.636	2.281	5.795	0.008
Combined	220	–7.272	4.562	11.591	0.0002

This result is not significant at the one-sided level of significance α of 0.05. The results of the individual studies, on the contrary, are highly significant as is shown in Table 16.18 ($p = 0.008$, one-sided). The paradox arises, on the one hand, from an unbalanced allocation of subjects to treatment groups. On the other hand there is a large difference between the proportions of positive outcomes in the two studies, as is shown clearly by Fig. 16.6. This latter phenomenon is called the *effect of study*. The effect of study may severely bias an estimate of the effect of treatment when data are simply pooled as we have done in this hypothetical case. We state that the effect of treatment is confounded by the effect of individual studies, or, in other words, that type of study is a *confounding factor* of the effect of treatment.

Mantel and Haenszel [21] have proposed a test which accounts for effects of study when combining independent 2×2 contingency tables. We use the term 'combining' to indicate a procedure which avoids the bias produced by simple 'pooling' of the data. Their approach is to combine the individual observed minus expected frequencies into $(O-E)_c$ and to combine the corresponding variances into V_c:

$$(O-E)_c = \sum_k (O-E)_k \tag{16.33}$$

$$V_c = \sum_k V_k \tag{16.34}$$

where the summation extends over all k individual 2×2 contingency tables.

We know that V_c is the variance of $(O-E)_c$, provided that the individual contingency tables are independent. This consideration leads to the *Mantel–Haenszel* χ^2 *test statistic* which possesses one degree of freedom:

$$\chi^2 = \frac{(O-E)_c^2}{V_c} \qquad \text{with df} = 1 \tag{16.35}$$

The result of the Mantel–Haenszel test in the case of the two independent studies is shown on the bottom line of Table 16.18. It is even more significant than in each of the individual ones. It is also quite different from the biased result which we have obtained from simple pooling of the individual studies.

16.2.9 Odds ratio

So far, we have been discussing tests of hypotheses about 2×2 contingency tables. It must be realized, however, that a statistically significant result does not necessarily correspond in practice to a relevant effect. Indeed, for any negligibly small effect one can find a sufficiently large sample size for which this irrelevant effect will become statistically significant. It is felt therefore that a statistical analysis of an effect should provide not only the probability of its occurrence under some hypothesis, but also provide an estimate of the size of this effect, together with its 95% confidence interval [3].

A common measure of effect in a 2×2 contingency table is the *odds ratio* which is defined as follows in the notation introduced in Section 16.2.1:

$$\text{Odds Ratio} = OR = \frac{n_{11}\, n_{22}}{n_{21}\, n_{12}} = \frac{\text{Odds}\,(n_{11}/(n_{11}+n_{21}))}{\text{Odds}\,(n_{12}/(n_{12}+n_{22}))} \tag{16.36}$$

Note that the expected value of the odds ratio OR is 1 under the assumption of homogeneity. In a comparison of two treatment groups, we can interpret the odds ratio as the odds of obtaining a positive outcome in the treatment group divided by the odds of obtaining a positive outcome in the control group (eq. (16.36)). In this sense the odds ratio can be regarded as a measure of the size of the effect produced by treatment.

Another interpretation can be derived from the notation used in Section 16.1.3 for sensitivity and specificity:

$$OR = \frac{\text{tp tn}}{\text{fn fp}} = \frac{\text{Odds}\,(\text{tp}/(\text{tp}+\text{fn}))}{\text{Odds}\,(\text{fp}/(\text{fp}+\text{tn}))} = \frac{\text{Odds (Sensitivity)}}{\text{Odds}\,(1-\text{Specificity})} \tag{16.37}$$

We find here a striking resemblance between the expression for the odds ratio OR (eq. (16.37)) and the one which we have derived above for the likelihood ratio LR (eq. (16.13)):

$$LR = \frac{\text{Sensitivity}}{1-\text{Specificity}}$$

In Section 16.1.8 on receiver operating characteristics (ROC) we have made a graphical interpretation of the likelihood ratio. The slope of the line which joins the origin to a point on the ROC curve equals the likelihood ratio of the assay at the corresponding decision limit (Fig. 16.2). We derive a similar interpretation of the odds ratio from an ROC plot in which the horizontal and vertical scales are expressed in odds rather than proportions. The transformation from odds to proportions has already been defined by eq. (16.15). Figure 16.7 represents the transformed ROC curve of the SCK assay for myocardial infarction [7] which we discussed in Subsection 16.1.7. Note that the vertical and horizontal axes now

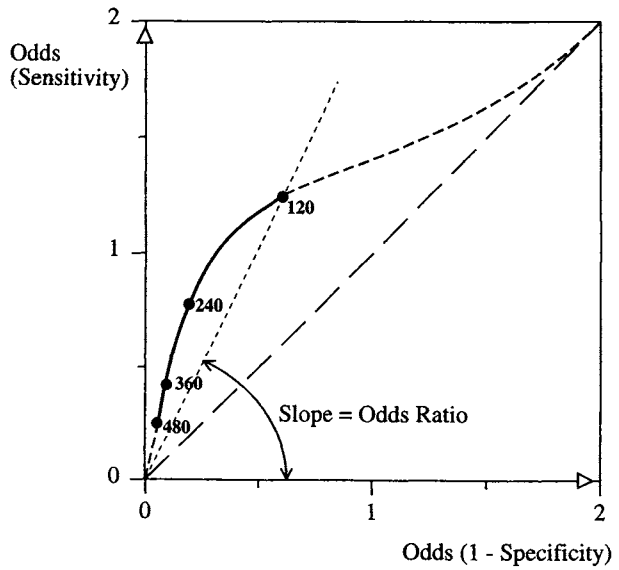


Fig. 16.7. Receiver operating characteristic curve (ROC) of SCK assay for myocardial infarction (data by Radack et al. [7]) on double odds axes.

express odds of sensitivity and odds of (1 – specificity), respectively. The slope of the line through the origin and a particular point of the curve equals the odds ratio of the assay at the corresponding decision limit.

16.2.10 Log odds ratio

For reasons that will become evident hereafter, it is more convenient to work with the natural logarithms of odds, also called *logits*. According to the definition of odds (eq. (16.15)) we obtain for any probability or proportion P:

$$\text{Log Odds}\,(P) = \ln\left(\frac{P}{1-P}\right) = \text{logit}(P) \tag{16.38}$$

Using natural logarithms in eq. (16.36) we obtain a definition for the log odds ratio β:

$$\begin{aligned}\beta = \text{Log Odds Ratio} &= \ln\left(\frac{n_{11}}{n_{21}}\right) - \ln\left(\frac{n_{12}}{n_{22}}\right) \\ &= \text{logit}\left(\frac{n_{11}}{n_{11}+n_{21}}\right) - \text{logit}\left(\frac{n_{12}}{n_{12}+n_{22}}\right)\end{aligned} \tag{16.39}$$

In a study of the effect of a treatment (medication or control) on the outcome of an assay (positive or negative) we can interpret the log odds ratio β geometrically.

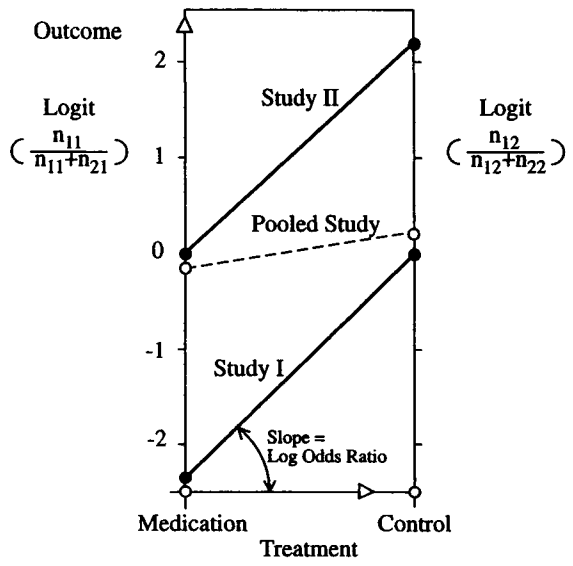


Fig. 16.8. Outcome of an assay (positive or negative) as a function of treatment (medication and control) in two studies labelled I and II. The vertical scale is expressed in logits. The slope of the response line is equal to the log odds ratio β and is a measure for the effect of treatment.

In Fig. 16.8 we have represented the dichotomous variable treatment along the horizontal axis. The variable outcome is represented along the vertical axis on a logit scale.

From eq. (16.39) follows immediately that the slope of the response line is equal to the log odds ratio β. It is natural therefore to consider β as a measure of effect size, which can be readily derived from a 2×2 contingency table. Woolf has shown that the log odds ratio is approximately normally distributed with variance as defined below [3]:

$$V(\beta) \approx \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \tag{16.40}$$

provided that none of the cell frequencies is zero.

Woolf's formula (eq. (16.40)) allows us to define approximately the lower (l) and upper (u) limits of the confidence interval:

$$\beta_{l,u} \approx \beta \pm z_{\alpha/2} \sqrt{V(\beta)} \tag{16.41}$$

where $z_{\alpha/2}$ is the standard normal deviate that corresponds with the one-sided probability $\alpha/2$. By convention, the level of significance is often defined at 0.05, 0.01 or 0.001. We are 100 $(1 - \alpha)$% confident that the log odds ratio of the population from which the sample has been drawn lies within the limits of the confidence interval. If zero is outside the confidence interval then we reject the null

hypothesis of homogeneity at the two-sided significance level α. Otherwise, if the confidence interval contains zero then we retain the null hypothesis.

After exponentiation of β, together with its lower (l) and upper (u) limits, we obtain the estimate of the odds ratio and its confidence interval:

$$OR = \exp(\beta) \qquad \text{and} \qquad OR_{l,u} = \exp(\beta_{l,u}) \tag{16.42}$$

In the case of odds ratios we reject the null hypothesis at the level of significance α, if 1 lies outside the confidence limits. Otherwise, if the confidence interval contains 1 then we retain the null hypothesis.

By way of illustration, we compute the odds ratio and Woolf's approximate confidence interval for the data of Table 16.12. In this example we related outcome of an assay (positive or negative) to type of treatment (medication or control) in 20 subjects:

$$OR = \frac{2 \times 5}{8 \times 5} = 0.25 \qquad \text{and} \qquad \beta = \ln(OR) = \ln(0.25) = -1.386$$

$$V(\beta) = \frac{1}{2} + \frac{1}{5} + \frac{1}{8} + \frac{1}{5} = 1.025$$

$$\beta_{l,u} = -1.386 \pm 1.96 \sqrt{1.025} = (-3.370, 0.598)$$

where the factor 1.96 is the standard normal deviate $z_{0.025}$. Since zero is contained within the 95% confidence interval, we retain the null hypothesis of homogeneity. On the basis of this experiment, we have no reason to assume that the proportion of positive outcomes is different in the two treatment groups. After exponentiation we obtain the confidence interval of the odds ratio:

$$OR_{l,u} = \exp(-3.370, 0.598) = (0.034, 1.818)$$

Because the value of 1 is within the confidence interval, we retain the null hypothesis of homogeneity.

The log odds ratio leads to yet another variant of the ROC diagram in which the vertical and horizontal axes represent logits of sensitivity and logits of (1 – specificity), respectively (Fig. 16.9). When sensitivity and specificity are distributed according to the *logistic distribution* function, then ROC curves are transformed into straight lines in a diagram of double logit axes. This property resembles that of the ROC diagram with double normal deviate axes in Fig. 16.3. The quality of an assay operating at a given detection limit can now be estimated from the distance between the corresponding point on the ROC curve and the diagonal line. We show that this distance is equal to the log odds ratio of the assay. Taking natural logarithms on both sides of eq. (16.37) produces:

$$\beta = \ln(OR) = \text{logit(Sensitivity)} - \text{logit}(1 - \text{Specificity}) \tag{16.43}$$

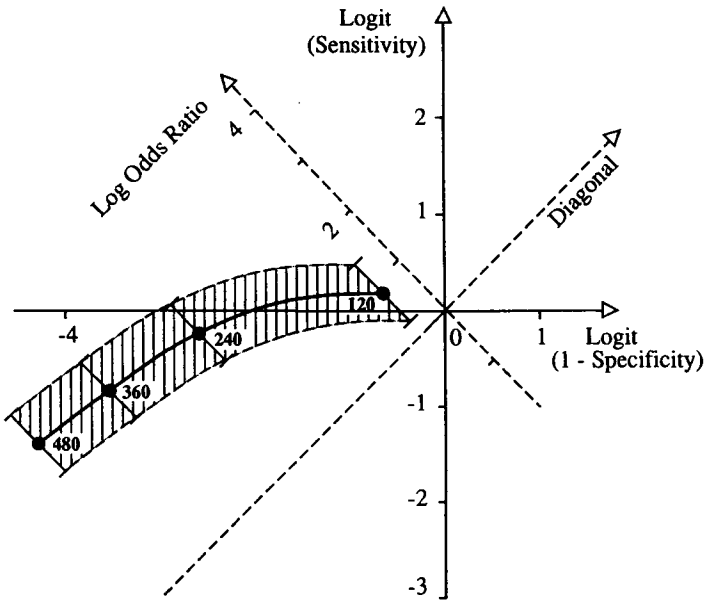


Fig. 16.9. Receiver operating characteristic (ROC) diagram with double logit axes. The distance of a point on the curve from the diagonal is equal to the log odds ratio. Log odds ratios can be read off directly from the axis which is drawn perpendicularly to the diagonal line. The intervals shown are the 95% confidence intervals for the log odds ratios.

Let us assume that X represents logits of sensitivity and that Y represents logits of (1 – specificity). The contrast $X - Y$ is equal to β, according to eq. (16.43). It is measured along a line through the point defined by X and Y and drawn perpendicularly to the diagonal line. Points with zero contrast are located on the diagonal line. This shows that β is estimated as the distance of a point to the diagonal line in the double logit ROC diagram. Using Woolf's approximation to the variance of β (eq. (16.40)) we can also display the 95% confidence interval along the same perpendicular on which β is measured. The shaded area in Fig. 16.9 has been drawn through the endpoints of the 95% confidence intervals of the four experimental points.

The diagonal line represents the geometrical average of the odds for a positive outcome in either of the treatment groups. If the odds in both treatment groups are high then the corresponding point will be found high up in the direction of the diagonal line. In this case we state that the elevation (along the diagonal) is large. Note also that the double logit ROC diagram of Fig. 16.9 is similar to the 'elevation-contrasts' diagram of Fig. 16.5. In both diagrams the diagonal line expresses the average positive outcome (elevation). Both diagrams also represent the differential outcomes in two treatment groups (contrasts). The main difference is that Fig. 16.5 is constructed from log proportions, while Fig. 16.9 is expressed in terms of log odds. The former uses tests of hypotheses, while the latter displays

confidence intervals. Both diagrams suggest that an optimal choice for the decision level is between 240 and 480 IU/ml on the scale of SCK.

16.2.11 Multiple 2×2 contingency tables, meta-analysis

Peto has extended Mantel–Haenszel's approach to the analysis of multiple 2×2 contingency tables to include effect size [22]. It is assumed that the different tables have been observed independently of each other. This method makes straightforward use of the log odds ratio β.

A maximum likelihood estimate for the log odds ratio β has been derived by Peto under the assumption of homogeneity of effects produced by a treatment in two groups. We refer to this estimate by means of the symbol β′ and to the corresponding odds ratio by OR':

$$\beta' = \ln OR' = \frac{O - E}{V} \tag{16.44}$$

where the variance V of 'observed minus expected' cell frequencies $(O - E)$ is derived from the hypergeometric distribution and is given by eq. (16.26).From the above equation we can derive the variance of β′:

$$V(\beta') = \frac{1}{V} \tag{16.45}$$

Using these results we can now define an approximate 100 $(1 - \alpha)$% confidence interval for β′ with lower (l) and upper (u) limits given by:

$$\beta'_{l,u} = \frac{O - E}{V} \pm z_{\alpha/2} \sqrt{\frac{1}{V}} \tag{16.46}$$

In the case of multiple 2×2 contingency tables we apply Mantel–Haenszel's approach which has been described in Section 16.2.8. Here, too, we accumulate the $(O - E)$ and V values from the individual 2×2 contingency tables into a combined $(O - E)_c$ and a combined V_c. Since we may add up individual variances provided that the individual tables have been observed independently from each other, we obtain:

$$\beta'_c = \frac{(O - E)_c}{V_c} \tag{16.47}$$

$$V(\beta'_c) = \frac{1}{V_c} \tag{16.48}$$

$$(\beta'_c)_{l,u} = \frac{(O - E)_c}{V_c} \pm z_{\alpha/2} \sqrt{\frac{1}{V_c}} \tag{16.49}$$

TABLE 16.19

Result of meta-analysis using the fixed effects model. Combined analysis of individual studies presented in Table 16.16. The symbol CI means confidence interval.

Study	$O-E$	V	β' (95% CI)	OR' (95% CI)
I	-3.64	2.28	-1.59 (-2.91, -0.28)	0.20 (0.05, 0.75)
II	-3.64	2.28	-1.59 (-2.91, -0.28)	0.20 (0.05, 0.75)
Combined	-7.27	4.56	-1.59 (-2.51, -0.68)	0.20 (0.08, 0.51)

By means of exponentiation (in base e) we can transform eq. (16.49) into the maximum likelihood estimates for the odds ratio OR' and its associated confidence interval.

We can compute Peto's maximum likelihood estimates of the odds ratio for the case which has been presented in Table 16.16. This analysis shows that effect sizes of the two studies (I and II) as well as that of the combined study are equal ($OR' = 0.203$). The confidence intervals show that the effect is significant at the level of significance α of 0.05 in the individual and combined studies (Table 16.19). Note that the value of 1 is outside the 95% confidence intervals of OR'. The confidence interval of the combined study, however, is smaller than that of the individual studies which demonstrate the utility of combining studies into a single large one using appropriate statistical methods.

Meta-analysis, or overview of studies as it is called by Peto, is the combination of 2×2 contingency tables that result from experiments in which two treatment groups are selected at random [22]. As meta-analysis combines information from several independent sources, it is deemed to be good scientific practice. A similar argument has been presented in favour of the Bayesian approach against traditional hypothesis testing in Section 16.1.5. In order to avoid *recall bias* it is mandatory that all available studies are included. Serious recall bias occurs, for example, when unfavourable results are not included in the meta-analysis.

Meta-analysis is different from so-called *omnibus tests*, such as the *log P test* of Fisher [23]. In the log P test one adds up the natural logarithms of the one-sided probabilities of the individual studies. After multiplication by the constant -2, this produces a χ^2 test statistic for the significance of the combined study. The number of degrees of freedom is equal to twice the number of individual studies in a one-sided test of significance. While an omnibus test only produces the significance, meta-analysis also estimates the size of the combined effect and provides a confidence interval for it.

Peto's approach to meta-analysis assumes that effects of study are fixed, i.e. not subject to random fluctuations, although they can differ from one study to another.

For this reason, the approach is called the fixed effects model of meta-analysis [22]. A random effects model for meta-analysis has been described by DerSimonian and Laird [24]. A similar distinction between fixed and random effects models arises in the analysis of variance (ANOVA) and is discussed in Sections 6.3 and 6.4.

16.2.12 Logistic regression, confounding, interaction

In the previous section we have discussed the case of a pooled 2×2 contingency table in which the effect of treatment was confounded by the type of study. In data gathered from subjects whose outcome (positive or negative) was studied under different treatments (medication or control), we can also regard age and gender as possible confounding factors. In such a case of multiple confounding factors we can set up a *linear logit model* [25] which relates outcome to treatment, age, gender, time of study and so on. More generally, in the case of independent variables we defined the logit model as:

$$\hat{\mathbf{y}} = \text{logit}(P) = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \ldots + \beta_j\mathbf{x}_j + \ldots + \beta_p\mathbf{x}_p \qquad (16.50)$$

where $\mathbf{x}_j$ represents the jth independent variable and where β_j is the coefficient of the jth independent variable which can be estimated by maximum likelihood regression. The coefficient β_0 denotes the constant term of the regression model.

Note that $\hat{\mathbf{y}}$ represents the maximum likelihood estimate of the log odds or logit of the outcome P, hence the name logit model. The dependent variable $\mathbf{y}$ takes values between minus and plus infinity. A variable $\mathbf{x}_j$, other than the treatment variable is determined to be a confounding factor if the corresponding coefficient β_j is significantly different at a stated level of significance α (e.g. 0.05 in a two-sided test). This information is available from statistical computer programs such as provided by SAS [26]. If the independent variable $\mathbf{x}_j$ is a dichotomous variable, then the corresponding coefficient β_j represents the log odds ratio corrected for confounding factors [27].

The logit model can also be defined with second degree terms:

$$\hat{\mathbf{y}} = \text{logit}(P) = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \ldots + \beta_{12}\mathbf{x}_1\mathbf{x}_2 + \ldots \qquad (16.51)$$

The linear terms in eq. (16.51) which represent the treatment variable and its possible confounding factors, account for the *main effects*, while the product terms represent *interactions* [28]. This distinction between main effects and interaction terms is illustrated in Fig. 16.10. It has an analogy in the analysis of variance (ANOVA) which is covered in Chapter 6. By way of example we assume that the two independent variables $\mathbf{x}_1$ and $\mathbf{x}_2$ are dichotomous and represent treatment (medication or control) and age (less than 50 years or 50 years or more).

In Fig. 16.10a we have the case where $\mathbf{x}_2$ is a pure confounding factor of $\mathbf{x}_1$. This situation has been encountered already in the discussion of Mantel–Haenszel's

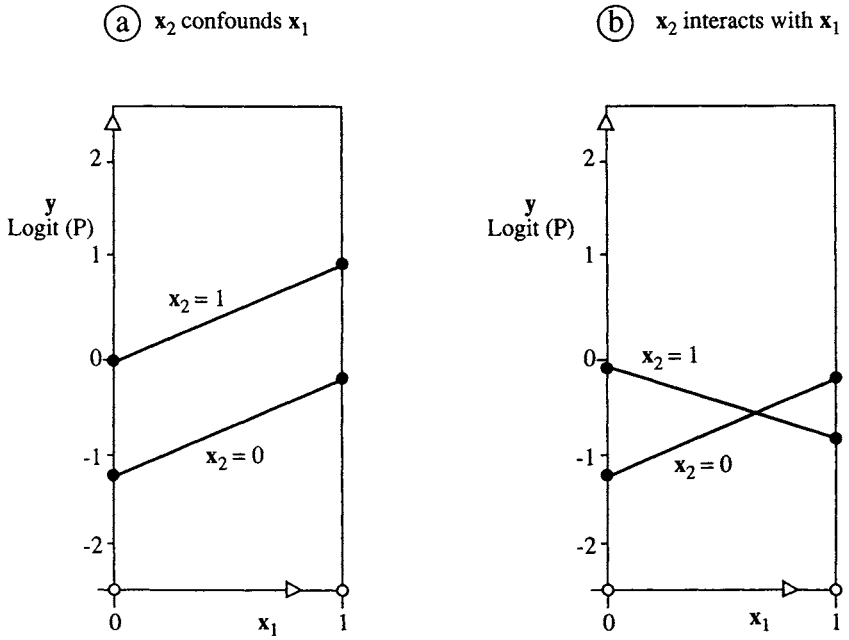


Fig. 16.10. Illustration of the cases in which the relation between outcome (**y**) and treatment ($\mathbf{x}_1$) is (a) confounded by another variable ($\mathbf{x}_2$) and (b) interacting with another variable ($\mathbf{x}_2$). Each response line represents a particular study.

method and Peto's approach to the analysis of multiple 2×2 contingency tables. Here the response variable **y** is expressed as logit(P) and appears to be linearly related to the design variables $\mathbf{x}_1$ and $\mathbf{x}_2$ by means of a linear logit model. Note that the slope of the response lines in Fig. 16.10a is equal to the log odds ratio β which has been defined by eq. (16.39). This so-called *common odds ratio* can also be derived from the simplified linear logit model:

$$\hat{\mathbf{y}} = \text{logit}(P(\mathbf{x})) = \beta_0 + \beta_1 \mathbf{x} \tag{16.52}$$

where the dichotomous independent variable **x** takes the values 0 and 1. We can show that the slope of the line is given by:

$$\beta = \text{logit}(P(1)) - \text{logit}(P(0)) = (\beta_0 + \beta_1) - \beta_0 = \beta_1 \tag{16.53}$$

The logistic model finds applications in retrospective studies which are also called *case-control studies* in epidemiology [28]. For each case that underwent a specific exposure (such as a therapeutic, dietary, environmental or other factor) one searches through the historical files for a number of matching controls (1 to 4 controls for each case). Matching is performed for obvious confounding factors such as age, gender and time of exposure. The logistic model then includes all other

relevant confounding factors and interactions. This approach has been used for the retrospective study of relationships between diseases and possible causative factors (for example between lung cancer and smoking) which cannot be studied by means of designed prospective experiments (or *cohort studies*).

16.2.13 Venn diagram

A 2×2 contingency table defines a relationship between two dichotomous variables, say **x** and **y** which may represent, for example, treatment (medication or control) and outcome (positive or negative). Each variable defines a set on the universe of discourse, which is in our case the total numbers of subjects included in the assay. Treatment defines the set of subjects with and without medication. Outcome defines the set of positives and negatives in the sample. In total there are 2×2 possible subsets in this sample. John Venn around 1880 has devised a diagram which is widely used for representing subsets in the form of connected areas. There are two variants of the *Venn diagram* as shown in Fig. 16.11a and b. Both diagrams represent the same four subsets defined by the dichotomous variables **x** and **y**. These are identified by different types of shading as indicated in the insert. In the Venn diagram of Fig. 16.11a one of the subsets, (not **x**) ∩ (not **y**), is represented by an unbounded area, while in that of Fig. 16.11b all subsets are bounded. The latter idea seems to be attributed to a contemporary of John Venn, the famous writer Lewis Carroll [29].

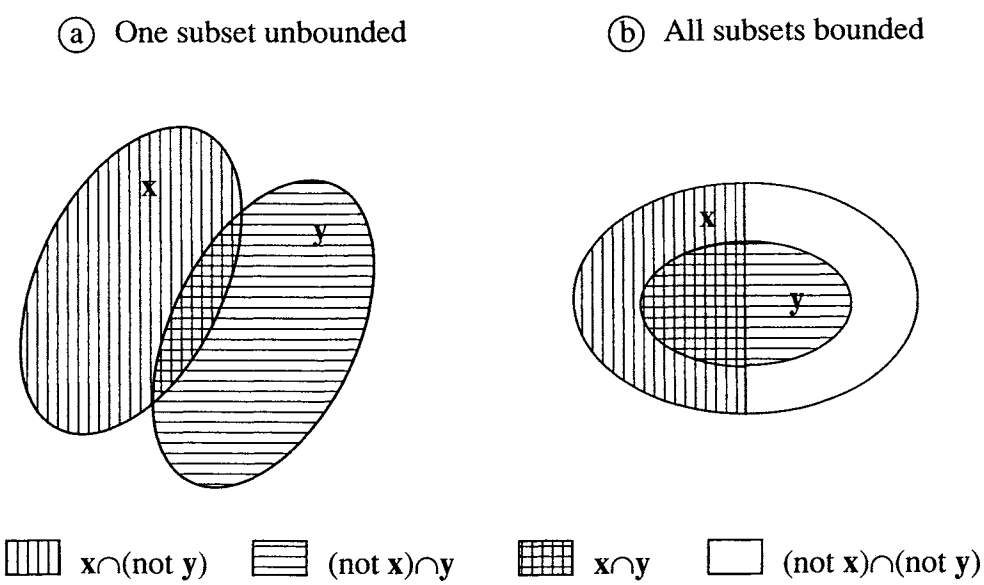


Fig. 16.11. Two types of Venn diagram for the visual display of a 2×2 contingency table. In (a) one of the four subsets is unbounded. In (b) all four subsets are bounded.

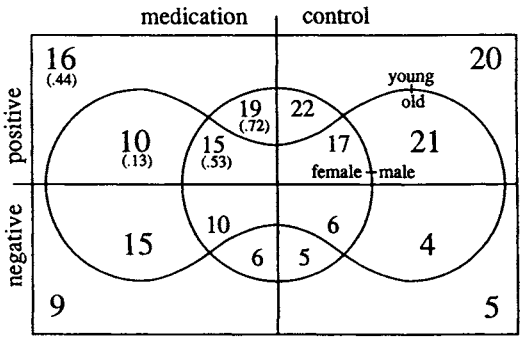


Fig. 16.12. Edwards's map of a 4-set Venn diagram showing the 16 subsets produced by an assay of outcome vs treatment stratified by gender and age. The number of subjects is indicated in each subset [29]. Numbers within brackets represent the common odds ratios of the four 2×2 contingency tables resulting from the stratification by gender and age.

Extensions of the Venn diagram to higher order sets produce tangled and obtrusive displays. Recently, however, Edwards [29] discovered how the Venn diagram can be used in a very attractive way to represent multiple sets for the illustration of the general case of a 2×2×2×2×... contingency table. In *Edwards's map* of the Venn diagram higher order sets are threaded with increasing periodicity around the boundary of a central circular set.

Figure 16.12 represents the case of a 2×2×2×2 contingency table in which the four sets are defined by the variables outcome (positive or negative), treatment (medication or control), gender (male or female) and age (old or young). This assay of outcome with respect to treatment has been *stratified* for gender and age. Hence, the figure is the equivalent of four 2×2 contingency tables, representing a total of 16 subsets of a sample of 200 subjects. Each of these subsets corresponds with one of the 16 fields of Edwards's map. For example, 16 young and male subjects showed a positive outcome with medication, 4 older and male control subjects obtained a negative outcome, etc.

In order to interpret the diagram correctly, one has to pick one number in a particular subset and find the corresponding numbers in the three subsets that are similarly shaped. For example, the elements of the 2×2 contingency table for older males are 10, 15, 21 and 4. In the upper left quadrant of Fig. 16.12 we have added (between brackets) the common odds ratio (eq. (16.36)) corresponding to the four 2×2 contingency tables that resulted from the stratification by gender and age. In the case of older males this common odds ratio is (10/15)/(21/4) or 0.13 as indicated in the figure.

An important property of this diagram is that the crossing of the boundaries between two subsets changes only one variable at a time. Thus we can observe readily that a change from young to old produces a drop in the odds ratio (from

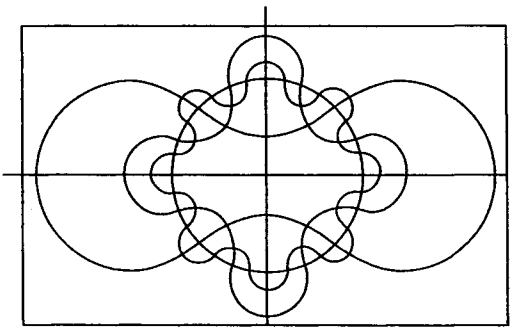


Fig. 16.13. Edwards's map of the 7-set Venn diagram showing the 128 possible subsets in a complete 2^7 factorial design [29].

0.44 to 0.13 for male and from 0.72 to 0.53 for female subjects). Similarly a change from female to male also produces a drop in the odds ratio (from 0.72 to 0.44 for young and from 0.53 to 0.13 for older subjects). The smallest odds ratio is obtained in the case of older males (0.13).

Edwards's map can also be used for visual exploration of the results of complete 2^k *factorial designs*. The 7-set diagram is shown in Fig. 16.13.

16.2.14 General contingency table

A general $n \times p$ contingency table $\mathbf{X}$ is constructed by crossing the n categories of one variable with the p categories of another one. Each cell n_{ij} of the table contains the number of joint occurrences of category i and of category j of the two variables. By analogy with eq. (16.21) we derive the expected value of x_{ij} from the marginal totals of the table which are assumed to be fixed:

$$E(x_{ij}) = \frac{x_{i+}\, x_{+j}}{x_{++}} \qquad \text{with } i = 1, \ldots, n \text{ and } j = 1, \ldots, p \tag{16.54}$$

where x_{i+} and x_{+j} represent the totals of row i and of column j and where x_{++} is the grand total of table $\mathbf{X}$. The Pearson χ^2 statistic for the $n \times p$ contingency table is then defined by means of:

$$\chi^2 = \sum_{i}^{n} \sum_{j}^{p} \frac{(x_{ij} - E(x_{ij}))^2}{E(x_{ij})} \tag{16.55}$$

which possesses $(n - 1) \times (p - 1)$ degrees of freedom.

As an example, we analyze the data already presented in Table 16.8 on the presence and absence of myocardial infarctions in persons complaining of chest pain and from five categories of serum creatine kinase (SCK). The observed frequencies x_{ij}, the corresponding expected frequencies $E(x_{ij})$ according to eq.

TABLE 16.20

Observed, expected and observed-expected frequencies of the presence and absence of myocardial infarction for various categories of serum creatine kinase (SCK in IU/ml) in persons complaining of chest pain [7]

	SCK (IU/ml)	Present	Absent	Total
Observed myocardial infarctions	0–120	23	471	494
	121–240	6	201	207
	241–360	7	24	31
	361–480	6	12	18
	>480	9	14	23
	Total	51	722	773
Expected myocardial infarctions	0–120	32.6	461.4	494
	121–240	13.7	193.3	207
	241–360	2.0	29.0	31
	361–480	1.2	16.8	18
	>480	1.5	21.5	23
	Total	51	722	773
Observed–expected myocardial infarctions	0–120	–9.6	9.6	0
	121–240	–7.7	7.7	0
	241–360	5.0	–5.0	0
	361–480	4.8	–4.8	0
	>480	7.5	–7.5	0
	Total	0	0	0

(16.54) and the observed minus expected frequencies are displayed in Table 16.20. It can readily be observed from the table that the persons with SCK-values larger than 240 (IU/ml) have a greater incidence of myocardial infarction than would be expected from the marginal frequencies. The χ^2 statistic (eq. (16.55)) associated with this 5×2 contingency table amounts to 80.9 and is to be tested with $4\times1 = 4$ degrees of freedom. From a table of critical values of the χ^2 distribution we may conclude that the SCK assay is capable of predicting myocardial infarction in persons complaining of chest pain. The probability that the observed differences in the contingency table have arisen by chance is less than 0.001.

The χ^2 statistic is but a global indicator of interaction between the two variables that define the table. It does not reveal which particular categories of the two variables are interacting. In a large table this is difficult to perceive without the use of multivariate analysis such as explained in Chapters 31 and 32. Two particular approaches are useful for the analysis of general contingency tables, namely correspondence factor analysis and log-linear models (which includes spectral map

analysis). The former can be considered as the multivariate analysis of chi-square χ^2, more exactly distance of chi-square χ^2/x_{++}. The latter can be regarded as a multivariate analysis of the sum of squares c of the data after a transformation by logarithms:

$$c = \frac{1}{np} \sum_{i}^{n} \sum_{j}^{p} (y_{ij} - E(y_{ij}))^2 \qquad \text{with } i = 1, ..., n \text{ and } j = 1, ..., p \tag{16.56}$$

where

$y_{ij} = \ln (x_{ij})$ or, equivalently, $\mathbf{Y} = \ln \mathbf{X}$

$E(y_{ij}) = y_{i.} + y_{.j} - y_{..}$

where $y_{i.}$ and $y_{.j}$ represent the row and column means and where $y_{..}$ is the global mean of the logarithmically transformed values in the table **Y**.

In Chapter 32 it will be shown that correspondence factor analysis and log-linear models yield similar results, unless strong interactions between rows and columns are present, i.e. when observed frequencies are far from their expected values.

References

1. B.S. Everitt, The Analysis of Contingency Tables. Chapman and Hall, London, 1977.
2. X, Science Briefings, Testing for HIV in urine. Science, 249 (1990) 121.
3. M.J. Gardner and D.G. Altman, Statistics with Confidence. British Medical Journal Publ., London, 1989.
4. H. Kaiser, Zur Definition von Selektivität und Empfindlichkeit von Analysenverfahren. Z. Anal. Chem., 260 (1972) 252–260.
5. H. Kaiser, Zum Problem der Nachweisgrenze. Z. Anal. Chem., 209 (1965) 1–18.
6. C. Howson and P. Urbach, Bayesian reasoning in science. Nature, 350 (1991) 371–374.
7. K.L. Radack, G. Rouan and J. Hedges, The likelihood ratio. Arch. Pathol. Lab. Med., 110 (1986) 689–693.
8. H. Delooz, P.J. Lewi and the Cerebral Resuscitation Study Group, Early prognostic indices after cardiopulmonary resuscitation (CPR). Resuscitation, 17 (1989) 149–155.
9. P.J. Lewi and R. Marsboom, Toxicology Reference Data — Wistar Rat. Elsevier/ North-Holland, Amsterdam, 1981.
10. Ch.E. Metz, Basic principles of ROC analysis. Seminars in Nuclear Medicine, 8 (1978) 283–298.
11. Ch.E. Metz, ROC methodology in radiologic imaging. Investigative Radiol., 21 (1986) 720–733.
12. W.G. de Ruig and H. van der Voet, Is there a tower in Ransdorp? Harmonization and optimization of the quality of analytical methods and inspection procedures. In: D. Littlejohn and D. Thorburn Burns (Eds.), Reviews on Analytical Chemistry — Euroanalysis VIII. The Royal Society of Chemistry, Cambridge, 1994.
13. M.G. Kendall and A. Stuart, The Advanced Theory of Statistics. Vol. 2. Ch. Griffin, London, 1967.

14. S. Siegel and N.J. Castellan, Nonparametric Statistics for the Behavioral Sciences. McGraw Hill, New York, 1988.
15. W.G. Cochran, Some methods for strengthening the common χ^2 tests. Biometrics, 10 (1954) 417–451.
16. W.J. Connover, Some reasons for not using the Yates continuity correction on 2×2 contingency tables. J. Am. Statist. Assoc., 69 (1974) 374–382.
17. Federal Register, Part II, Office of Science and Technology Policy, Chemical Carcinogens. Evaluation of Statistical Significance. Vol. 49, May 22, 1989, p. 21637.
18. P.J. Lewi, Graphical assessment of statistical significance and clinical-biological relevance. Drug Devel. Res., 15 (1988) 419–425.
19. Encyclopedia of Statistical Sciences. S. Kotz and N.L. Johnson (Eds.), Vol. 1. Wiley, New York, 1982, pp. 300–301.
20. M.R. Spiegel, Statistics, Schaum's Outline Series. McGraw Hill, New York, 1972.
21. N. Mantel and W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst., 22 (1959) 719–748.
22. S. Yusuf, R. Peto, J. Lewis, R. Collins and P. Sleight, Beta blockade during and after myocardial infarction: an overview of the randomized trials. Progress in Cardiovascular Diseases, 27 (1985) 335–371.
23. R.A. Fisher, Statistical Methods for Research Workers. Oliver and Boyd, London, 1932.
24. R. DerSimonian and N. Laird, Meta-analysis in clinical trials. Controlled Clin. Trials, 7 (1986) 177–186.
25. D.R. Cox, Analysis of Binary Data. Chapman and Hall, London, 1970.
26. SAS Institute Inc., SAS/STAT User's Guide. Version 6, Fourth Edition, Vol. 2, Chapter 27, PROC Logistic, Cary, NC, 1990.
27. N.E. Breslow and N.E. Day, Statistical Methods in Cancer Research. Vol. I. The Analysis of Case-control Studies. International Agency for Research on Cancer (IARC Scient. Publ. No. 32), Lyon, 1980.
28. J.J. Schlesselman, Case-control Studies. Design, Conduct, Analysis. Oxford University Press, New York, 1982.
29. A.W.F. Edwards, Venn diagrams for many sets. New Scientist, 121 (1989) 51–56.

Chapter 17

Principal Components

17.1 Latent variables

This chapter is a first introduction to *principal components* and *principal component analysis (PCA).* The approach here is mainly intuitive and non-mathematical. It also illustrates some applications of the method. The principal components concept is very important to chemometrics. The soft modelling methods and the multivariate calibration methods described in Part B are based on it. The first sections of this text are based on an audiovisual series about PCA [1]. A more systematic mathematical introduction is given in Part B, Chapters 29, 31 and 32.

One of the reasons for the use of PCA resides in the enormous amount of data produced by modern computers and measurement techniques. For instance, inductively coupled plasma emission allows the analysis of many elements simultaneously, capillary gas chromatography can easily yield concentrations of 100 compounds in a single run and near-infrared spectrometry yields absorbances at several hundreds of wavelengths in a very short time. If many samples are analyzed in that way, so many data result that it is virtually impossible to make intelligent use of the data without further treatment.

One obvious way of organising the data of the kind described earlier is to construct a table, in which the n objects constitute the rows and the m variables constitute the columns. However if, as in one example that will be discussed later, 300 air samples (the objects) had been measured and the concentrations of 150 substances (the variables) had been determined in each, the table would contain 45000 data entries and it would be very difficult to extract meaningful information from it. One would like to visualize the information and PCA provides the means to achieve this.

The data table of Fig 17.1 is, in chemometrical terms, a data matrix and many of the calculations will be based on matrix algebra. A first introduction will be given in Section 17.6 and a detailed account in Chapter 31. The table can also be described as a two-way table. Three-way tables also exist and can be treated by *three-way PCA*. This will be described in Chapter 31.

A few terms specific to this field should first be defined. One is *feature*, which means variable. When one says that principal components is used for *feature reduction* this means therefore that PCA reduces the number of variables in some

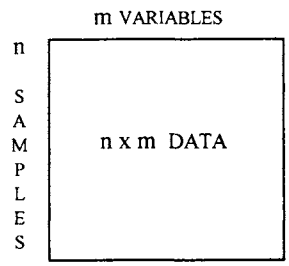


Fig. 17.1. The data matrix.

way. In Section 17.3, we will explain that this is done by making linear combinations of the original variables. The term should not be confused with *feature selection*. The latter term means that one selects some variables, e.g. only a few wavelengths from a spectrum are used. In feature reduction, as applied in PCA, all wavelengths are used, but the number of variables is diminished because of the linear combinations. How will be discussed later, but the term makes it clear that PCA in some way simplifies the presentation of the data. The data of a data matrix are also called *multivariate* or *multidimensional*. These last terms merit some further clarification.

Chemists, like many other scientists, like to draw graphs to understand better the data they have obtained. Let us suppose that a chemist has determined the concentration of a single substance in a few samples. The concentration, x_1, is then considered as a variable or feature to be plotted. The resulting plot could look as in Fig. 17.2a and in this case would tell the chemist that the samples really belong to two groups. Such a one-dimensional plot and the data themselves are called *univariate*.

Usually, more than one variable is measured in the hope of obtaining more information. When two variables, x_1 and x_2, are measured, the chemist can plot the samples in a plane of x_1 versus x_2 (Fig. 17.2b). The two-dimensional plot and the data are now called *bivariate*. Plots are still possible in three dimensions (Fig. 17.2c). The number of dimensions is equal to the number of variables measured for each sample or object. It follows that data from a data matrix in which m variables have been measured for each sample, are m-dimensional. They are called *multivariate* and to visualize them we would need m-dimensional plots. As this is not possible for dimensions larger than three, we must reduce the number of features to three or less.

Before investigating how to reduce features from m-dimensional space to two or three dimensions, let us consider the simplest possible case of feature reduction, namely the situation where only two variables are present and we want them reduced to one. We imagine that we are able to see along one dimension only. This would mean that we would not be able to perceive visually the structure of the two-dimensional data in Fig. 17.3. An obvious solution would be to project the points from the two-dimensional space (the plane) to the one-dimensional space of the line. The direction of that line is important. In Fig. 17.3a the projections on the

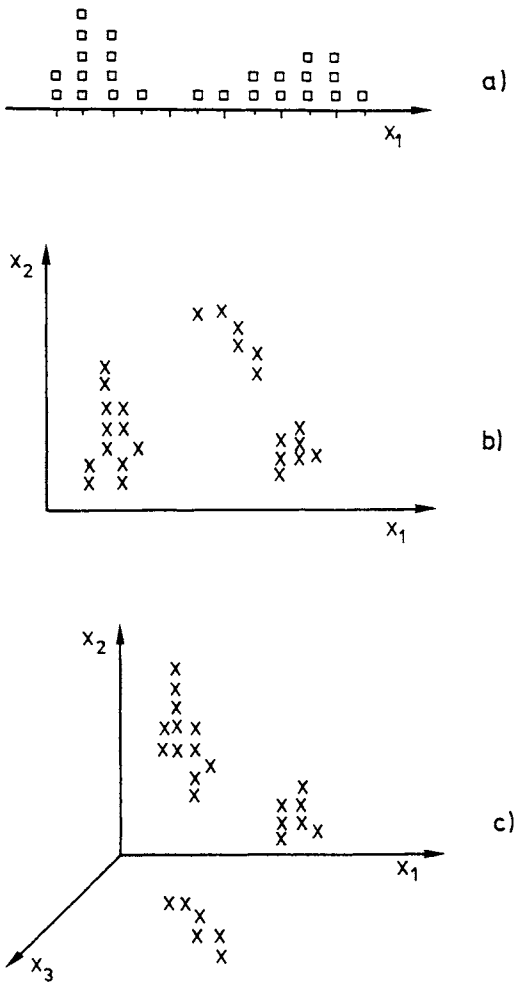


Fig. 17.2. Plots in one (a), two (b) and three (c) dimensions.

line do not yield much information about the structure of the original data set. It is for instance not possible to observe from the projections, the crosses along the line, that there are two groups in the data. In Figure 17.3b, the projections do allow us to observe the most important characteristic of the data structure: two groups of crosses are clearly present. A good direction to draw the line is along the axis of largest variation in the data. This line is called the first *principal component*, PC1, and one can say that *PC1 explains the largest possible variation in the data* and therefore that *PC1 accounts for most information*. The projections of the points from the original x_1–x_2 space on PC1 are called the *scores* (of the objects) on PC1.

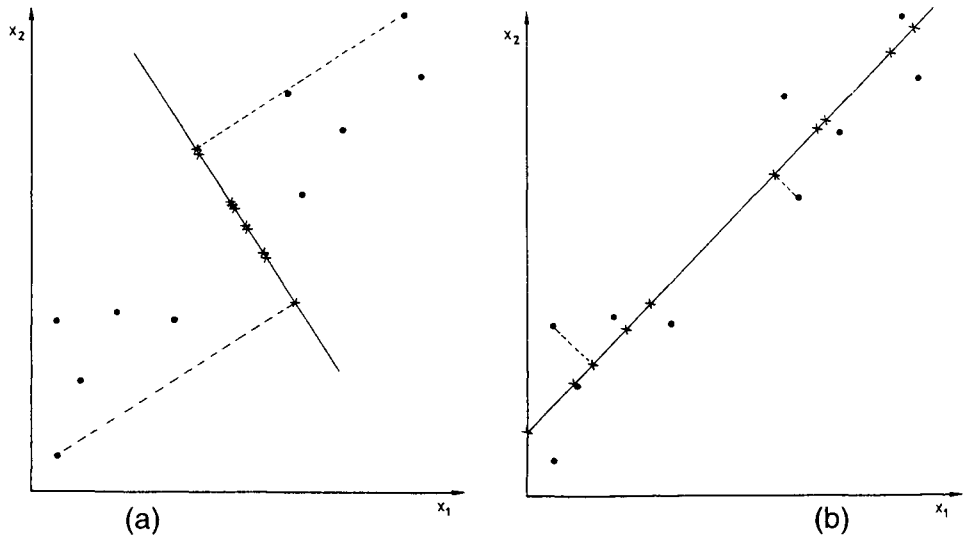


Fig. 17.3. Projections from two to one dimension: (a) the information is lost on projection; (b) the information is retained. The line in (b) is PC1.

The objects are dispersed around the PC line. The residuals, r_i (Fig. 17.4), express the remaining or *unexplained variation*. We can express that variation on a second axis, by definition orthogonal to the first, on which one also projects the data from the original space. This is the second principal component, PC2 (Fig. 17.4a) and the projections are the scores (of the objects) on PC2.

PC1 and PC2 can be considered as the new axes in the same two-dimensional space. If we work on mean-centred data (see Chapter 9), at the same time the origin of the new coordinate system is translated to a more natural location, namely the *barycentre* or centre of mass of the data (the barycentre is the coordinate corresponding with the mean for each variable). One can plot the scores of the objects on PC1 against those on PC2. This is done in Fig. 17.4b for the objects originally present in the x_1–x_2 coordinates. The two groups of objects that can be distinguished in the x_1–x_2 space can also be seen in the two-dimensional space PC1–PC2.

Let us now return to our supposition that we are able to see only along one dimension. One would then choose to look along PC1, and obtain the result of Fig. 17.4c. The crosses are the scores and therefore the projections from the original x_1–x_2 space on PC1. They tell us almost as much about the original data structure as the original x_1–x_2 graph. The feature reduction is successful, since the number of variables was reduced from two (x_1 and x_2) to one, namely PC1, without significant loss of information. PC1 can be described as a new and perhaps more fundamental variable than the original variables, x_1 and x_2, separately. PC1 is called a *latent variable* in contrast to x_1 and x_2. These are called the *manifest variables*. In

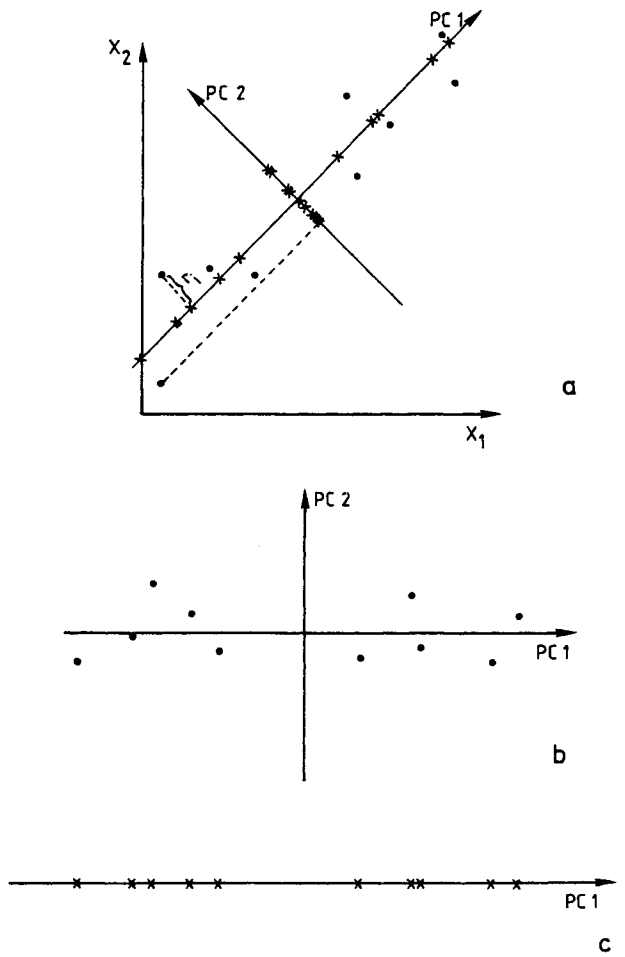


Fig. 17.4. (a) PC2 is orthogonal to PC1. (b) The PC plot: the projections of each object of (a) on PC1 are plotted against those on PC2. (c) Feature reduction by retaining only one PC.

this specific case PC2 essentially expresses noise. We can separate information (PC1) from noise (PC2).

Another example is given in Fig. 17.5a. The direction of PC1 is determined here to a large extent by point 7. This point is an outlier compared to the 6 others and therefore it is responsible for a large part of the variation. On PC1 it is distinguished easily from the 6 other points. This also shows that the principal components plot allows us to find outliers in a data set. It also means that outliers can mask the structure of the other data. Indeed, the two groups of three points cannot be distinguished along PC1. When outliers are found in a PC plot, one should first identify and then eliminate the outlier and carry out the PCA again. The PC1 of the

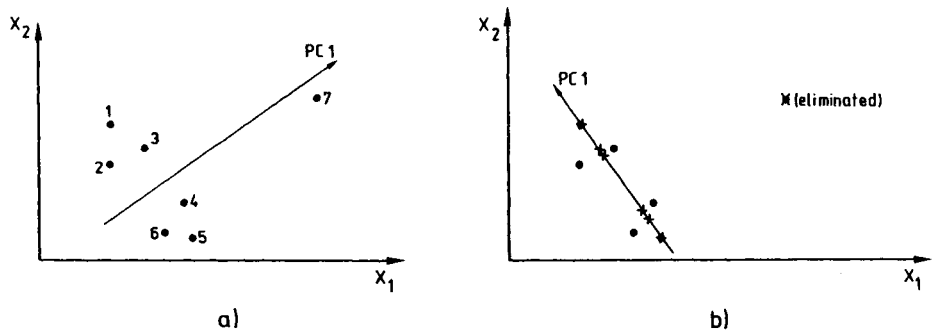


Fig. 17.5. (a) PC1 for a two-dimensional data set containing an outlier (point 7). (b) PC1 for the same data set after elimination of point 7.

six points remaining after elimination of the outlier is shown in Fig. 17.5b. The residual structure residing in the remaining data points is now revealed, since one observes two groups of three points along the new PC1.

In Fig. 17.6a a three-dimensional case is presented. The direction of largest variation of the data is indicated. This is then PC1. The projection of the data points on PC1 is shown on Fig. 17.6b. The data structure is well preserved. However, this is only so because the data are more or less linear which means that there is only one main direction of variation.

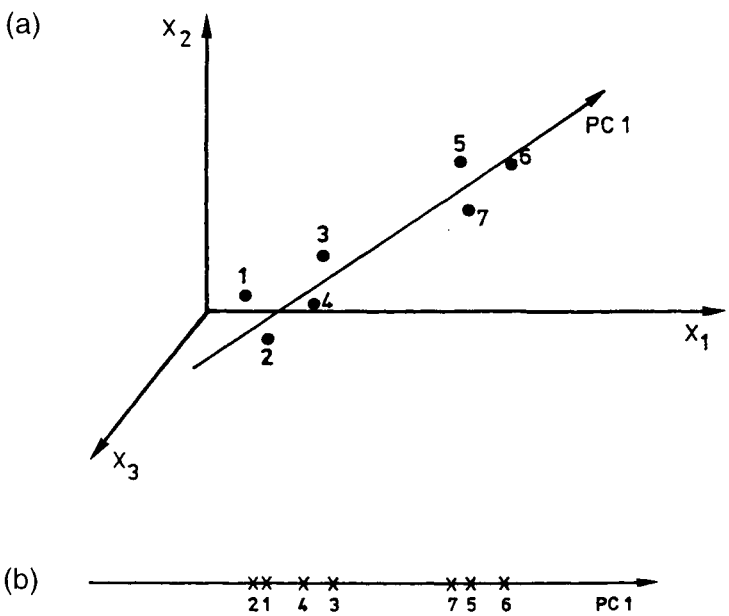


Fig. 17.6. (a) PC1 for a three-dimensional data set. (b) Projection of points on PC1 for the data set from (a). One PC is sufficient for a good representation of the original data.

In Fig. 17.7a, the data are structured to a smaller extent around a straight line. The projection of the data on PC1 (Fig. 17.7b) shows some differentiation between the different categories, but for example the fact that group 5–6–7 is separate from 8–9–10 cannot be observed. There is quite some residual variation around PC1, so that one proceeds by determining PC2. It must be orthogonal to PC1, but, in contrast with the two-dimensional case, its direction is not determined *a priori* because there are now three dimensions and PC2 is chosen so that it is drawn in the direction of largest residual variation around PC1. The different points can be projected on PC1 and PC2 and the resulting scores used as the new coordinates to picture each point in the PC1–PC2 plane (Fig. 17.7c).

We have now applied PCA in the way it is used most frequently, namely to present in two dimensions the information present originally in more than two dimensions. In this case, the information is well preserved and one can distinguish correctly three groups of objects. In fact, three principal components can be determined, but PC3 would in this case not have added information. One has therefore carried out feature reduction, since the number of features was reduced from 3 to 2.

(a)

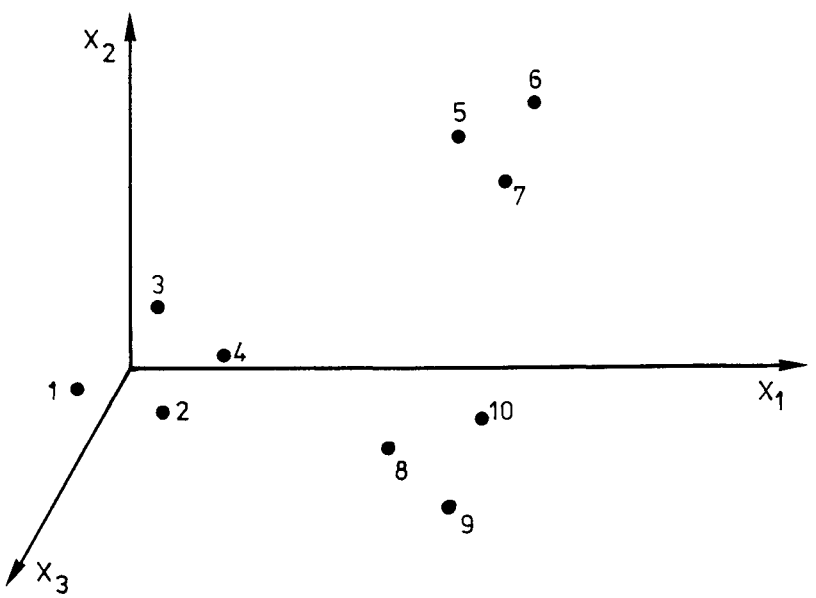
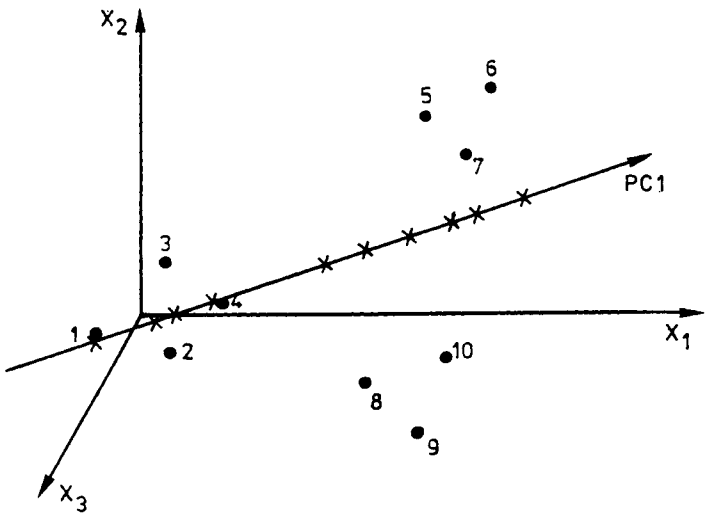
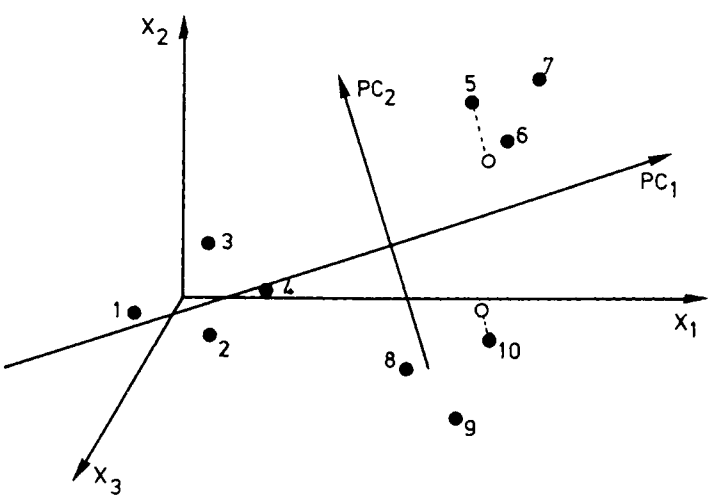


Fig. 17.7. (a) PC for a three-dimensional data set.

(b)





(c)

Fig. 17.7 (continued). (b) Projection of points on PC1 for the data set from (a). (c) The PC1–PC2 plane. Projections on this plane would yield a good representation of the original data.

More generally, when one analyzes a data matrix consisting of n objects for which m variables have been determined, m principal components can then be extracted (as long as $m < n$). PC1 represents the direction in the data, containing the largest variation. PC2 is orthogonal to PC1 and represents the direction of the largest residual variation around PC1. PC3 is orthogonal to the first two and represents the direction of highest, residual variation around the plane formed by PC1 and PC2. Although it becomes difficult or even impossible to visualize more components, one can continue extracting principal components in this way, until m principal components have been obtained. These will contain less and less variation and therefore less information. The projections of the data on the plane of PC1 and PC2 can be computed (or sometimes also on the PC1–PC3 or the PC2–PC3 plane) and shown in a plot. This plot is called a *score plot.*

17.2 Score plots

A first example comes from the field of food authentication. One would like to verify the origin (geographical, plant or animal species, etc.) of a foodstuff by carrying out appropriate chemical determinations. One then needs to determine a large number of variables and hopes that the resulting patterns of data will permit to differentiate between the different samples (objects). Pattern recognition methods (see Chapter 33) are used in food authentication and PCA is nearly always applied to make a preliminary study of the structure of the data.

The example of Fig. 17.8 concerns 3 Italian wines: Barolo, Barbera and Grignolino [2]. About 100 certified samples of known origin of the three wines were analyzed for 8 variables, namely alcohol, total polyphenols and 6 measurements of optical densities at several wavelengths. The latter describe the colour of the wine. The resulting data matrix consists therefore of 100 by 8 data and to plot the original data one would need therefore an 8-dimensional plot. Figure 17.8 shows the plot of PC1 against PC2. One observes that the Barolo wines can be distinguished to a large extent from the two others and one concludes that the 8 variables are appropriate for the discrimination of this type of wine from the two others. It should be noted that the discrimination as such will then be performed using other methods such as linear discriminant analysis (see Chapter 33) or perhaps neural networks (see Chapter 44).

Figure 17.9 is a score plot where the objects are 14 tablets for which near infra-red (NIR) spectra (1050 wavelengths) were measured [3]. The absorbances at these wavelengths are the variables. This is therefore a high-dimensional, namely a 1050-dimensional, situation. The final aim of the data analysis was to be able to analyze the active substance in the tablet by relating absorbance to concentration in a multivariate calibration model, but the purpose of the score plot was a

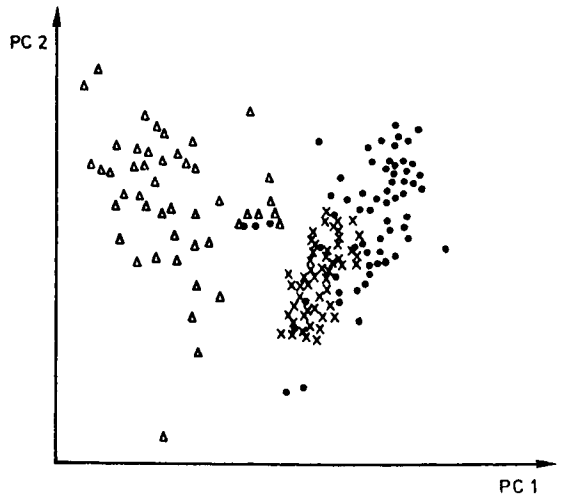


Fig. 17.8. Score plot of 3 Italian wines (Δ = Barolo, ● = Barbera, × = Grignolino) in the PC1–PC2 plane (from Ref. [2]).

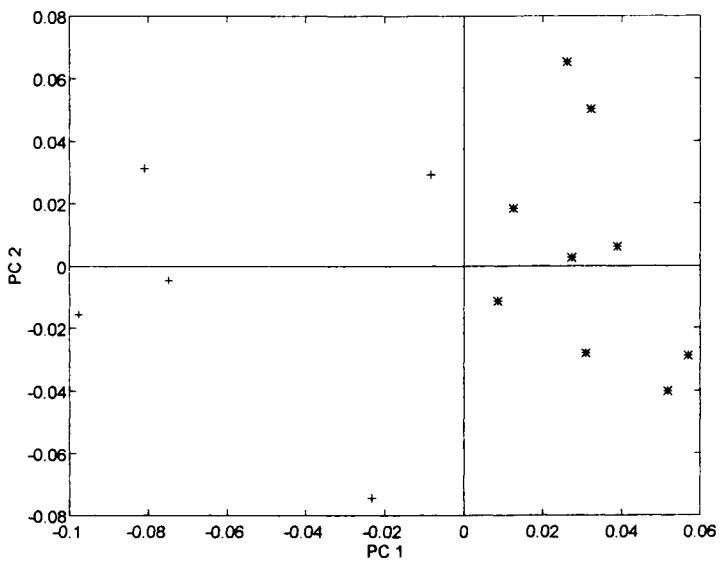


Fig. 17.9. Score plot (PC1–PC2) of 14 tablets characterized by their NIR spectra. The meaning of the symbols is explained in the text.

preliminary evaluation of the data. There are two sets of tablets: the tablets indicated with + were obtained from the production line, those indicated with a * were produced in the laboratory to obtain a larger concentration range. It was therefore expected that the + objects should fall inside the group of * objects. The

14 tablets together would then serve to develop the multivariate calibration model. Unfortunately the score plot does not show the expected pattern since the two groups are clearly separated along PC1. This means that there is a difference in the NIR spectra of the two sets and it turned out that the reason for this is that the shape of the two sets of tablets was not exactly the same. As we will see later, it however still proved possible to apply principal components to obtain a multivariate calibration model (see further in this section).

Another application concerns organic air pollution. At four monitoring stations in the Netherlands, air samples were collected every week, once in the morning and once in the afternoon during a period of three years [4]. In these samples nearly 150 volatile organic compounds were analyzed by means of gas chromatography. Moreover, 12 meteorological parameters were measured at the same time. Nearly half a million data were therefore collected. The interpretation of such a big data set absolutely requires visualization and feature reduction. To show how PCA does this, a subset concerning data for about one year at one monitoring station will be studied and only 26 of the more important substances are considered. Figure 17.10 shows the air samples plotted in a plane defined by the first two principal components. Three clusters of samples can be distinguished: a rather compact cluster situated to the left of the figure and two more or less elongated clusters both situated in different directions from this central cluster. The central cluster contains most of the samples. It therefore defines the most common condition, the normal level of pollution situation. The two elongated clusters, on the contrary, determine pollution situations that are less frequent. In fact further analysis of the PC plot (see Section 17.3) shows that these are samples in which the pollution is larger. It is also very instructive to give different symbols to the samples according to some other parameter, here the wind direction at the time of sampling. One observes that the

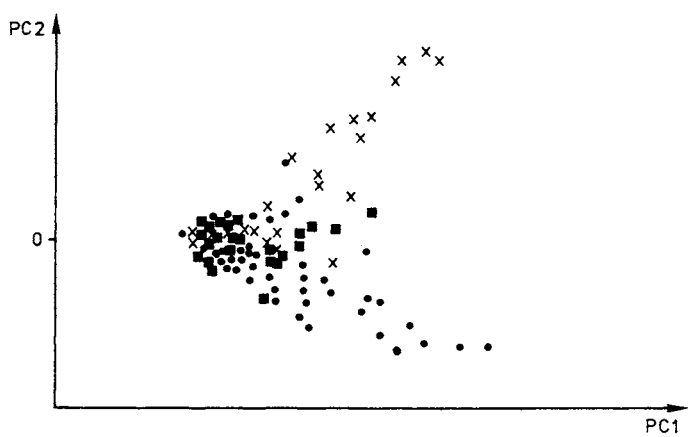


Fig. 17.10. Score plot for the air pollution example [3]. The symbols indicate different wind directions.

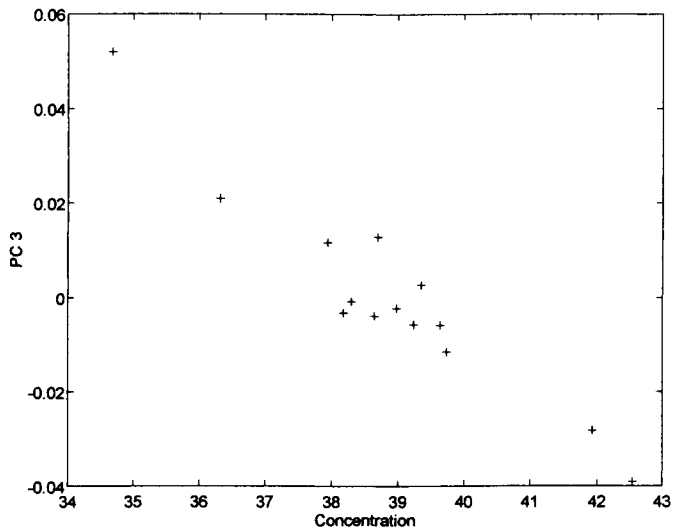


Fig. 17.11. Scores on PC3 in function of concentration for the tablets of Fig. 17.9.

three clusters correspond with three different wind directions. The first elongated cluster corresponds with air samples coming from the industrialised south and south west directions; the second elongated cluster represents air samples collected when the wind was coming from the north, a greenhouse area, and the central cluster contains the samples obtained when the wind blew from the urban zone to the east of the monitoring station.

Instead of making two-dimensional plots of scores, it is often useful to plot the scores of single PCs against one or other characteristic, such as concentration. This was done with the scores of the PCs for the NIR data we described earlier. Plots were made of different PCs against the concentration of the active substance. The scores on PC1 and PC2 are not related to concentration but those on PC3 are (Fig. 17.11). This fact can be used when building multivariate calibration models (see Section 17.8).

17.3 Loading plots

The examples in Section 17.2 show that PCA, through feature reduction and visual display, allows us to observe the sources of variation in complex data sets. It is, however, possible to extract much more information from a PCA. So far, we have focused on the score plots and the relationships among objects. In the same way, we can wonder about relationships between variables. Some additional theory about principal components must first be introduced to be able to do this.

As explained in Section 17.1 the value of object i (its projection) along a principal component PCp is called the score, s_{ip}, of that object on that principal component. In the air pollution example the objects were air samples for which many chemical substances were measured. The concentration of these substances such as benzene, toluene, etc. constitute the manifest variables. The scores on the principal components are a weighted sum of the original variables, the concentrations. They are given by: $s_{ip} = v_{\text{benzene},p} \cdot [\text{benzene}]_i + v_{\text{toluene},p} \cdot [\text{toluene}]_i + v_{\text{decane},p}$ $[\text{decane}]_i + \ldots$ where [benzene] is the concentration of benzene. The v-values are the weights and they contain information about the variables. This can be written for each of the principal components that are considered. These weights are called *loadings* and in general one can write the equation as follows:

$$s_{ip} = \sum_j v_{jp} \cdot x_{ij} \tag{17.1}$$

where v_{jp} is the loading of variable j on PCp and x_{ij} is the value of the ith object for manifest variable j. The scores are *linear combinations* of the manifest variables.

In matrix notation :

$$\underset{n\times m}{\mathbf{S}} = \underset{n\times m}{\mathbf{X}} \quad \underset{m\times m}{\mathbf{V}} \tag{17.2}$$

where **S** is the matrix of scores, **X** the matrix of manifest variables and **V** the matrix of loadings. This equation relates the principal components to the original variables.

Table 17.1 shows the loadings of the more important variables, i.e. those with the highest loadings on PC1. A word of warning must be added here. The scores of the objects are weighted sums of the manifest variables and the units in which original variables are expressed determine therefore the value of the scores. The weights or loadings are constrained by mathematical techniques; as will be explained in Chapter 31 the sum of the squared loadings is equal to 1. There is no such constraint on the manifest variables and, if one does not pretreat them in some way, this will make the result dependent on the scale of the original variables. Preliminary operations on the data are important. Chapter 31 describes methods such as column-centring, column-standardization, log column-centring, log double-centring and double-closure. In this specific case, column-standardization was applied to eliminate scale effects. For the moment, this will not be explained further.

The loadings in Table 17.1 are rather similar and it is striking that all loadings have the same sign: they influence PC1 in the same direction. This phenomenon is often observed. When it occurs, it can be interpreted by considering that PC1 is an indicator of general size, in this case a variable describing how high the total pollution is. Objects with a high score show high pollution. Indeed, a high score means that the values of many of the original variables, i.e. the concentrations of the pollutants, are also high.

TABLE 17.1

Loadings for PC1 of the more important variables of the example of Fig. 17.10

m-ethylmethylbenzene	0.300	ethylbenzene	0.230
toluene	0.288	*n*-dodecane	0.227
1,2,4-trimethylbenzene	0.277	*n*-octane	0.226
n-nonane	0.265	*p*-xylene	0.216
o-xylene	0.232	styrene	0.202
n-decane	0.231	tetrachloroethylene	0.201
n-tridecane	0.231		

Only loadings with absolute value >0.200 are given.

TABLE 17.2

Loadings for PC2 of the more important variables of the example of Fig. 17.10

Largest negative scores		Largest positive scores	
o-xylene	–0.310	*n*-decane	0.260
ethylbenzene	–0.234	*n*-dodecane	0.260
p-xylene	–0.274	*n*-undecane	0.315
styrene	–0.204	*n*-nonanal	0.281
m-xylene	–0.226	isobutylacetate	0.200
benzene	–0.289	acetophenone	0.247

The loadings for PC2 (Table 17.2) yield a very different picture. Some of these loadings are positive, while others are negative. One also observes that those substances that have the same sign are chemically related: substances with negative loadings are all aromatic, while most of the substances with a positive loading are aliphatic.

One concludes that PC2 apparently differentiates between samples with pollution of an aromatic nature and samples with a more aliphatic pollution. Highly positive scores along PC2 are indicative of aliphatic pollution. The samples with highly positive scores on PC2 and those with highly negative scores were obtained for different wind directions (Fig. 17.10). It can now be concluded that samples from the wind direction, indicated by a cross, are characterized by a higher aliphatics to aromatics ratio than the samples taken when the wind comes from the direction symbolized with a black dot. PC1 tells us that total pollution is higher when the wind blows from the cross or dot direction than when it comes from the square direction. PC2 shows that when high pollution occurs it is characterized by different aliphatic to aromatic ratios and the ratio depends on wind direction. While PC1 is a general *size component*, PC2 expresses *contrast*. It should be noted here that, by using certain types of pretreatment (see Chapter 31), the size effect can be eliminated.

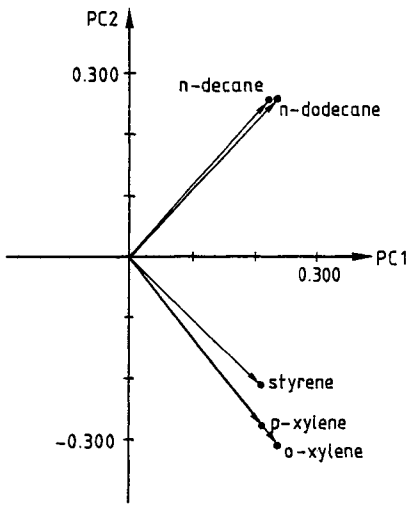


Fig. 17.12. Loading plot of a few of the variables of the air pollution example.

The information present in the loadings can of course be displayed also in two-dimensional *loading plots*. One can for instance plot the loading of each variable on PC1 against the loading of that variable on PC2. The interpretation is based on the direction in which the variables lie on this plot as seen from its origin. When two variables show a high correlation coefficient, one can predict the values for one variable to a large extent from the other one. It also means that when one has determined one variable, the other does not yield much additional information (see also Chapter 18). It is interpreted in the same way as a vector. Two variables are strongly correlated when there is a small angle between the lines connecting them with the origin. Figure 17.12 shows a few loadings on PC1 and PC2 for the variables of the air pollution example. *O*- and *p*-xylene are very strongly correlated since they fall on the same line from the origin. Styrene is somewhat less, but still strongly correlated with the two xylenes. Decane and dodecane are also strongly correlated with each other, but not with the three other substances shown.

In the same way that it is useful to plot the scores of PCs one by one (see the preceding section) it can also be interesting to plot the loadings of single PCs. This is certainly the case when the variables are wavelengths (or rather an optical measurement, such as absorbance, at different wavelengths). In that case the rows in the data matrix are spectra and the plots can be described as loading spectra. For the tablet example introduced in the preceding section, we plotted the loadings on PC2 against wavelength. The result is shown in Fig. 17.13. One observes that the highest loadings were obtained in the visible wavelength range: PC2 is largely determined by the colouring agent that was added to the tablets, since the wavelengths at which this substance absorbs receive the highest weights in determining

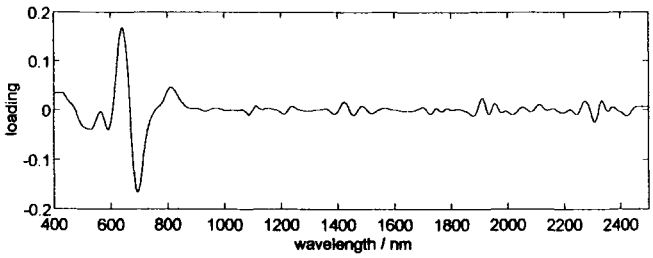


Fig. 17.13 Loadings on PC2 in function of wavelength for the tablets of Fig. 17.9.

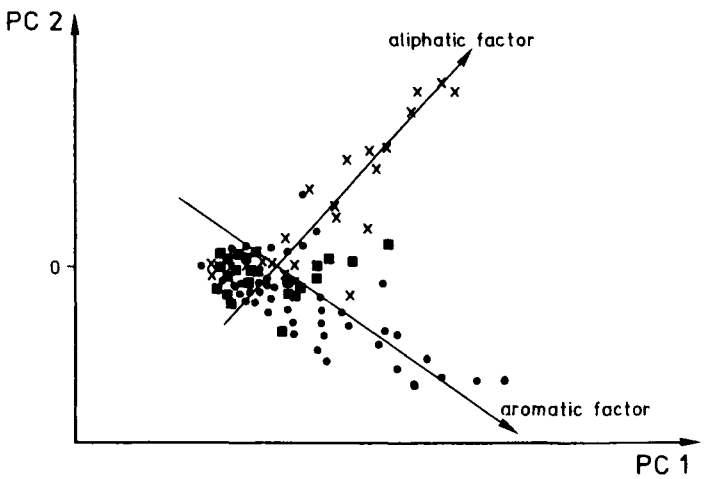


Fig. 17.14. Rotation of PCs for the air pollution example to obtain more meaningful factors.

the scores. Incidentally, it should be noted that we have now interpreted the first three PCs of the tablet example. The first describes differences in shape, the second in colour and the third in concentration of active substance.

We have explained the main PCs for both the air pollution and the tablet example. However, it should be noted that such an interpretation is not always possible and that the PC axes are not necessarily the best way to describe certain underlying characteristics of the data. Let us consider again the air pollution example. Although PC1 and PC2 allow to represent the objects, and the loadings can be interpreted to say something about the variables, the new axes introduced in Fig. 17.14 would physically be more meaningful. One could then identify an aliphatic variable and an aromatic variable and scores of the objects on these coordinates would give immediately a value for aromatic and aliphatic character. This requires additional rotations of the PCs. The techniques that are used to perform this are collectively called *factor analysis,* while the new meaningful

variables are called *fundamental variables* or *factors*. Factor analysis is one of the more important fields of chemometrics, and is described further in Chapter 34.

We now know that the scores, s, contain the information about the objects and that the loadings, v, contain the information about the variables. A very important result in this context is that:

$$\underset{n\times m}{\mathbf{X}} = \underset{n\times m}{\mathbf{S}}\ \underset{m\times m}{\mathbf{V}^{\mathrm{T}}} \tag{17.3}$$

The data matrix **X** can be decomposed into a product of two matrices, one of which contains the information about the objects (**S**) and the other about the variables (**V**). The decomposition is based on the singular value decomposition (see further Section 17.6). An illustration of the usefulness of this decomposition is given by Migron et al. [4]. They studied a 28×9 matrix consisting of the partition coefficients of 28 solutes between 9 solvents and water and decomposed the data into a matrix of solute properties and one of solvent properties.

The **S** matrix contains the scores of n objects on m principal components or latent variables. The **V**-matrix is a square matrix and contains the loadings of the m manifest variables on the m latent variables (see Fig. 17.15). By retaining only the significant components PC1 to PCk one restricts the information in the two matrices to structural information, i.e. one eliminates the irrelevant noise.

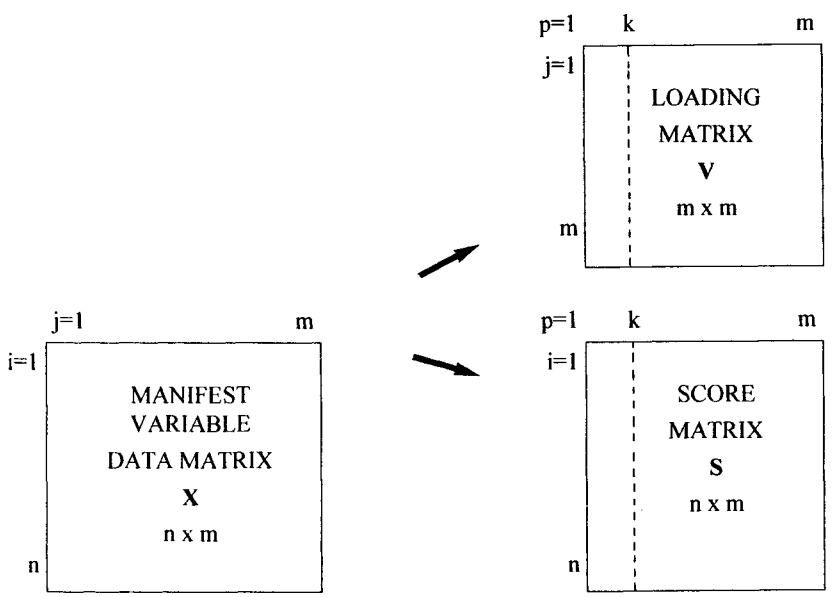


Fig. 17.15. Decomposition of the matrix of manifest variables in a score matrix and a loading matrix. Noise is eliminated by retention of the first k significant principal components, i.e. restricting **V** and **S** to the vectors to the left of the broken line.

17.4 Biplots

The methods we have described until now allow us to look at plots of either the objects or the variables. When variables and objects are displayed on the same plot this is called a *biplot*.

An example comes from the field of structure–activity analysis [6]. There are several biplot methods (see Chapters 31 and 32). The method applied here is called the method of spectral maps [7]. The activity investigated is the inhibition of rhinoviruses, i.e. the viruses that cause the common cold. Several candidate drugs were tested for their activity and one wants to know whether there is any relationship between the structure of the chemical compounds and the activity they exert. This is not a simple problem to solve because there are so many different rhinoviruses (serotypes). One hundred such serotypes were isolated and 15 different chemical compounds were tested with all one hundred serotypes. Their minimal inhibitory concentration (MIC value) was then measured. This is the minimal concentration of the drug required to inhibit a standard culture of a serotype. The reciprocal MIC values are a measure of antiviral activity. This yields a table consisting of 100×15 data. The biplot of the serotypes and the drugs is shown in Fig. 17.16. The chemical formulas of five of these substances are also given in the figure.

Let us first look at the objects (the serotypes). They are split up in two groups, A and B, along PC1. In this application, a log double-centring pretreatment of the data was carried out so that PC1 is not an indicator of general size (see Chapter 31). One then concludes that group A and group B have different patterns for the values of the variables; certain substances are more successful in inhibiting group A serotypes and others in inhibiting group B objects.

We should recall that, to interpret the effect of the variables, we must look at their direction from the origin. Certain drugs point in more or less the same direction, for instance, DCF and MDL. This means that they are correlated, so that they must be most effective against the same serotypes, namely the objects of group B. Indeed, DCF and MDL have high loadings on PC1. Because scores are sums of products of loadings and manifest variables, the scores of the objects along PC1 will be positive only if the values of the manifest variables, which exhibit positive loadings, are generally higher than those of the manifest variables with negative loadings. Therefore, DCF and MDL must indeed inhibit group B serotypes to a greater extent than group A serotypes. More generally, variables that point towards certain objects are more important for those objects. One says that they interact to a greater extent.

It is instructive to look at the structure of the substances interacting respectively with group A and B serotypes. Two of the chemical substances (SDS and WIN 51711) shown in Fig. 17.16 point to group A and the other three to group B. The general similarity of the structures of substances in group A on the one hand and

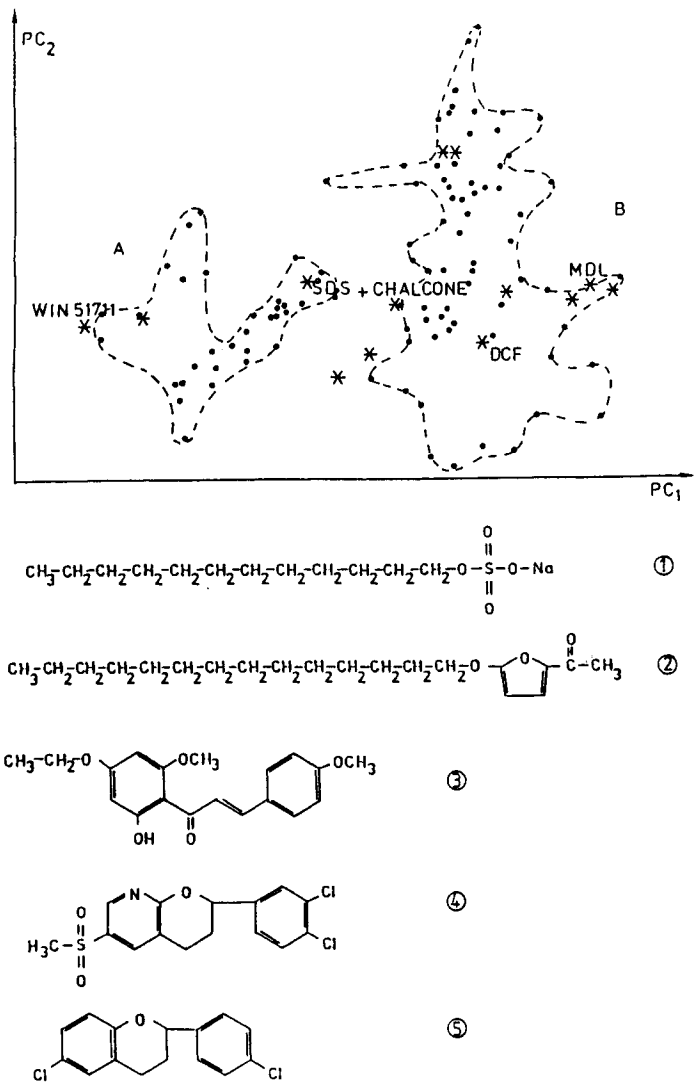


Fig. 17.16. Biplot of the rhinovirus data set [7]. The dots are the serotypes, the stars the variables and the cross is the origin of the biplot. Substances 1 and 2 are respectively SDS and WIN 51711. Substance 3 is chalcone and substances 4 and 5 are analogues of chalcone that interact with group B serotypes.

those in group B on the other is striking. The group A substances can be characterized as having a long aliphatic chain. Those of group B can be characterized as bulky and polycyclic. The spectral map analysis strongly suggests two different classes of serotypes interacting with different types of compounds. Group A serotypes are particularly sensitive to long aliphatic chain molecules. Group B serotypes are more sensitive to bulky polycyclic molecules.

Before explaining the hypothesis derived from this conclusion, it is necessary to explain somewhat more about rhinoviruses. The protein envelope of these viruses possesses a regular icosahedral structure. Each of the 20 faces of this icosahedron is built up of three different proteins. These proteins form canyons around each of the twelve symmetry axes located at the twelve vertices of the polyhedron and drugs can inhibit the virus by lodging themselves in these canyons. They thereby prevent the uncoating of the viral envelope, which is a necessary step for its replication inside the host cell. The hypothesis derived from the spectral map analysis is that there must be two different forms of canyons, which have evolved from a common ancestor. The A class of viruses possesses canyons which can accommodate elongated aliphatic molecules. The B class accepts only shorter and bulkier polycyclic molecules. The existence of two groups of rhinoviruses has recently been confirmed by means of sequence analysis of the envelope proteins.

One can conclude that biplots allow to investigate trends in a data table and that they show

- relationships or contrasts between objects, in our example between serotypes,
- relationships or contrasts between variables, in our example between compounds,
- relationships between variables and objects, in our example between serotypes and compounds.

17.5 Applications in method validation

17.5.1 Comparison of two methods

In Section 8.2.11 we studied the situation in which both the predictor and the response variable are subject to error. We concluded that instead of using ordinary least squares which minimizes residuals parallel to y (see Fig. 8.15a) we should minimize residuals orthogonal to the regression line (see Fig. 8.15b). Clearly, Fig. 8.15b is very similar to Fig. 17.3b. It should therefore not be a surprise that we concluded that computing PC1 of a data matrix consisting of 2 columns (the x and y values) and n rows (n = number of samples) yields the desired regression line. This was applied in method validation in Section 13.5.6 to compare results obtained with two different methods or in two different laboratories.

This application allows us to make an important comment about principal components in general. In Fig. 17.3b PC1 can be considered as a method to carry out regression between x_1 and x_2 by minimizing residuals orthogonal to the regression line. In general, therefore, obtaining principal components can be described as a regression procedure.

17.5.2 Intercomparison studies

In previous chapters we have often stressed that visual analysis of results should as much as possible precede or accompany statistical analysis. In Section 14.3.1 we described several visual aids in laboratory-performance (proficiency) studies. When several materials are analyzed by each laboratory or several methods are involved, PCA-based methods can be used for the same purpose.

An example can be found in the work of Rej et al. [8]. Their aim was to study clinical reference materials (CRM). These materials are not always well characterized. One needs to determine the concentration for the analyte for which the CRM is used and it frequently happens that there are several alternative methods available. It is not clear which of these methods is better or whether they are equally suitable for the analysis of different types of material. Rej et al. set up an experiment for the analysis of theophyline. The objects are 35 samples of three types (bovine serum CRM, human serum CRM and patient's material). The variables are the five different methods used.

Figure 17.17 gives the result of a correspondence factor analysis, which is a PCA after a pretreatment of the data (see Chapter 32). It is not our purpose here to discuss this figure in full detail, but merely to point out its usefulness by some examples. Along PC 1 we find for instance that fluorescent polarization immuno-assay (FPIA), high performance liquid chromatography (HPLC), substrate-labelled fluorescent immuno-assay (SLFIA) and the enzyme-multiplied immuno-assay technique (EMIT) are separated. It is found that this is due to consistently low results of EMIT compared with the others. The bovine and human CRMs, which

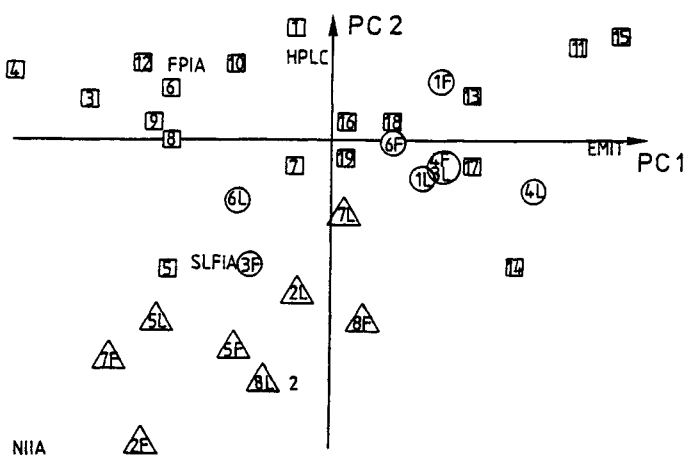


Fig. 17.17. Biplot for the determination of theophylline with five methods in 35 CRMs of 3 types (patient □, human (Δ) and bovine (○)) (adapted from Ref. [8]).

should give the same result, are also separated and one therefore knows that they yield different results with at least some of the methods. PC 2 is largely determined by NIIA and human serum CRMs. It was concluded that NIIA is sensitive to a certain type of background interference. SLFIA also points in the direction of the human CRM's and therefore interacts specifically with them.

Feinberg and Bugner [9] analyzed the results obtained for the determination of nitrate by 39 laboratories (the objects) on 7 types of plant material (the variables). The laboratories belonged to two different organizations, each organizing collaborative studies for their members, and employed different methods. Figure 17.18 shows a score plot. We can see that all ion selective electrode methods are found on the right and that the laboratories belonging to the two organizations (1–28, 29–38, respectively) are almost completely segregated along PC 1. The group of laboratories 29–38 delivers lower results than the other and the ion-selective methods give higher results than the other methods.

In both examples PC-based methods point out discrepancies or trends. They can therefore supplement or replace the visual methods of Chapter 14.

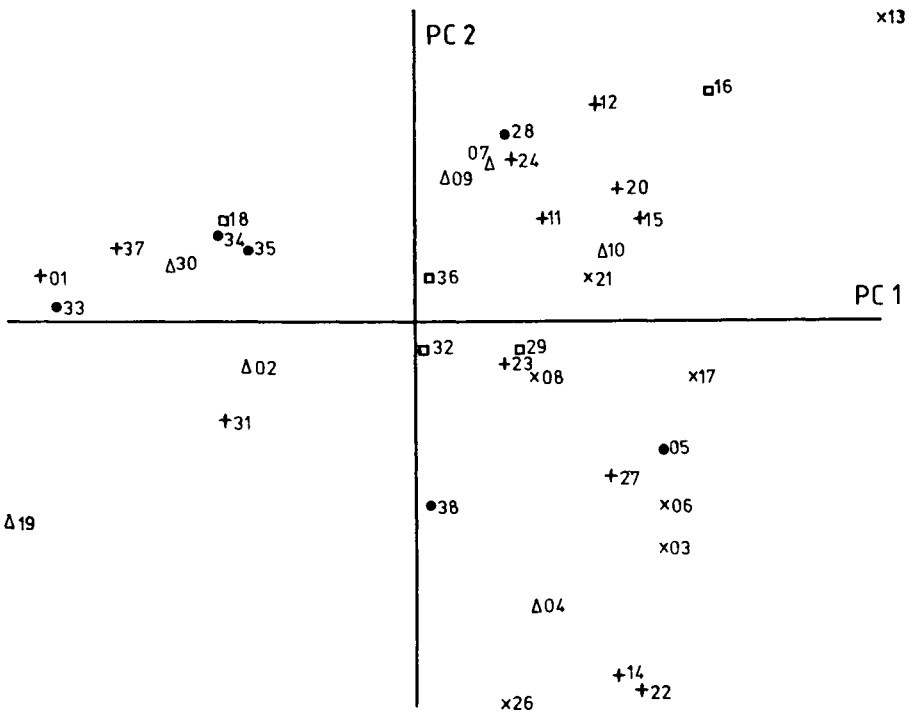


Fig. 17.18. Score plot for 39 laboratories characterized by their results for nitrate on 7 materials. Methods used: colorimetry (+), ion chromatography (●), distillation (Δ), ion-selective electrode (×), gas chromatography (□) (adapted from Ref. [9]).

17.6 The singular value decomposition

17.6.1 Eigenvalues

Any data matrix **X** can be decomposed according to the relationship

$$\mathbf{X} = \mathbf{U}\,\mathbf{W}\,\mathbf{V}^{\mathrm{T}} \tag{17.4}$$

where **U** is related to the scores of the objects, **V** is related to the loadings of the manifest variables and **W** is related to the variation explained by successive latent variables. Equation (17.4) describes what is known as the *singular value decomposition*. To understand what this relationship means and what **U**, **W** and **V** are exactly, we will consider a simple example.

The example is the one of Table 9.1. **X** is therefore a matrix consisting of $n = 5$ rows (or objects) and $m = 4$ columns (or variables). The variables are concentrations of the elements Al, Si, Mn and Fe, respectively.

$$\mathbf{X} = \begin{bmatrix} 200 & 300 & 100 & 360 \\ 380 & 580 & 420 & 840 \\ 200 & 320 & 400 & 380 \\ 500 & 760 & 250 & 1060 \\ 50 & 70 & 25 & 10 \end{bmatrix} \tag{17.5}$$

As already mentioned, it is customary to pretreat data before carrying out PCA and as we want to apply singular value decomposition to explain PCA, we will do this here too. The pretreatment we have chosen is column-centring (see also Chapters 9 and 31). This yields:

$$\mathbf{Z} = \begin{bmatrix} -66 & -106 & -139 & -188 \\ 144 & 174 & 181 & 292 \\ -66 & -86 & 161 & -168 \\ 234 & 354 & 11 & 512 \\ -216 & -336 & -214 & -448 \end{bmatrix} \tag{17.6}$$

Z has the same dimensions as **X** and we can carry out the singular value decomposition on **Z**:

$$\underset{n\times m}{\mathbf{Z}} = \underset{n\times m}{\mathbf{U}}\ \underset{m\times m}{\mathbf{W}}\ \underset{m\times m}{\mathbf{V}^{\mathrm{T}}} \tag{17.7}$$

We will not explain how to carry out singular value decomposition but only describe the results. Let us first consider **W**. This is a square diagonal matrix with dimensions $m \times m$:

$$\mathbf{W} = \begin{bmatrix} 1030.6 & 0 & 0 & 0 \\ 0 & 285.5 & 0 & 0 \\ 0 & 0 & 47.2 & 0 \\ 0 & 0 & 0 & 3.0 \end{bmatrix} \tag{17.8}$$

Only the elements on the diagonal are different from zero. From top to bottom we will call them $w_1, w_2, \ldots, w_m$. We observe that they are ranked such that $w_1 > w_2 > \ldots > w_m$. We also remember from Chapter 9 that diagonal matrices are often used for weighting purposes and we can therefore describe **W** as a weight matrix in which the weights are ranked in descending order of magnitude along the diagonal from top to bottom. This matrix is called the *singular values matrix.*

In PCA we will not use **W** as such but rather **Λ**. The matrix **Λ** is called the *eigenvalue matrix* and is given by:

$$\mathbf{\Lambda} = \begin{bmatrix} 1062000 & 0 & 0 & 0 \\ 0 & 81500 & 0 & 0 \\ 0 & 0 & 2230 & 0 \\ 0 & 0 & 0 & 9 \end{bmatrix} \tag{17.9}$$

It is a square matrix with as elements the *eigenvalues* λ for which:

$$\lambda_i = w_i^2 \text{ or in matrix notation } \mathbf{\Lambda} = \mathbf{W}^2 \tag{17.10}$$

The eigenvalues λ_k are of course ordered in the same way as the singular values w_k. The dimensions are again $m \times m$. The dimension of the square eigenvalue matrix is given by the number of manifest variables or, since this is equal to the number of latent variables, by the number of latent variables. It turns out that the eigenvalues are associated each with a PC. For instance λ_1 is associated with PC1 and is called the eigenvalue of PC1.

The eigenvalues represent the variation of the data along PC1, PC2, respectively. In Section 17.1 it was concluded that the variance along PC1 must be larger than that along PC2, etc. It can now be seen that the eigenvalue matrix is indeed constructed such that λ_1, associated with PC1 is larger than λ_2, associated with PC2, etc. In fact, the eigenvalues are the variances of the scores on PC1, PC2, etc.

The eigenvalues explain successively less information and are therefore associated with successively less important principal components. The sum of the eigenvalues is equal to the sum of all the variances along the principal components. In other words, the trace of the eigenvalue matrix is equal to the total variance of the data. A measure of the importance of each principal component can then be obtained by expressing it as a % variance explained. The % variance explained by PCk = $(\lambda_k/\text{trace}(\mathbf{\Lambda}))\cdot 100$.

This measure is sometimes used as a criterion to delete principal components in a feature reduction process. All components with less than X% explained variance (often $X = 80$, but there is no theoretical reason for that) are considered unimportant

and deleted. How to use eigenvalues to decide that PCs are significant is described in Section 17.7 and Chapter 34.

17.6.2 Score and loading matrices

U and **V** consist of a set of vectors called the left and right singular vectors. **V** is, in fact, the matrix of eigenvectors associated with the matrix of eigenvalues **Λ**. For our example it is given by

$$\mathbf{V} = \begin{bmatrix} 0.3392 & -0.0876 & 0.3684 & -0.8611 \\ 0.5185 & -0.0805 & 0.6850 & 0.5055 \\ 0.2082 & 0.9775 & -0.0227 & -0.0272 \\ 0.7568 & -0.1744 & 0.6282 & 0.0471 \end{bmatrix} \tag{17.11}$$

The columns of this matrix are the successive eigenvectors for PC1 to PC4 and the rows are the loadings for the manifest variables. The loading plot for PC1 and PC2 is given in Fig. 17.19a. Al, Si, Fe are found in the same direction from the origin and are therefore correlated. This can easily be verified by computing the correlation coefficient between the columns in **Z**. For 1 and 2 it is for instance equal to 0.997. Mn, on the other hand, is situated in another direction and therefore less correlated. Indeed r(Mn, Al) = 0.58. All loadings on PC1 are positive: PC1 is a size component. PC2 is dominated by a high loading for Mn (0.9775) and shows contrast, since there are both positive and negative loadings.

U is the normed score matrix. For our example, it is given by:

$$\mathbf{U} = \begin{bmatrix} -0.2412 & 0.3108 & 0.5150 & -0.6164 \\ 0.3761 & 0.3572 & -0.5578 & -0.4689 \\ -0.1558 & 0.6982 & 0.3951 & 0.3635 \\ 0.6333 & -0.4467 & 0.1443 & 0.4225 \\ -0.6124 & -0.2979 & -0.4967 & 0.2993 \end{bmatrix} \tag{17.12}$$

The columns are the normed scores on PC1 to PC4 for the five objects. The normed score plot is shown in Fig. 17.19b. One observes that, as expected, PC1 is a size component. The order from left to right of the scores on PC1 (5–1–3–2–4) is the order of the sums of concentrations of the trace elements from low to high. From the study of the loadings we expect PC2 to show contrast, mainly due to Mn, and, indeed, high scores on PC2 are found for objects 2 and 3 that have relatively high Mn concentrations.

Both **V** and **U** are orthonormal matrices. This means that they are *orthogonal* and *normed* (or normalised). A matrix is orthogonal when the sum of the cross-products of all pairs of column vectors is zero. One can, for instance, verify for the

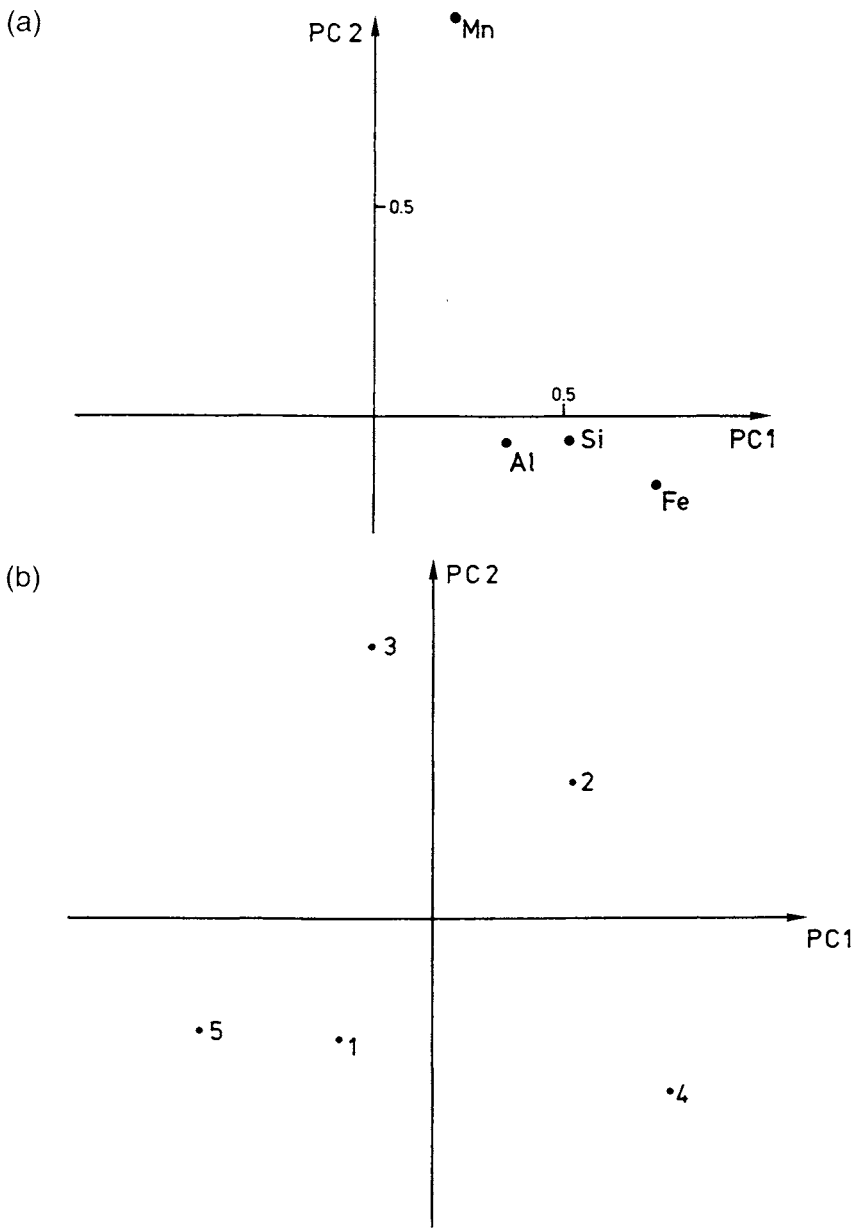


Fig. 17.19. Loading plot (a) and (normed) score plot (b) for the data of Table 9.1.

first two eigenvectors of **V** that ((0.3392 × – 0.0876) + (0.5185 × – 0.0805) + (0.2082 × 0.9775) + (0.7568 × – 0.1744) = 0. Two mean-centred vectors, whose sum of cross products is equal to zero are uncorrelated: PCA *decorrelates* the

variables (see also Section 17.8). A matrix is normed when the sum of the squared elements of a column vector is one. One can verify that, for instance, for the first eigenvector of **V** $(0.3392)^2 + (0.5185)^2 + (0.2082)^2 + (0.7568)^2 = 1$.

Multiplying the normed score matrix **U** with the weight matrix **W** yields the (un-normed) score matrix **S**

$$\underset{n\times m}{\mathbf{S}} = \underset{n\times m}{\mathbf{U}} \; \underset{m\times m}{\mathbf{W}} \tag{17.13}$$

so that we obtain the equivalent of eq. (17.2)

$$\underset{n\times m}{\mathbf{S}} = \underset{n\times m}{\mathbf{Z}} \; \underset{m\times m}{\mathbf{V}} \tag{17.14}$$

for the case where the PCA is done on pretreated data (**Z** instead of **X**). Since **W** is diagonal, this means that

$$s_{ip} = u_{ip}\, w_p \tag{17.15}$$

For our example, for instance, $s_{11} = u_{11} \cdot w_1 = -0.2412 \times 1030.6$, so that for object 1 the score on PC1 is –248.6. In practice, "score plots" are often normed score plots to take into account the difference in scale that would occur if one were to plot the un-normed scores.

One can summarize now that any matrix **X** (after transformation, **Z**) can be decomposed into a set of normed and orthogonal projections **U**, describing the locations of the objects in space, the normed and orthogonal vectors **V**, describing the relation between the old and new variables, and **W**, describing the amount of variance present in each eigenvector. The latter are related to the standard deviation of the projections on **V**.

By replacing **U W** by **S** in eq. (17.7) one obtains:

$$\underset{n\times m}{\mathbf{Z}} = \underset{n\times m}{\mathbf{S}} \; \underset{m\times m}{\mathbf{V}^{\mathrm{T}}} \tag{17.16}$$

This is the equivalent for **Z** of eq. (17.3) which we already proposed and commented on in Section 17.3. By multiplying both sides with **V**:

$$\underset{n\times m}{\mathbf{Z}} \; \underset{m\times m}{\mathbf{V}} = \underset{n\times m}{\mathbf{S}} \; \underset{m\times m}{\mathbf{V}^{\mathrm{T}}} \; \underset{m\times m}{\mathbf{V}} \tag{17.17}$$

For an orthogonal square matrix such as **V** one can write $\mathbf{V}^{\mathrm{T}} = \mathbf{V}^{-1}$ and by replacing $\mathbf{V}^{\mathrm{T}}$ by $\mathbf{V}^{-1}$ in eq. (17.17), we obtain again:

$$\underset{n\times m}{\mathbf{Z}} \; \underset{m\times m}{\mathbf{V}} = \underset{n\times m}{\mathbf{S}} \tag{17.18}$$

When not all eigenvectors are included, but only the k eigenvectors that are considered significant, eq. (17.16) is no longer correct, and we should write:

$$\underset{n\times m}{\mathbf{Z}} = \underset{n\times k}{\mathbf{S}} \quad \underset{m\times k}{\mathbf{V}^{T}} + \underset{n\times m}{\mathbf{E}} \tag{17.19}$$

where **E** is the matrix of errors, due to not including all eigenvectors.

17.7 The resolution of mixtures by evolving factor analysis and the HELP method

So far we have focused on PCA as a display method. However, PCA and related latent variable methods have found many more uses. In this and the following sections an introduction will be given to some of these. They will be discussed in much more detail in several chapters of Part B.

In this section, we will describe applications to the resolution of mixtures. Let us consider the case that the variables are wavelengths and the data absorbances measured at these wavelengths. Each object is then characterized by a spectrum, i.e. each row in the matrix is a spectrum. A possible question is whether all these spectra belong to the same substance and, if not, to how many species they belong. The problem is not simple because one has very little information to start with. We do not know the number of species, nor their pure spectra. How to solve such problems is described in Chapter 34. Here we will study a special case in which we have one additional element of information, namely that the objects follow some logical order. For instance, we are interested in acid–base or complex equilibria in function of pH and, to investigate this, we measure optical absorption for samples with increasing pH. The complexes have different spectra. This yields a table of optical absorption values at different wavelengths versus increasing (or decreasing) pH. The object of the experiment is to detect at which pH-values there is more than one species.

Another situation is found in HPLC with diode array detection (DAD). When a pharmaceutical company has produced a new active substance, it is necessary to make sure that that substance is sufficiently pure and one wants to develop a method to separate all impurities from the main substance. To validate this separation, one would like to establish that no impurity is hidden below the main peak. The DAD permits measurement of the full absorption spectrum at very short time intervals during the elution of a mixture. The data obtained can be written down as a matrix with, as rows, the spectra observed at t_1, t_2, etc., respectively.

The columns are the chromatograms as they would have been observed with fixed wavelength detectors at λ_1, λ_2, etc. The objects correspond to elution times, the variables correspond to spectral wavelengths. The objective of the experiment is to decide which rows (spectra) are pure, i.e. at which times only one species is present and which rows are not, meaning that there is more than one substance present.

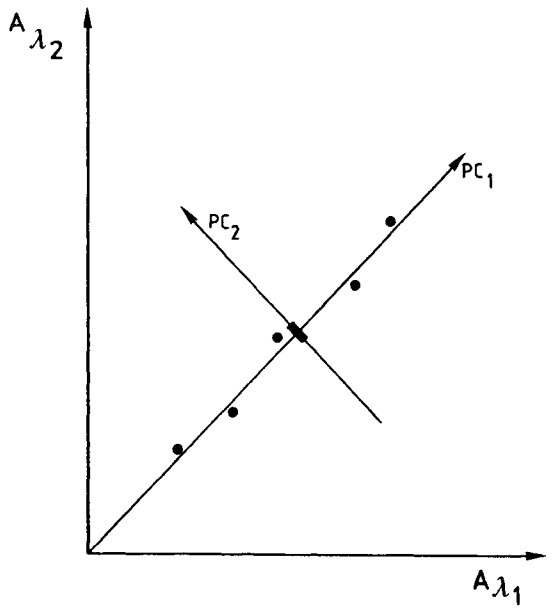


Fig. 17.20. Measurement points for one pure substance. The fat line is the range of scores on PC2.

To understand how to extract this information from such a data matrix, we will first turn to a simpler problem and suppose that there are only two wavelengths, λ_1 and λ_2. First we consider the situation where there is only one species in solution (Fig. 17.20). In the absence of measurement error or noise, the ratio of absorbances at the two wavelengths should be constant. In a plot of the absorbance at λ_1 against the absorbance at λ_2, the measurement points should therefore fall on a straight line. Because of measurement error there will be some spread around the line. A PC analysis of the data would yield one significant PC, PC1, as shown in the figure. The second PC, orthogonal to the first and having a small variance (see the range of scores), would explain only the noise and hence is not considered to be significant. The important thing to note is that there is one species and also only one significant principal component.

In Fig. 17.21a measurements for two different species A and B are plotted separately. Their spectra are different so that the ratios between absorbance at λ_1 and λ_2 are also different. If we consider each substance separately, the measurement points for each of the two species fall on two different lines. While pure species yield measurements situated around one of the two, such as X in Fig. 17.21b, mixtures of these species, such as point Y, yield points situated between them. The PC1 scores and PC2 scores obtained by PCA for all the points together are shown in Fig. 17.21c. The range of scores on PC2 is now much larger so that there are two significant principal components. More generally, if k species are

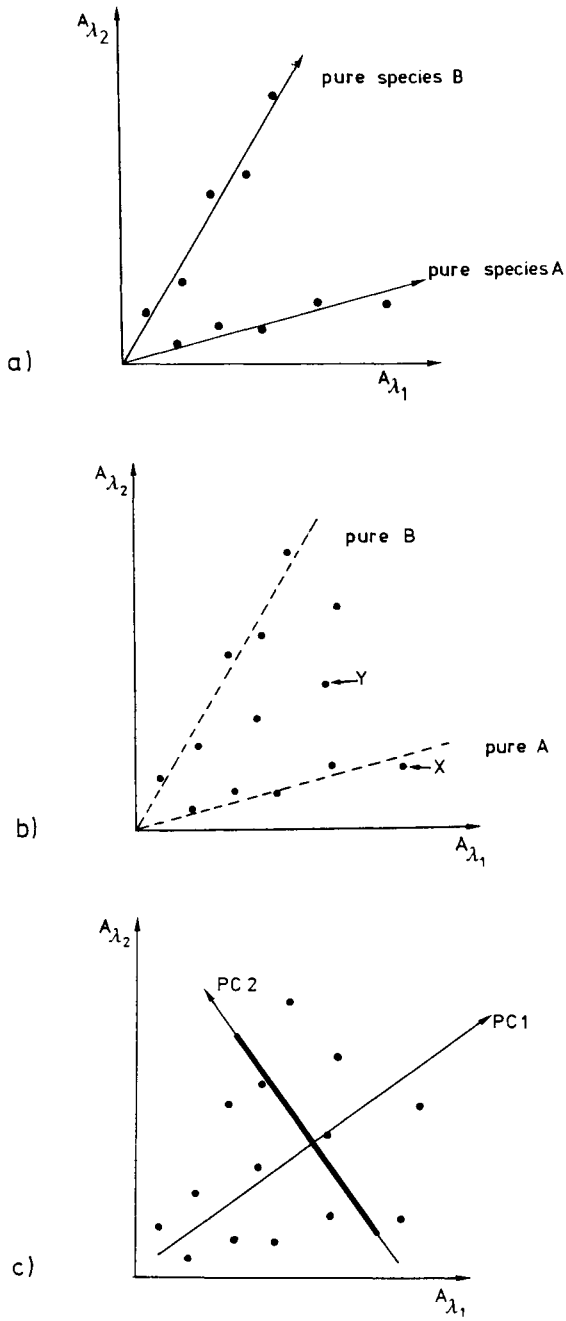


Fig. 17.21. (a) Measurement points for two pure substances. (b) Measurement points for two pure substances and mixtures of the two. (c) PC1 and PC2 for the data of b. The fat line is the range of scores on PC2.

present and one measures at at least k wavelengths, then one will obtain k significant components. It is now possible to state that more than one significant PC will be found if some of the time rows measure mixture spectra, and, more precisely that k significant PCs will be found when k species are present.

If one turns this around, it can be concluded that k significant PCs mean that k species are present. This is an interesting result but it leaves one question unanswered: namely how to decide that a PC is significant. Moreover, one would like to know at which times more than one species is present. This can be done, for instance, with *evolving factor analysis* (EFA) [11].

In Fig. 17.22 we give an example. If the major substance and the impurity have different absorption spectra then at t_6 a change in measured spectrum would occur, indicating that an impurity is present. In the simplified synthetic example shown here [10], detecting an impurity would not be difficult at all. However, one would still like to be able to detect impurities of the order of 0.1 to a few percent and in cases where the resolution between both substances is much smaller, so that overlap of the substances is complete. Moreover, it is probable that the substances are chemically related so that the spectra can be very similar.

EFA determines the eigenvalues of the first few PCs in matrices that describe increasing time windows. Suppose that we analyze first the time window consisting of the times t_1–t_5 as objects. As only one substance is present, one significant PC will be found. We then increase the time window by one unit, i.e. we now have 6 objects: t_1 to t_6. We then observe the emergence of a second significant PC, indicating that there is a second substance. This is still clearer when adding t_7, etc.

We are now faced with how to determine how many PCs are significant. The basic theory on that subject can be found in Malinowski [12]. However, in the type of situation described here, the visual approach of EFA performs better for practical reasons (too lengthy to describe here [10]). In EFA, the eigenvalues (or

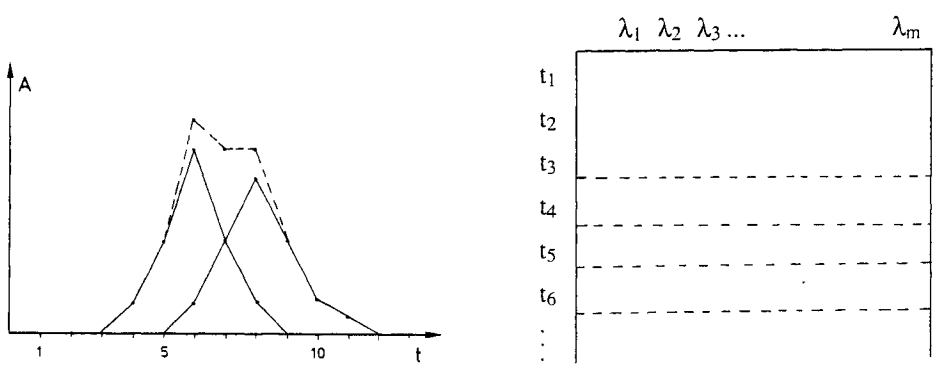


Fig. 17.22. Chromatogram and data table obtained with diode array detection. The broken lines denote successive matrices analyzed by evolving factor analysis.

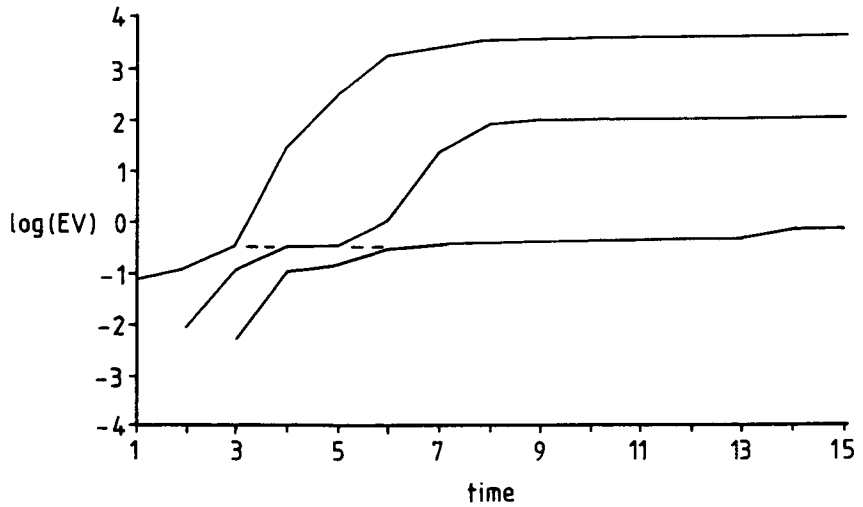


Fig. 17.23. Evolving factor analysis for the chromatogram of Fig. 17.22.

the log eigenvalues) of the first PCs are plotted as a function of time (Fig. 17.23). The first eigenvalue is not really of interest since we know that at least one substance must be present, but its first 3 values are representative for noise, since no substance is present yet. From t_4 there is a sharp increase: the first substance leaves the column. This means that the absorbance is larger, therefore also the variance in the data matrix and the first eigenvalue. The value of eigenvalue 2 is now representative for noise. At t_6 there is a clear increase out of the noise level of eigenvalue 2, indicating the presence of a second substance and noise is represented by eigenvalue 3. EFA is described more fully in Chapter 34. The method as applied originally to chromatography consists of a forward pass yielding the result shown in Fig. 17.23 and a backward run starting from the last spectra. The two are combined to decide where exactly impurities are found.

There is another simple way of looking at these data. Until now, we have placed the origin of the PCs in the centre of the data. This means that we have centred the data as described in Section 17.6. If we do not centre the data, i.e. we work on the original data, then the origin of the latent variable coordinate system is the same as that of the manifest variables and PCA is a simple rotation of the axes. In this case, this can be used to advantage. Figure 17.24 shows another chromatogram. Again, the chromatogram is simplified for didactical reasons, it would be very easy to decide that there are two substances present, and, again, we will first look at only two wavelengths. Points 7–8–9 are pure and therefore fall on a straight line through the origin and so do points 16–19. The other points are mixtures. The PC1 and PC2 through the origin (i.e. for uncentred data) are also shown. Figure 17.25 is the score plot on PC1 and PC2. We see that 7–8–9 and 16–19 are still situated on straight

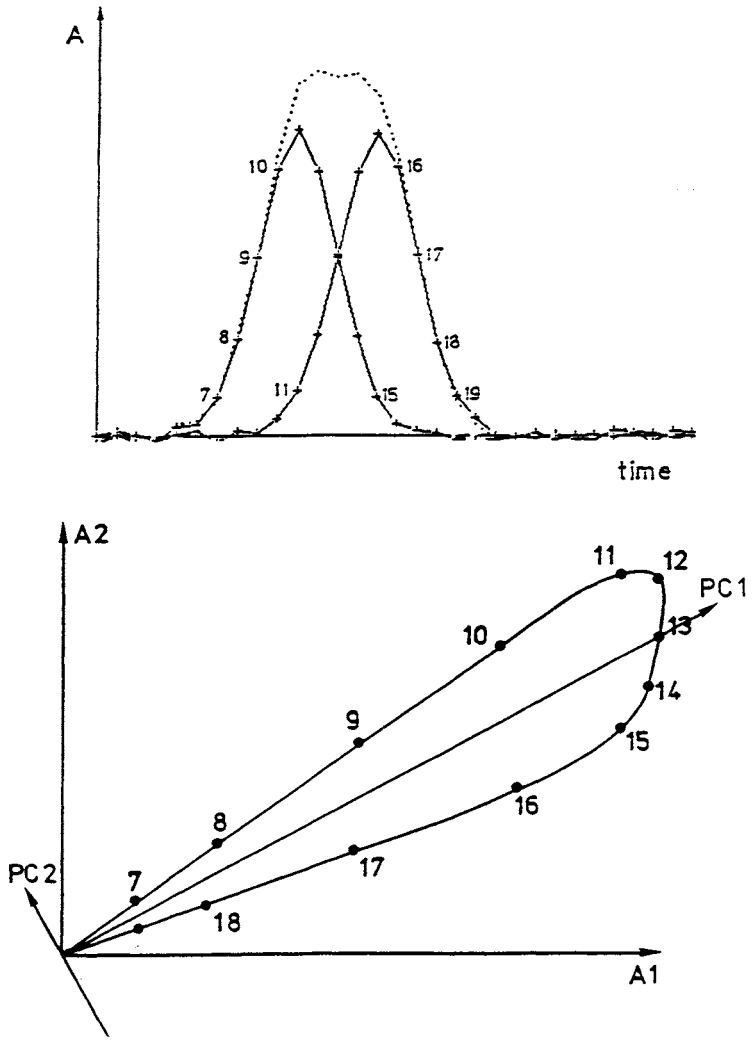


Fig. 17.24. Chromatogram measurement points obtained for two wavelengths and PC1 and PC2 for the uncentred data.

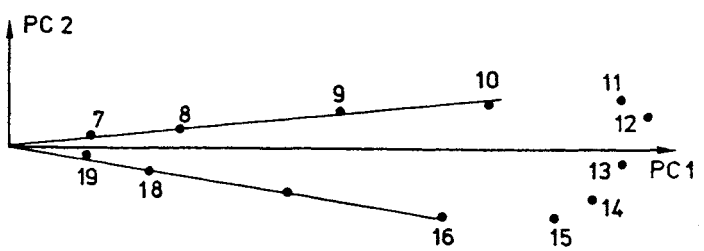


Fig. 17.25. PC1–PC2 score plot for the uncentred data of Fig. 17.24.

lines through the origin of this plot. We can again turn this around: points on a straight line through the origin indicate a pure zone in the chromatogram and in this case we would infer from the PC1–PC2 plot that there are two pure zones. This can be generalized to more than two wavelengths and is the basis of the *heuristically evolving latent projections (HELP)* method of Kvalheim et al. [13]. This method will be discussed further in Chapter 34.

17.8 Principal component regression and multivariate calibration

In Section 17.2, we studied an example concerning the score plot for a pharmaceutical example and we showed that the scores on PC3 are related to the concentration of the active substance. The principal components are latent variables and these can be entered into a (multiple) regression instead of the manifest variables. In the pharmaceutical example, it was eventually shown [3] that a good calibration can be obtained with PC3 and PC6 as the variables. Multiple regression is applied with the scores on these two principal components as the **X** data and the concentrations of the calibration samples as the **y** data. For the rest, one applies the techniques described in Chapter 10. For instance, the decision about how many principal components to include in the calibration equation is taken using the PRESS criterion described in Section 10.3.4. This type of regression is called *principal component regression (PCR)* and the application described is an example of *multivariate calibration*. Multivariate calibration is applied not only for the determination of the concentration of certain substances, but also for the direct determination of certain quality parameters. For instance, the sensory tenderness of beef was determined by PCR using near-infrared spectra as the manifest data [14]. Chapter 36 describes multivariate calibration in more detail.

There are two main advantages to PCR compared with multiple regression. The first is that the number of variables is reduced to only a few (feature reduction). In the pharmaceutical example there are 1050 manifest variables to choose from, but only 14 principal components, the number of principal components being limited to the number of objects (14 tablets were measured).

The second advantage is decorrelation. As explained in Chapter 10, the quality of prediction by multiple regression is adversely affected by correlation between the **X** variables. As we already stated in Section 17.6 the PCs are orthogonal, i.e. not correlated. Because of the fundamental importance of this concept, it is useful to elaborate it somewhat further. One of the many ways to describe the mathematical operations required to obtain the principal components is to state that one decorrelates the original variables. In Fig. 17.4a the data are plotted as a function of the original variables x_1 and x_2 and we observe that x_1 and x_2 are correlated. This is no longer the case in the score plot of Fig. 17.4b. The transformation from the

system of coordinates x_1, x_2 to PC1, PC2 can be described as a decorrelation of the variables (see also Section 17.6.2).

Let us describe this in a somewhat more formal way. Since $\mathbf{S} = \mathbf{U}\,\mathbf{W}$ (see Section 17.6.2) and, since for a diagonal matrix $\mathbf{W}^T = \mathbf{W}$:

$$\mathbf{S}^T\mathbf{S} = \mathbf{W}\mathbf{U}^T\,\mathbf{U}\mathbf{W} \tag{17.20}$$

Because $\mathbf{U}$ is an orthogonal matrix $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and

$$\mathbf{S}^T\mathbf{S} = \mathbf{W}^2 = \mathbf{\Lambda} \tag{17.21}$$

$\mathbf{S}^T\mathbf{S}$ is related to the covariance matrix (see Chapter 10) of the scores, so that we can indeed conclude, as already stated earlier, that $\mathbf{\Lambda}$ is related to the covariance matrix of the scores. $\mathbf{Z}^T\mathbf{Z}$ is related in the same way to the covariance matrix of the manifest (pretreated) variables and therefore one way of describing PCA is to say that the diagonal covariance matrix $\mathbf{\Lambda}$ is derived from the matrix $\mathbf{Z}^T\mathbf{Z}$ by *diagonalization*. That $\mathbf{\Lambda}$ is diagonal follows from the fact that principal components are not correlated, so that the covariances between principal components must be zero. The diagonal elements are the variances of the scores along the successive principal components. All the other elements of the matrix, the covariances, are zero.

17.9 Other latent variable methods

The use of latent variables is not restricted to PCA and the factor analytical techniques derived from it. In certain cases latent variables can be constructed according to different criteria than in PCA. We will briefly discuss two such techniques here, namely *partial least squares (PLS)* and *linear discriminant analysis (LDA)*.

PLS is an alternative to PCR for multivariate calibration. It can be used, for instance, in the tablet example. The PLS latent variables are also linear combinations of manifest variables, but the criterion to select weights is that the latent variable(s) describing the $\mathbf{X}$ data should have maximal covariance with the $\mathbf{y}$ data. One observes that the criterion is chosen keeping the aim of the method in mind, namely to model $\mathbf{y}$ as a function of $\mathbf{X}$.

When the $\mathbf{y}$ data consist of a single vector of data (in the tablet example, concentrations of an active compound), the method is called *PLS1*. When one relates two data tables to each other, one can apply *PLS2*. This can be explained with the following pharmacological example [15] (Fig. 17.26). The neuroleptics are a family of drugs used in psychiatry for the containment of many psychotic disorders. Each of these has different therapeutic effects on different disorders, that can be studied by clinical trials. This leads to different clinical spectra for each drug. These spectra together constitute one table. Such a table was collated for 17

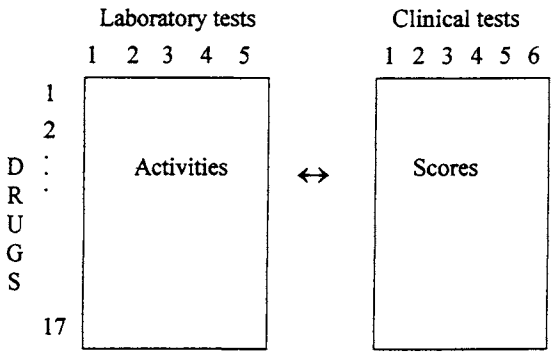


Fig. 17.26. Relating two tables that respectively describe the laboratory pharmacology profiles and the clinical spectra of 17 neuroleptic drugs.

neuroleptic drugs (the objects). The variables consisted of scores on a scale ranging from 0 to 5 for four different main effects (ability to antagonize delusion, for instance) and two side effects. The same neuroleptics were also characterized in the laboratory by tests on rats. The activity for five behavioral tests were determined. This constitutes a second table of 17 objects this time by 5 variables. The question then is whether one can relate the two tables and, more particularly whether one can predict the clinical spectrum from the laboratory one. In this case, latent variables are extracted from each of the tables with as criterion that the latent variables should show the highest covariance possible. Such problems are quite common. For instance, one can try to predict sensory characteristics of foodstuffs from physicochemical observations (see also Chapter 38) or pharmacological spectra of drugs from structural characteristics and/or molecular descriptors (see Chapter 37). Clearly, this type of problem is of very general importance and, for that reason, PLS and other methods that allow to relate two tables are described in Chapter 35.

One of the best known pattern recognition methods is linear discriminant analysis (LDA), which is also based on selecting appropriate latent variables. Let us go back to the food authentication example which we introduced in Section 17.2. Suppose that, instead of merely investigating whether we can distinguish 3 types of wine based on the measurement of 8 variables, we had been asked to develop a rule or set of rules that would have allowed us to make the discrimination. These rules should be based on the results of the 100 certified samples of known origin which have been analyzed. This question can be rephrased as follows: Use the samples of known origin (the learning samples) to derive a classification rule which allows us to classify samples of unknown origin (the test samples) in one of the (three) known classes. This is called *supervised pattern recognition* (Chapter 33). In LDA, one starts by reducing the number of variables. Consider Fig. 17.27. Two known classes are represented in a two-dimensional space and, as in Section

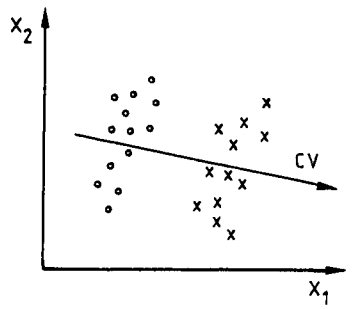


Fig. 17.27. Canonical variate (CV) describing the discrimination of two classes of objects of known classification.

17.1, we want to reduce the number of features to one. This means that we again need to determine a one-dimensional space (a line) on which the points will be projected. However, while PCA selects a direction which retains maximal structure in a lower dimension among the data, the criterion in LDA is a maximum discrimination among the given classes. The so-called *canonical variate* obtained in this way is also a linear combination of the original variables.

In Fig. 17.28 two canonical variates are plotted against each other for an example from clinical chemistry. The example [16] concerns the thyroid gland and the distinction between eu-, hypo- and hyperthyroid patients. To make the distinction five chemical determinations such as serum thyroxine were made. From the

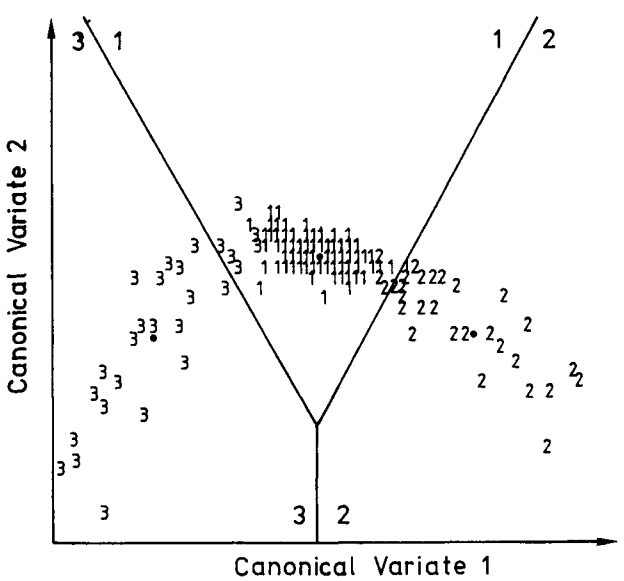


Fig. 17.28. Plot of the two canonical variates for three classes with different thyroid status (from [16]).

five manifest variables, the two canonical variates were obtained. In the two-dimensional space boundaries half-way between the centroids of each pair of classes can be drawn and new patients can be classified according to the scores on the canonical variates in one of the three classes.

References

1. D.L. Massart and P. Lewi, Principal Components, a Videoseries. Elsevier, Amsterdam, 1994.
2. M. Forina, C. Armanino, M. Castino and M. Ubigli, Multivariate data analysis as discriminating method of the origin of wines. Vitis, 25 (1986) 189–199.
3. D. Jouan-Rimbaud, B. Walczak, D.L. Massart, I.R. Last and K.A. Prebble, Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of NIR spectroscopic data. Anal. Chim. Acta, 304 (1995) 285–295.
4. J. Smeyers-Verbeke, J.C. Den Hartog, W.H. Dekker, D. Coomans, L. Buydens and D.L. Massart, The use of principal component analysis for the investigation of an organic air pollutants data set. Atmos. Environ., 18 (1984) 2471–2478.
5. Y. Migron, Y. Marcus and M. Dodu, Computer search for solvent and solute parameters that determine partition data. Chemom. Intell. Lab. Syst., 22 (1994) 191–197.
6. K. Andries, B. Dewindt, J. Snoecks, R. Willebrords, R. Stockbroeckx and P.J. Lewi, A comparative test of 15 compounds against all rhinoviruses as a basis for a more rational screening. Antiviral Res., 16 (1991) 213–225.
7. P.J. Lewi, Spectral map analysis. Factorial analysis of contrasts, especially from log ratios. Chemom. Intell. Lab. Syst., 3 (1988) 277–300.
8. R. Rej, R.W. Jenny and J.P. Bretaudière, Quality control in clinical chemistry: characterization of reference materials. Talanta, 31 (1984) 851–862.
9. M. Feinberg and E. Bugner, Chemometrics and food chemistry: data validation. Anal. Chim. Acta, 223 (1989) 223–235.
10. H.R. Keller and D.L. Massart, Evolving factor analysis. Chemom. Intell. Lab. Syst., 12 (1992) 209–224.
11. M. Maeder and A.D. Zuberbühler, The resolution of overlapping chromatographic peaks by evolving factor analysis. Anal. Chim. Acta, 18 (1986) 287–291.
12. E.R. Malinowski, Factor Analysis in Chemistry, 2nd Edn. Wiley, New York, 1991.
13. Y.Z. Liang, O.M. Kvalheim, H.L. Keller, D.L. Massart, P. Kiechle and F. Erni, Heuristic evolving latent projections: Resolving two-way multicomponent data. 2. Detection and resolution of minor constituents. Anal. Chem., 64 (1992) 946–953.
14. K.I. Hildrum, B.N. Nilsen, M. Mielnick and T. Naes, Prediction of sensory characteristics of beef by near-infrared spectroscopy. Meat Sci., 38 (1994) 67–80.
15. P.J. Lewi, B. Vekemans and L.M. Gypen, in: E.J. Karjalainen (Ed.), Scientific Computing and Automatisation (Europe). Elsevier, Amsterdam, 1990, pp. 199-209.
16. D. Coomans, M. Jonckheer, D.L. Massart, I. Broeckaert and P. Blockx, The application of linear discriminant analysis in the diagnosis of thyroid disease. Anal. Chim. Acta, 103 (1978) 409–415.

Chapter 18

Information Theory

18.1 Uncertainty and information

Information theory was initially developed to describe the amount of information in signals consisting of zeros and ones. In chemometrics, its initial use was to describe the amount of information yielded by a qualitative analysis. In data analysis, it is used to describe the uniformity of distributions. To introduce the subject, we will consider mainly qualitative analysis. Let us suppose we have a data bank with the identity of n substances described by a binary code (0 or 1), i.e. each substance has a binary identification number. How many substances can we describe with a single digit or bit? The answer is $2(2^1)$: we can distinguish the numbers 0 and 1. If two bits are available, the possible combinations are 00, 01, 10 and 11, in the decimal system 0, 1, 2 and 3. We can now distinguish 2^2 different substances. With N bits of 0/1 information we can distinguish $n = 2^N$ substances. This reasoning is not restricted to binary numbers. For instance when carrying out qualitative tests, we could say that if we have N tests available each with two outcomes (e.g. blue colour, no colour), then we have 2^N different combinations possible and this would allow us to distinguish between 2^N types of compounds.

Going back to the data bank example we can also state that if we have n substances to distinguish, we need $N = \log_2 n$ bits. The *information* I is a measure of the reduction of uncertainty. One of n possible substances can occur with equal probability and each is identified by a number from 0 to n. The probability that a certain event occurs, i.e. a certain identification number is obtained, is $p = 1/n$. Before an observation is done, each of the n instances can occur. After observing it, the substance in question is known and only one instance is possible.

The reduction of uncertainty is equal to:

$$I = \log_2 n - \log_2 1 = \log_2 n \text{ (bits)} \tag{18.1}$$

Since $p = 1/n$, we can re-write eq. (18.1) as

$$I = \log_2 (1/p) = -\log_2 p \tag{18.2}$$

Thus, we see that information theory is related to classical probability theory. In this connection, one also often uses the term *entropy* (H) and equates this to I.

Boltzmann has shown that entropy is a measure of the disorder in a physical system. In such systems the numerical value of disorder also measures the uncertainty about the state of an individual particle. To summarize we can therefore write that

$$I = H = \log_2 n = -\log_2 p \tag{18.3}$$

when all instances are equally probable and maximal information is obtained. More generally, when there is a set of possible identities before the experiment ($x_1, x_2, \ldots, x_j, \ldots, x_n$), each having a probability $p(x_j)$, the uncertainty before the experiment and thus also the information that is obtained by reducing this uncertainty to zero is given by

$$H = \sum_{j=1}^{n} - p(x_j) \log_2 p(x_j) \tag{18.4}$$

$$\text{with } \sum_{j=1}^{n} p(x_j) = 1$$

which is known as *Shannon's equation.*

Qualitative analytical methods are often referred to as "good", "valuable", "excellent", "specific", etc., with no further explanation of these terms. An objective interpretation of such terms is not easy and therefore the resulting choice of a method often does not have a fully rational basis. Whereas quantitative methods can be evaluated by using criteria such as precision, accuracy, robustness, detection limit, selectivity, and others discussed in the preceding chapters, no generally accepted criteria exist for qualitative analysis.

Information theory permits a mathematical evaluation of qualitative methods by calculation of the expected or average amount of information obtained from the analysis. Information theory can be applied to quantitative analysis but this has led to less practical results than its application to qualitative problems. A very complete overview is given by Eckschlager and Danzer [1]. A tutorial review was published by Clegg and Massart [2].

The aim of an analysis is to reduce the uncertainty with respect to the material (and therefore the system) to be analyzed. In qualitative analysis the analysis is carried out because there is an uncertainty about the identity of the components in the sample. After the analysis, the state of uncertainty is (hopefully) turned into a state of certainty (or, at least, of less uncertainty); in other words, the analysis has yielded a certain amount of information.

Let us apply the concepts we have introduced to derive the information yielded by a specific signal (or procedure). Again we assume a simple model of the analytical problem in which each of the possible identities has the same probability before analysis. Before the experiment, the uncertainty can be expressed in terms

of the number of possible identities, n_0, each identity having a probability $1/n_0$. Due to outcome i of the experiment (signal y_i), the number of possible identities is reduced to n_i, each with probability $1/n_i$. The information I_i obtained from this outcome is defined by

$$I_i = \text{initial uncertainty } (I_{\text{before}}) - \text{reduced uncertainty } (I_{\text{after}}) = \log_2 n_0 - \log_2 n_i = \log_2(n_0/n_i) = -\log_2 p(y_i) \quad (18.5)$$

I_i is called the specific information of outcome i; y_i is the signal that leads to outcome i. When all identities are equally probable, then $p(y_i) = n_i/n_0$. Therefore $I_i = -\log_2 p(y_i)$ as indicated in eq. (18.5).

A numerical example will illustrate the concepts introduced so far. Let us assume that the analyte to be identified is known to be one of 100 possible substances and that the measurement yields a signal, which is known to be obtained with 10 of the possible substances. We will call obtaining this signal a positive (+) result. Then, the application of eq. (18.5) leads to the specific information $I_i = \log_2(100/10)$ or $I_i = -\log_2(0.1) = 3.32$ bits.

If the test is negative (–), i.e. one does not obtain the signal which is considered to be positive, this also yields some information. It excludes that any of the 10 substances yielding that signal is present and reduces uncertainty in the sense that instead of 100 possible identities, there are now only 90 left. The specific information when one obtains a (–) result, is therefore $I_i = \log_2 (100/90) = 0.15$ bit. The two results are not equally probable: on average the probability of obtaining the (+) signal is 0.1, the probability of a (–) result is 0.9.

The average of the specific information obtained, sometimes also called *information content of the test (signal, procedure)* is then $I = (0.1 \times 3.32) + (0.9 \times 0.15) = 0.47$ bits (assuming that all 100 substances occur with the same probability). In equation form this can be written as

$$I = \frac{n_1}{n_0} \log_2 \frac{n_0}{n_1} + \frac{n_2}{n_0} \log_2 \frac{n_0}{n_2}$$

with values of n_0, n_1 and n_2 of 100, 10 and 90, respectively. It is customary to write n_0 in the denominator

$$I = -\frac{n_1}{n_0} \log_2 \frac{n_1}{n_0} - \frac{n_2}{n_0} \log \frac{n_2}{n_0} \quad (18.6)$$

Sometimes it is convenient to write this as

$$I = -p^+ \log_2 p^+ - p^- \log_2 p^- \quad (18.7)$$

where p^+ is the probability for the (+) outcome.

After generalization to more than two possible responses, eqs. (18.6) and (18.7) become:

$$I = \sum_i -\frac{n_i}{n_0} \log_2 \left(\frac{n_i}{n_0}\right) = -\sum_{i=1}^{n} p(y_i) \log_2 p(y_i) \quad (18.8)$$

where I is the information content of the procedure, n the number of possible outcomes (classes), n_0 the number of possible identities before the experiment (with equal probabilities), n_i the number of possible identities after interpretation of the experiment with result y_i, and $p(y_i)$ is the probability of measuring a signal y_i. This equation has the same form as eq. (18.4). It is Shannon's equation, now for the information content of a certain procedure. It should be noted that only those classes where $n_i > 0$ are taken into account. It should be stressed here that the value of the information content is only of interest when used in a relative way, i.e. as a means to compare the performance of one qualitative procedure with another.

A more general model can now be introduced. This model represents a set of possible identities before the experiment (x_1, x_2, ..., x_j, ..., x_n), each having a probability $p(x_j)$. As explained before, the uncertainty before the experiment, and thus the information that is obtained by reducing this uncertainty to zero, I_{before}, can be expressed by means of the Shannon equation

$$I_{\text{before}} = \sum_{j=1}^{n} - p(x_j) \log_2 p(x_j) \tag{18.9}$$

After the experiment with result y_i

$$I_i = \sum_{j=1}^{n} - p(x_j|y_i) \log_2 p(x_j|y_i) \tag{18.10}$$

where $p(x_j|y_i)$ is the (conditional) probability (also called Bayes' probability, see Chapter 16) of identity x_j, provided that the experiment has yielded a signal y_i (i = 1, ..., m). The uncertainty, I_i, remaining after a signal y_i was obtained depends, of course, on the signal measured. The difference $I_{\text{before}} - I_i$ is equal to the specific information. In order to derive an equation for the information content, we have to subtract the weighted average I_i from I_{before}, which leads to the expression

$$I = I_{\text{before}} - \sum_{i=1}^{m} p(y_i) I_i \tag{18.11}$$

where $p(y_i)$ is the probability of measuring a signal y_i. By making use of eqs. (18.9) and (18.10), eq. (18.11) can be written as

$$I = \sum_{j=1}^{n} - p(x_j) \log_2 p(x_j) - \sum_{i=1}^{m} p(y_i) \sum_{j=1}^{n} - p(x_j|y_i) \log_2 p(x_j|y_i) \tag{18.12}$$

Calculation of the information content generally requires a knowledge of the following probabilities.

(a) The probabilities of the identities of the unknown substance before analysis, $p(x_j)$. The first term on the right-hand side of eq. (18.12) represents what is known about the analytical problem in a formal way, or the prior information. A definition of the analytical problem in terms of the probabilities $p(x_j)$ is essential for calculating

the information content. An infinite number of possible identities each having a very small probability (approaching zero) represents a situation without pre-information. The uncertainty is infinitely large and solving the analytical problem requires an infinite amount of information.

(b) The probabilities of the possible signals, $p(y_i)$. These probabilities depend on the relationship between the identities and the signals (tables of melting points, R_f values, spectra, etc.) and also on the substances expected to be identified, $p(x_j)$. If an identity is not likely to be found, the corresponding signal is not likely to be measured. It should be noted that, in replicate experiments, one identity can lead to different signals because of the presence of expérimental errors.

(c) The probabilities of the identities when the signal is known, $p(x_j|y_i)$. In fact, these probabilities are the result of the interpretation of the measured signals in terms of possible identities. To this end, the following relationship can be used for interpretation.

$$p(x_j|y_i) = \frac{p(x_j) \cdot p(y_i|x_j)}{\sum_j p(x_j) \cdot p(y_i|x_j)} \tag{18.13}$$

This relationship shows that the probabilities for the identities after analysis can be calculated from the pre-information, $p(x_j)$, and the relationships between the identities and the signals, $p(y_i|x_j)$. Eq. (18.13) is in fact Bayes' theorem, already described in Chapter 16. This result again stresses the relationship between information theory and probability theory.

18.2 An application to thin layer chromatography

If we assume that substances of which the R_f values differ by 0.05 can be distinguished, the complete range of R_f values can be divided into 20 groups (0–0.05, 0.06–0.10, ...). Such a simplified model leads to a situation where substances with R_f values of, for instance, 0.05 and 0.06 are considered to be separated, which clearly is not true. However, it is not important here to distinguish exactly which substances are separated and which are not, as the purpose is rather to see how well the substances are spread out over the plate.

Each of the 20 groups of R_f values can then be considered as a possible signal (y_1, y_2, ..., y_{20}) and there is a distinct probability [$p(y_1)$, $p(y_2)$, ..., $p(y_{20})$] that an unknown substance will have an R_f value within the limits of one of the groups. Let us consider a TLC procedure that is used to identify a substance belonging to a set of n_0 substances; n_1 substances fall into R_f group 1, n_2 into group 2, etc. If all substances have the same *a priori* probability to be the unknown compound, eq. (18.8) can be used for calculating the information content. To understand further the meaning of the information content, let us investigate some extreme conditions.

TABLE 18.1
R_f values (× 100) of DDT and related compounds and information content of the proposed separation [2]

Solvent system	*p,p'*-DDT	*o,p'*-DDT	*p,p'*-DDE	*o,p'*-DDE	*p,p'*-DDD	DDA	DDMU	DBP	Kelthane	DPE	DBH	BPE	DDM	*I*
1	25	35	41	36	10	0	32	2	0	27	0	0	35	2.28
2	48	55	61	65	28	0	55	8	2	44	0	0	54	2.47
3	69	72	75	72	52	0	67	24	3	63	0	0	67	2.35
4	76	76	77	75	64	0	74	45	8	72	2	4	75	2.50
5	67	69	72	69	52	0	70	69	7	66	1	2	71	1.98
6	67	70	75	68	51	0	72	45	10	66	2	4	70	2.87
7	70	70	75	70	63	0	74	61	23	66	6	13	69	2.78
8	78	79	83	79	77	6	83	75	53	77	27	39	80	2.56
9	69	71	76	69	60	0	73	56	19	67	4	9	71	3.03
10	35	44	49	45	16	0	48	4	0	35	0	0	42	2.62
11	58	63	65	62	42	0	62	18	3	55	0	0	61	2.41
12	60	64	71	64	43	0	69	50	8	60	2	5	67	2.71
13	63	66	71	64	48	0	68	55	16	62	4	9	67	2.97
14	73	74	77	72	65	0	75	68	48	77	24	36	76	2.57
15	83	84	85	83	78	5	83	81	58	82	46	55	83	1.47
16	84	84	84	83	80	0	84	82	75	84	71	74	83	1.70
17	87	88	89	87	81	20	87	83	64	86	59	63	86	1.47
18	59	68	73	69	39	8	67	34	34	60	33	34	67	2.31
19	92	94	96	93	83	32	92	82	72	91	61	70	93	2.35
20	78	80	80	77	69	17	78	67	43	74	24	35	78	2.50
21	82	82	82	80	81	0	81	81	81	81	80	81	82	0.39
22	80	80	80	77	79	5	78	78	79	79	79	79	80	0.39
23	35	45	50	45	13	0	40	2	0	36	0	0	39	2.35
24	40	51	52	48	16	0	44	3	0	42	0	0	42	2.28
25	71	74	77	72	59	0	74	64	21	70	5	13	72	2.72
26	77	77	78	77	72	30	78	72	56	78	43	47	78	2.03
27	27	40	43	42	10	0	35	2	0	32	0	0	35	2.08
28	70	75	77	75	51	2	74	20	4	69	4	4	75	2.04
29	65	69	76	68	48	0	72	57	7	64	2	5	70	2.93
30	85	85	85	84	84	4	80	83	84	84	83	84	84	0.77
31	54	67	74	69	22	0	61	6	6	54	6	6	62	2.62
32	100	100	100	100	94	35	100	94	92	100	83	92	100	1.57
33	93	94	96	94	90	7	96	93	75	94	67	70	93	1.88

(a) All substances fall into the same group n_i. In this instance $n_i/n_0 = 1$ and thus $I = 0$. As all of the substances yield the same R_f value, the experiment does not indicate anything to the observer. No information is obtained because there is no uncertainty as to which event (signal, R_f value) will occur: whatever the unknown substance, the result will always be the same.

(b) All n_0 substances fall into different groups. The information content is maximal as each substance yields a different R_f value. The information content, from eq. (18.8) with all $n_i = 1$, is equal to

$$I = -n_0 \frac{1}{n_0} \log_2\left(\frac{1}{n_0}\right) = \log_2(n_0)$$

This is the maximum value which can be obtained. It is equal to the information necessary to obtain an unambiguous, complete identification of each substance (eq. 18.1). In general, the maximum value for I is obtained when the substances are spread as evenly as possible over the 20 signal categories, i.e. the distribution is as uniform as possible.

From these extreme conditions, it follows that in order to obtain a maximum information content the TLC system should cause an equal spread of the R_f values over the entire range. The results for an application [4] concerning DDT and related compounds are summarized in Table 18.1. The computation of I for system 13 is given as an example in Table 18.2. The results in Table 18.1 were based on slightly different and more complex definitions of R_f groups, which explains the difference between the 2.82 bit of Table 18.2 and the 2.97 bit of Table 18.1. The best separations are obtained with solvents 9, 13 and 29.

TABLE 18.2

Example of computation of the information content

R_f groups	Substances falling in the group	n_i/n_0	$\log_2(n_i/n_0)$	$(-n_i/n_0)\log_2(n_i/n_0)$
0–0.05	DDA,DBH	2/13	–2.697	0.415
0.06–0.10	BPE	1/13	–3.715	0.286
0.16–0.20	Kelthane	1/13	–3.715	0.286
0.46–0.50	*p,p'*-DDD	1/13	–3.715	0.286
0.51–0.55	DBP	1/13	–3.715	0.286
0.61–0.65	DPE, *p,p'*-DDT, *o,p'*-DDE	3/13	–2.114	0.488
0.66–0.70	*o,p'*-DDT, DDMU, DDM	3/13	–2.114	0.488
0.71–0.75	*p,p'*-DDE	1/13	–3.715	0.286
				$I = 2.82$

18.3 The information content of combined procedures

The objective of a qualitative analysis is to obtain an amount of information permitting unambiguous identification. In practice, this is often not possible with a single test and therefore experiments have to be combined. For example, in toxicological analysis of basic drugs, one will combine techniques such as UV and IR spectrometry, TLC and GLC or one will use two (or more) TLC procedures, etc. in order to obtain the necessary amount of information. Hence, the next question is how to calculate the information content of two or more methods.

When the information of two TLC systems is combined, one can consider the combination of two R_f values which fall in the range 0.00–0.05 as one event (signal y_{11}), an R_f value of 0.00–0.05 for system 1 and 0.05–0.10 for system 2 as a signal y_{12}, etc. As before, one can define a probability $p(y_{ij})$ for signal y_{ij} so that $n_{ij}/n_0 = p(y_{ij})$. In the general case of system 1 containing m_1 classes and system 2 containing m_2 classes, eq. (18.8) can be converted into

$$I = -\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{n_{ij}}{n_0} \log_2 \left(\frac{n_{ij}}{n_0} \right) \tag{18.14}$$

At first sight, one might assume that I is the sum of the information content of the systems 1 and 2

$$I = I(1) + I(2) \tag{18.15}$$

This is true only if the information yielded by systems 1 and 2 is not correlated, i.e. if no part of the information is redundant. This can be understood more easily by considering a simple example represented by the R_f values for eight substances in three different solvents (Table 18.3).

TABLE 18.3

R_f values of eight substances in three different solvents

Substance	Solvent I	Solvent II	Solvent III
A	0.20	0.20	0.20
B	0.20	0.40	0.20
C	0.40	0.20	0.20
D	0.40	0.40	0.20
E	0.60	0.20	0.40
F	0.60	0.40	0.40
G	0.80	0.20	0.40
H	0.80	0.40	0.40
Information content (bit)	2	1	1

With solvent I, one obtains 2 bits of information, while 3 bits are necessary for the complete identification of each possible substance. Solvents II and III each yield 1 bit of information. First running a plate with solvent I and then with solvent II does, indeed, permit complete identification: 3 bits are obtained with this combination. Although solvents I and III have clearly different R_f values, the combination of I and III does not yield any more information than that obtained with solvent I. The information content of a procedure in which both solvents are used is still 2 bits. Both of these cases are extreme. The combination of two TLC procedures, and in general of any two procedures, will lead to an amount of information less than that which would be obtained by adding the information content of both procedures but at least equal to the information content of a single procedure. In practice, it is improbable that two chromatographic systems would yield completely uncorrelated information and even when combining methods such as chromatography and spectrophotometry some correlation should be expected. As a result of these correlations, the measurement of two (or more) physical quantities yields partly the same information (also called *mutual information*).

For instance, both melting and boiling points usually increase with increasing molecular weight. When a high melting point has been observed, the boiling point is also expected to be high. If the correlation between melting point and boiling point were perfect, it would make no sense to determine both quantities for identification purposes. However, the correlation is not perfect because melting and boiling points are not determined solely by the size of the molecule but are also governed by factors such as its polarity. From this crude physical description, it is clear that the measurement of the boiling point will yield additional information even if the melting point is known. However, this additional amount of information is smaller than that obtained in the case of an unknown melting point.

The most important conclusions from this section are that the highest information content for individual systems is obtained when the substances are distributed evenly over the different classes and that, for combinations of methods, the correlated information should be kept as low as possible.

Neither conclusion is surprising. Analytical chemists know that a TLC separation is better when the substances are divided over the complete R_f range and they also understand that two TLC systems in combination should not be too similar. However, information theory allows one to formalize this intuitive knowledge and to quantify it, so that an optimal method can be devised.

Information theory can, for instance, be used as a tool for optimizing combinations of analytical methods [5,6]. For instance, in gas chromatography, many different liquid stationary phases have been described (several hundreds). Quite clearly, one does not want to keep in stock so many stationary phases because many of these have the same separation characteristics or, to put it in the terminology

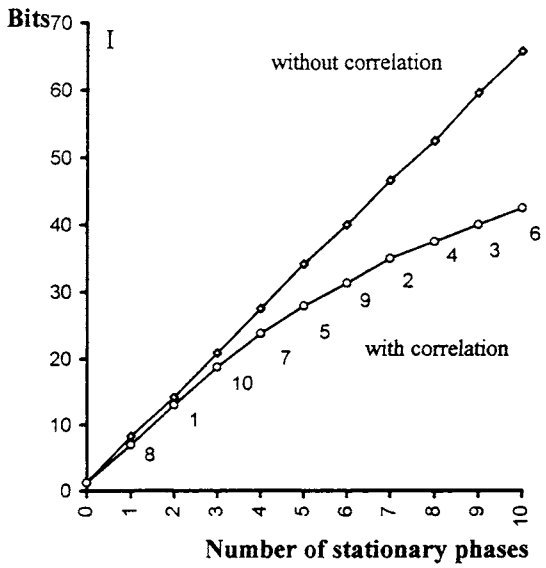


Fig. 18.1. Amount of information as a function of the number of gas-chromatographic stationary phases applied to a certain separation problem [6].

used in this chapter, yield correlated information. Figure 18.1 gives the amount of information which can be extracted by using gas chromatography for a particular set of compounds. The best stationary phase, i.e. the one with the highest information content, yields 7 bits for this particular group of compounds. In fact, in this particular instance, many of the stationary phases yield about the same quantity of information so that a combination of 5 stationary phases could theoretically yield 35 bits. However, as the figure shows, only 28 bits are obtained. The 7 bits of information lost are due to correlation. The mathematics used to obtain this result are quite complex and of minor importance for our purpose. Of greater importance, however, are the rules for qualitative analysis which can be derived from them. These rules are as follows.

(1) Optimal combined methods for qualitative analysis make use of individually good systems.

(2) Individually good systems are characterized by high spread of the analytical signals (for instance, many different R_f values) and low errors (substances with small difference in R_f value can be discriminated).

(3) Optimal combinations also require that the individual systems should yield uncorrelated information, or, in other words, show informational orthogonality [7] meaning that very dissimilar systems should be combined. This explains the power of combinations of methods with very different principles, such as GC/MS (combination of a chromatographic method with a spectrochemical one).

18.4 Inductive expert systems

In Chapter 43, we will see that deductive expert systems use rules, given by an expert, to make conclusions. We will also see that the requirement that rules should be given constitutes a bottleneck. To avoid this problem so-called inductive expert systems derive the rules from examples. The best known algorithm, Quinlan's Id3 algorithm [8], applies information theory for this purpose. It can also be described as a method for supervised pattern recognition (see Chapter 33).

Let us explain it with a simple example based on Ref. [2]. We need to make rules to decide whether a certain sample belongs to class A or to class B. One knows that the *a priori* probabilities are equal. A set of samples of known origin, which we will call the training set, representative for the population and consisting of 50% A samples and 50% B samples, is available to develop the rules. Each sample is characterized by a set of variables. The initial uncertainty is given by eq. (18.7)

$$I = -p(\mathrm{A}) \log_2 p(\mathrm{A}) - p(\mathrm{B}) \log_2 p(\mathrm{B})$$

$$= -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1 \text{ bit}$$

This is the maximum uncertainty possible for a 2 class situation and we want to reduce it, if possible, to 0, i.e. we want to develop rules such that we will be able to state with certainty (i.e. $p = 1$) that an unknown belongs to a certain class. We therefore need to gain:

$$I = 1 - 0 = 1 \text{ bit}$$

Let us suppose that we have selected a certain variable x and a threshold value x_T, which we think would give us some information. Let us call x_+ all values of $x > x_T$ and x_- all values $x < x_T$. Using the training set, it is found that 5/8 of the samples have the value x_+ and 3/8 have the value x_-. Of the x_+ class samples 60% are found to belong to A (i.e. 3/8 of the whole sample set) and 40% are B (i.e. 2/8 of all samples). Of the x_- class samples one third is A (i.e. 1/8 of all samples) and two thirds B (i.e. 2/8 of all samples).

The uncertainty remaining when one has obtained the x_+ result is (see eq. 18.10):

$$I_+ = -0.4 \log_2 0.4 - 0.6 \log_2 0.6 = 0.971 \text{ bit}$$

When the result was x_- then the remaining uncertainty is:

$$I_- = -0.33 \log_2 0.33 - 0.67 \log_2 0.67 = 0.918 \text{ bit}$$

Taking into account that the x_+ result is more probable than the x_- result, the average remaining uncertainty would be:

$$H_a = (5/8 \times 0.971) + (3/8 \times 0.918) = 0.951 \text{ bit}$$

and the information gained is

$$I = 1 - 0.951 = 0.049 \text{ bit}$$

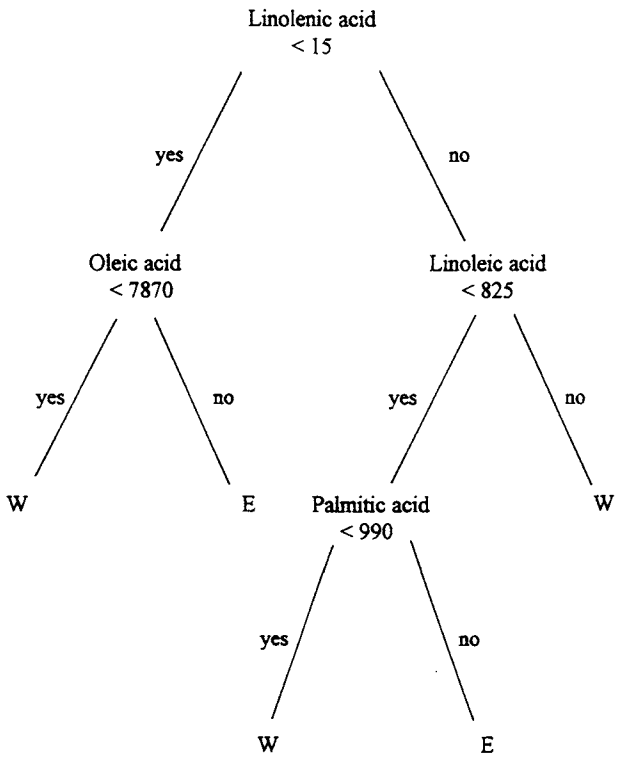


Fig. 18.2. Decision tree for the classification of East (E) and West (W) Ligurian olive oils, based on their fatty acid pattern [9].

Quinlan's algorithm sequentially selects variables and thresholds that cause the highest gain in information, until complete certainty is obtained or no further gain can be reached. First it selects the variable that yields the most information. For a continuous variable, this means that it finds the threshold at which the gain in information is highest. When this results in + or – classes that do not consist of pure A or pure B, that class is further separated in two new sub-classes using the variable and the threshold that increase the information most for that class. The result is a decision tree. An example is given in Fig. 18.2 taken from Derde et al. [9]. Hopke et al. [10] used the algorithm to develop rules for the classification of particles, collected during air-pollution studies on the basis of scanning electron microprobe results. Scott [11] applied it to classification of compounds of environmental interest and Tandler et al. [12] used it for monitoring products from polyethylene cracking. An iterative version of the algorithm, which allows among others, pruning away unnecessary branches of the tree is known as the C 4.5 algorithm [13]. Other tree-based classification methods for exploratory data analysis have been described in the statistical literature [14,15].

18.5 Information theory in data analysis

In Chapter 17 we studied PCA and described how it can be applied to display multivariate data. One of the main focuses of interest in the visual evaluation is then the occurrence of inhomogeneities e.g. clusters or outliers. PCA can detect these inhomogeneities, although it is not specifically directed towards it. Indeed, its aim is to describe variance. Inhomogeneities are sources of variance. They are, however, not necessarily the main source of variance in a data set. Statisticians [16–18] have developed a method that is specifically directed towards detecting inhomogeneities. The method is called *projection pursuit* and is similar to PCA in the sense that it projects data from multivariate space on what are called "interesting" directions (while PCA projects them on directions explaining maximal variance). An interesting direction is one that leads to non-uniform distributions of the projections. Fig. 18.3 shows a bivariate data set with an interesting and a non-interesting direction. *Uniformity* can be measured with the (negative) Shannon equation. In Section 18.2 we have shown that the Shannon equation leads to maximum values for the most uniform distribution. The negative of the Shannon

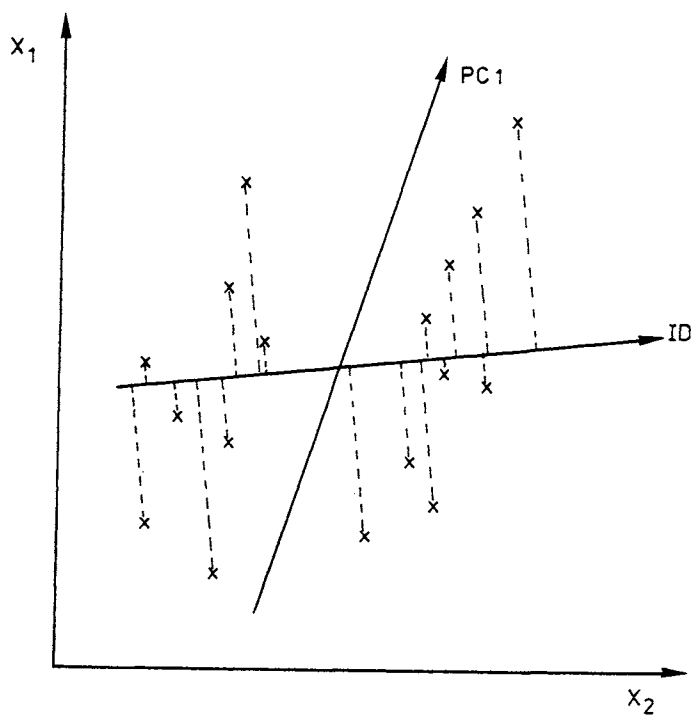


Fig. 18.3. An interesting direction (ID) for a bivariate data set, compared to the first principal component (PC1).

result thus leads to the maximum value when the distribution is "interesting", i.e. non-uniform. One can, as we did in Section 18.2, split up the axis indicating a certain direction in equidistant classes, note how many points project into each class and apply eq. (18.8) to obtain I and multiply by -1 to obtain the negative value. Another procedure is to create a potential function (see Chapter 33) around each projected point and to apply a continuous version of eq. (18.8). The reader is referred to references [14–16] for more details. Incidentally, it can be noted that the interesting direction of Fig. 18.3 resembles the canonical variate of Fig. 17.27, obtained by linear discriminant analysis. There is a fundamental difference however. In linear discriminant analysis one knows that there are two classes and which points belong to each class (supervised procedure). Here, one discovers that there are two classes (unsupervised procedure).

Another application of the Shannon equation is found in signal processing, where the entropy (H) is used as a criterion for the reduction of noise from a spectrum. In an iterative way, several possible spectra are calculated from which the one with maximum entropy is selected, or the one which contains the maximum amount of information (see Chapter 40).

References

1. K. Eckschlager and K. Danzer, Information Theory in Analytical Chemistry. Wiley, New York, 1994.
2. D.E. Clegg and D.L. Massart, Information theory and its application to analytical chemistry. J. Chem. Educ., 70 (1993) 19–24.
3. E. Shannon and W. Weaver, The Mathematical Theory of Information. University of Illinois Press, Urbana, IL, 1949.
4. D.L. Massart, The use of information theory in thin layer chromatography. J. Chromatogr., 79 (1973) 157–163.
5. A. Eskes, P.F. Dupuis, A. Dijkstra, H. De Clercq and D.L. Massart, The application of information theory and numerical taxonomy to the selection of GLC stationary phases. Anal. Chem., 47 (1975) 2168–2174.
6. F. Dupuis and A. Dijkstra, Application of information theory to analytical chemistry. Identification by retrieval of gas chromatographic retention indices. Anal. Chem., 47 (1975) 379–383.
7. P.J. Slonecker, X. Li, T.H. Ridgway and J.G. Dorsey, Informational orthogonality of two-dimensional chromatographic separations? Anal. Chem., 68 (1996) 682–689.
8. J.R. Quinlan and J. Ross, in: R.S. Michalski et al. (Eds.), Machine Learning: An Artificial Intelligence Approach. Tioga, Palo Alto, 1983.
9. M.P. Derde, L. Buydens, C. Guns, D.L. Massart and P.K. Hopke, Comparison of rule building expert systems with pattern recognition for the classification of analytical data. Anal. Chem., 59 (1987) 1868–1871.
10. P.K. Hopke and Y. Mi, in: E.J. Karjalainen (Ed.), Scientific Computing and Automation (Europe). Elsevier, Amsterdam, 1990.
11. D.R. Scott, A. Levitsky and S.E. Stein, Large scale evaluation of a pattern recognition/expert system for mass spectral molecular weight estimation. Anal. Chim. Acta, 278 (1993) 137–147.

12. P.J. Tandler, J.A. Butcher, H. Tao and P. de B. Harrington, Analysis of plastic recycling products by expert systems. Anal. Chim. Acta, 312 (1995) 231–244.
13. J.R. Quinlan, Induction of decision trees. Machine Learning, 1 (1986) 81–106.
14. L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, Classification and Regression Trees. Wadsworth and Brooks/Cole, Monterey, CA, 1984.
15. L.A. Clark and D. Pregibon, Tree-based Models, Ch. 9 in: J.M. Chambers and T.J. Hastie (Eds.), Statistical Models in S. Chapman and Hall, New York, 1992.
16. J.H. Friedman, Exploratory projection pursuit. J. Am. Statist. Assoc., 82 (1987) 249–266.
17. M.C. Jones and R. Sibson, What is projection pursuit. J. Roy. Statist. Soc., A 150 (1987) 1–36.
18. G.P. Nason, Entropy in multivariate analysis: Projection pursuit. UK Chemometrics Discussion Group Newsletter, Issue 18, November 1991.

Chapter 19

Fuzzy Methods

19.1 Conventional set theory and fuzzy set theory

Fuzzy methods are based on fuzzy set theory which was introduced by Zadeh [1] in 1965. It describes a way of dealing with vague statements or uncertain observations.

In a *conventional* or *crisp set* A the elements x of a given universe X either belong to the set A or they do not belong to A. We write $x \in A$ to indicate that x is a member or element of A and $x \notin A$ to indicate that x is not a member of A. Consider for example a universe X containing resolutions of chromatographic peaks, R_s, between 0 and 2. A crisp set labelled "well-resolved" would then for instance contain all resolutions at least equal to 1.5. They are a member of the set "well-resolved" while resolutions < 1.5 are no member of this set. A crisp set can be represented in different ways: (i) by listing the elements belonging to the set (only useful for finite sets containing a finite number of elements), (ii) by stating the conditions for membership; for our example this would be $A = \{x|X \geq 1.5\}$ where the symbol | means "such that", (iii) by using the characteristic function or *membership function* $m_A(x)$ that assigns a value 1 to all elements belonging to the set "well-resolved" and a value 0 to all elements that do not belong to this set:

$$m_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases} \tag{19.1}$$

In Fig. 19.1a the membership function is represented for the crisp set "well-resolved". In conventional set theory the transition between membership and non-membership is abrupt and for the variable well-resolved this might not be the most appropriate representation. Indeed an R_s value equal to 1.51 would indicate a good separation between the peaks while with an R_s value equal to 1.49 the conclusion would be that the chromatographic peaks are ill-separated. A more useful approach could be to consider R_s-values larger than 1.8 as really well-separated and those smaller than 0.8 as not well-separated at all. In between we would be rather vague about the amount of separation by specifying that the peaks are rather well-separated or rather ill-separated depending on whether R_s approaches 1.8 or 0.8 respectively. Fuzzy sets take this kind of vagueness into account by allowing membership values between 0 and 1. The fuzzy set A representing our concept of well-resolved assigns

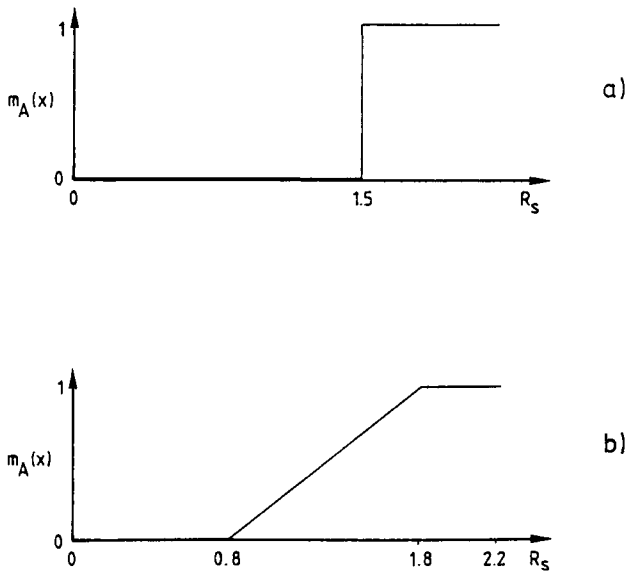


Fig. 19.1. Representation of the variable well-resolved (a) as a crisp set, (b) as a fuzzy set.

a membership value, $m_A(x)$, of 1 to R_s-values ≥ 1.8; $m_A(x)$ decreases when the retention becomes smaller, to reach a value of 0 at $R_s = 0.8$. The membership value thus indicates to which degree two chromatographic peaks are considered to be well-resolved. Therefore, as shown in Fig. 19.1b, the transition between membership and non-membership is not abrupt but gradual. This can be generalized in the following way:

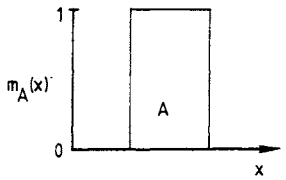
$$m_A(x) = \begin{cases} > 0 \text{ and } \leq 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{19.2}$$

where A represents a fuzzy set.

Fuzzy set theory not only allows us to deal with vague statements but it can also be applied with imprecise measurements. A simple analytical example from Bandemer and Otto [2] will illustrate this.

The presence of an element in a sample, analyzed by e.g. atomic spectrometry, has to be verified by means of a spectroscopic line. The position of the line is compared with a library of reference lines. Due to random variation the experimentally obtained line will not exactly match the tabulated line. Therefore an interval around the reference line is taken into account in order to decide whether both lines coincide. It will be concluded that they coincide if the observed line is within the interval and that they do not coincide if the observed line is outside the interval. Consequently only a yes–no or a 1–0 answer is obtained. This can be expressed by the membership function $m(x)$ which assigns a value 1 to each line x contained in

A. CONVENTIONAL SET:



B. FUZZY SETS:

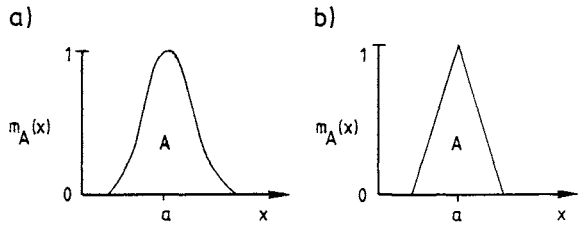


Fig. 19.2. Another example of (A) a conventional or crisp set with membership function specified in eq. (19.1). (B) fuzzy sets with (a) exponential membership function (eq. (19.4)) and (b) triangular membership function (eq. (19.5)).

the interval. The membership value is 0 for lines that are outside the interval. In this way a conventional or crisp set has been defined. Fig. 19.2a shows a generalized representation of such a conventional set A. For our example no difference is made between a line that is situated close to the border and inside the interval and a line that exactly coincides with the reference line. With fuzzy set theory a more detailed description of the coincidence of the two lines is possible by allowing membership values between 0 and 1. The membership value for a line is then 1 only if it exactly matches the reference line and it decreases the more the line deviates from the reference line.

The type of membership function to be used depends on the problem. The membership function that characterizes the fuzzy set "well-resolved" in the first example has the following form:

$$m_A(x) = \begin{cases} 0 & \text{for } x < a \\ \dfrac{x-a}{b-a} & \text{for } a \le x \le b \\ 1 & \text{for } x > b \end{cases} \tag{19.3}$$

where $a = 0.8$ and $b = 1.8$.

For the characterization of the imprecision of the position of a spectroscopic line in the second example, a symmetrical function, such as the following exponential membership function:

$$m_A(x) = \exp(-(x-a)^2/b^2) \tag{19.4}$$

or a simpler triangular function

$$m_A(x) = [1 - |x - a|/b]_+ \tag{19.5}$$

could be used. In these expressions a and b are constants and the + sign in expression 19.5 indicates that negative values are equated to zero. In the spectroscopic example a represents the wavelength or the position of the reference spectroscopic line. The constant b is a parameter of width of the membership function. The above defined membership functions are illustrated in Fig. 19.2b.

Different sources of information can be used to specify the membership function such as personal experience, specific knowledge about the actual problem, literature data and statistical information. One usually finds that the specific mathematical form of the membership function chosen has only a minor influence on the final conclusions [3].

19.2 Definitions and operations with fuzzy sets

There are two ways of representing a fuzzy set A, either by stating the membership function ($m_A(x)$), or as an ordered set of pairs:

$$A = \{(x, m_A(x)) \mid x \in X\}$$

where x represents the elements and $m_A(x)$ their membership value. Elements with a membership value equal to zero are generally not listed. Examples of the latter notation are found below.

Since fuzzy sets are represented by their membership function, operations with fuzzy sets are defined via these membership functions. In what follows some simple operations are compared to those applied to crisp sets. They will be illustrated by means of the following example. Consider the crisp and finite universal set of resolutions X:

$$X = \{0.4,0.6,0.8,1.0,1.2,1.4,1.6,1.8,2.0,2.2\}$$

The fuzzy set "well-resolved" will be described as:

$$A = \{(1.0,0.20),(1.2,0.40),(1.4,0.60),(1.6,0.80),(1.8,1),(2.0,1)(2.2,1)\}$$

where the first values between the brackets represent the elements (here resolutions) of the fuzzy set and the second values represent their membership value.

The fuzzy set "very well-resolved" could be described for example as:

$B = \{(1.8,0.33),(2.0,0.66),(2.2,1)\}$

The *union* of two crisp sets is the set containing the elements that belong to at least one of the sets. For the union, L, of two fuzzy sets A and B ($L = A \cup B$) the membership function is given by the maximum of the membership functions $m_A(x)$ and $m_B(x)$:

$$L = A \cup B\text{: } m_L(x) = \max[m_A(x), m_B(x)] \tag{19.6}$$

The union of the fuzzy sets "well-resolved" and "very well resolved" therefore is the fuzzy set defined as:

$L = \{(1.0,0.20),(1.2,0.40),(1.4,0.60)(1.6,0.80)(1.8,1)(2.0,1)(2.2,1)\}$

The *intersection* of two crisp sets is the set containing all the elements belonging to both sets simultaneously. For the intersection, I, of two fuzzy sets, A and B, the membership function is obtained from the minimum of both membership functions, $m_A(x)$ and $m_B(x)$:

$$I = A \cap B\text{: } m_I(x) = \min[m_A(x), m_B(x)] \tag{19.7}$$

The intersection of the fuzzy sets "well-resolved" and "very well-resolved" is therefore the fuzzy set I defined as:

$I = \{(1.8,0.33),(2.0,0.66),(2.2,1)\}$

The *complement* of a crisp set contains all the elements of the universe X which do not belong to the set. In analogy the complement $\overline{A}$ of a fuzzy set is defined as:

$$m_{\overline{A}}(x) = 1 - m_A(x) \tag{19.8}$$

The complement of the fuzzy set "well-resolved" produces the fuzzy set "not well-resolved" which is defined as:

$\overline{A} = \{(0.4,1),(0.6,1),(0.8,1)(1.0,0.80)(1.2,0.60),(1.4,0.40),(1.6,0.20)\}$

The number of elements that belong to a finite crisp set is called the *cardinality*. For a fuzzy set A the cardinality is defined as the sum of the membership values of all elements of X in A:

$$\text{card } A = \sum_{x \in X} m_A(x) \tag{19.9}$$

For infinite X the cardinality is obtained as:

$$\text{card } A = \int_X m(x)\, \mathrm{d}x$$

Thus the cardinality for the fuzzy set "well-resolved" is:

$$\text{card } A = 0 + 0 + 0 + 0.20 + 0.40 + 0.60 + 0.80 + 1 + 1 + 1 = 5$$

The comparison of the cardinalities of different fuzzy sets can be performed by considering the *relative cardinality*. This relative cardinality corresponds to a normalization of the cardinality of a fuzzy set to the interval [0,1]. It is the cardinality of set A divided by the cardinality of a standard set U, e.g. the universe X.

$$\text{rel}_U \text{ card } A = \text{card } A \,/\, \text{card } U \tag{19.10}$$

The relative cardinality represents the fraction of elements of U present in A, weighted by their degree of membership in A. To compare fuzzy sets by their relative cardinality of course the same standard set U has to be chosen.

The relative cardinality for the fuzzy set "well-resolved" is:

$$\text{rel}_X \text{ card } A = 5/10 = 0.5$$

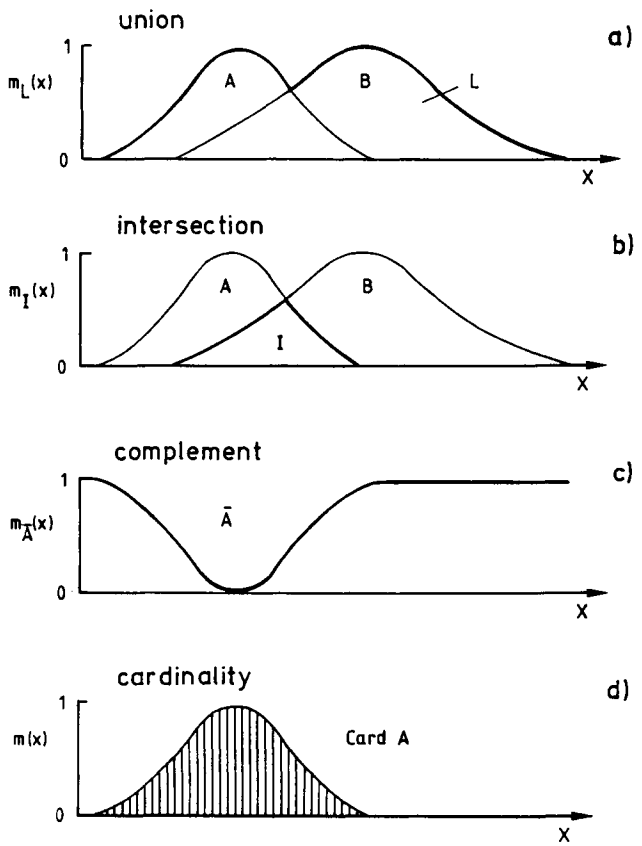


Fig. 19.3. Some operations on fuzzy sets. For explanation see text.

since the cardinality of the above described universal set X, taken here as standard set is equal to 10.

Finally the *support* of a fuzzy set A in the universal set X is the crisp set of all elements x that have a non-zero membership value in A:

$$\text{supp } A = (x \in X \mid m_A(x) > 0) \tag{19.11}$$

The support of the fuzzy set "well-resolved" therefore is

$$\text{supp } A = \{1,1.2,1.4,1.6,1.8,2.0,2.2\}$$

Some of these operations are also illustrated with another example in Fig. 19.3. In this figure A and B represent two fuzzy sets. The union and the intersection of the two sets is shown in Fig. 19.3a and Fig. 19.3b, respectively. The complement of fuzzy set A is shown in Fig. 19.3c and the cardinality in Fig. 19.3d.

19.3 Applications

19.3.1 Identification of patterns

Fuzzy theory can be applied for identification purposes to account for the possible uncertainty in the data patterns. The principle will be explained by a simple example concerning the identification of spectroscopic patterns based on peak positions. Fig. 19.4b gives the data pattern for an unknown sample with six peaks. This unknown sample pattern (U) is considered to be crisp with $m_U(x) = 1$ if a peak is present at wavelength x and $m_U(x) = 0$ otherwise. How well does that unknown pattern match the reference pattern from Fig. 19.4a? Peak positions are not always exactly reproducible due to e.g. measurement noise. Moreover, the reference spectrum might not have been recorded under the same conditions as the unknown spectrum. Therefore the reference pattern is fuzzified by assigning a membership function to the lines of the spectrum. Here the triangular membership function of eq. (19.5), in which b is taken equal to 2, is used:

$$m_i(x) = [1 - |x - a_i| / 2]_+ \tag{19.12}$$

with a_i the wavelength of the ith peak ($i = 1, ..., n$; n being the number of peaks in the reference pattern). The membership function of the whole reference pattern, $m_L(x)$, is obtained from the union (see Section 19.2) of the fuzzified lines:

$$m_L(x) = \max_i m_i(x)$$

For each wavelength x this corresponds to the maximum of the overlapping membership functions. For example

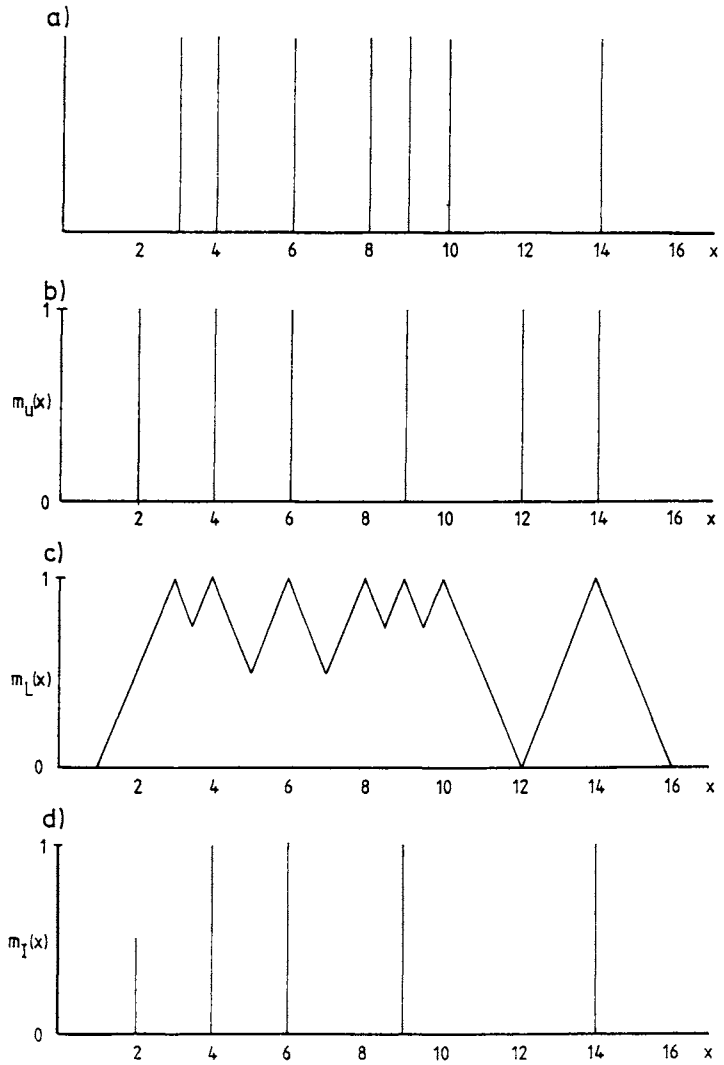


Fig. 19.4. (a) Reference pattern, (b) unknown data pattern, (c) the fuzzified reference pattern, (d) the intersection of the fuzzified reference pattern and the (crisp) unknown pattern.

for $a_1 = 3$ $m_1(x) = 0$ at $x = 1$

$m_1(x) = 0.5$ at $x = 2$

$m_1(x) = 1$ at $x = 3$

$m_1(x) = 0.5$ at $x = 4$

$m_1(x) = 0$ at $x = 5$

for $a_2 = 4$ $m_2(x) = 0$ at $x = 2$
$m_2(x) = 0.5$ at $x = 3$
$m_2(x) = 1$ at $x = 4$
$m_2(x) = 0.5$ at $x = 5$
$m_2(x) = 0$ at $x = 6$

for $a_3 = 6$ $m_3(x) = 0$ at $x = 4$
$m_3(x) = 0.5$ at $x = 5$
$m_3(x) = 1$ at $x = 6$
$m_3(x) = 0.5$ at $x = 7$
$m_3(x) = 0$ at $x = 8$

Therefore $m_L(x)$ is equal to 0; 0.5; 1; 1; 0.5; 1 for $x = 1, 2, 3, 4, 5, 6$ respectively. Fig. 19.4c represents the complete fuzzified reference pattern.

The comparison between the fuzzified reference pattern and the crisp unknown pattern is performed by intersecting (see Section 19.2) both data patterns, yielding the membership function for the intersection (I):

$m_I(x) = \min[m_L(x), m_U(x)]$

This gives a vector of m_I values that represent how well a peak in the unknown sample matches a peak in the reference sample.

All information necessary to obtain $m_I(x)$ is given in Table 19.1 and the intersection is represented in Fig. 19.4d. The $m_I(x)$ values can be aggregated to a single value that characterizes the goodness of fit between the unknown sample and the reference pattern by calculating the relative cardinality (see Section 19.2). Here this is done by referring the cardinality of the intersection to the cardinality of the unknown crisp set:

$$\text{rel}_U \text{ card } I = \text{card } I / \text{card } U$$
$$= \sum m_I(x) / \sum m_U(x)$$
$$= 4.5/6$$
$$= 0.75$$

This value represents the degree of containment and reflects the quality of coincidence between the fuzzified reference and the crisp unknown sample pattern. The larger the relative cardinality the better both patterns match.

The aggregation of the $m_I(x)$ values could also be performed by calculating the mean $m_I(x)$ value which here is equal to $4.5/16 = 0.28$.

TABLE 19.1

Summary of fuzzy set calculations for the example of Section 19.3.1

Wavelength (x)	Reference spectrum*	$m_L(x)$	Unknown spectrum $m_U(x)$	Intersection $m_I(x)$
1		0	0	0
2		0.5	1	0.5
3	x	1	0	0
4	x	1	1	1
5		0.5	0	0
6	x	1	1	1
7		0.5	0	0
8	x	1	0	0
9	x	1	1	1
10	x	1	0	0
11		0.5	0	0
12		0	1	0
13		0.5	0	0
14	x	1	1	1
15		0.5	0	0
16		0	0	0

*x means that a peak is present at that wavelength.

It is also possible to fuzzify the unknown spectrum and to consider the reference spectrum as being crisp. This might be a useful approach for the comparison of a sample spectrum with reference spectra from a spectroscopic library. The best match is searched for by intersecting the fuzzified sample spectrum with the different crisp reference spectra from the library.

The principle outlined above can be extended to complex classification problems in which several criteria are considered. It has been applied to the classification of patients with nephritis disease, based on chromatograms of urine samples [3,4]. Since retention data as well as signal response data were used, a two-dimensional membership function of the following form was specified:

$$m_i(x,y) = (1 - [(x - x_i)^2 / u_i^2 + (y - y_i)^2 / v_i^2]) \tag{19.13}$$

in which y is the peak height, x the retention time and u_i and v_i are the parameters of broadness which can be based on the uncertainties in peak height and retention time measurements.

Other applications concern the characterization of gasolines based on their capillary gas chromatograms [5], library search in infrared spectroscopy [6,7], identification of highly imprecise peaks from HPLC separation of vitamins [8], quality control of analgesic tablets based on monitoring their ultraviolet spectra, characterized by strongly overlapping and non-additive signals [4].

19.3.2 Regression

In Chapter 12 the problem of outliers in least squares regression was discussed and robust methods which are much less affected by outlying observations were introduced. Here it will be shown how a fuzzy approach, which also does not necessitate assumptions about the residuals, except for the definition of the membership function can be used to advantage in regression.

As an example consider the following x, y values: (1, 1.1); (2, 2.0); (3, 3.1); (4, 3.8) and (5, 6.5). The data are considered as fuzzy observations, which means that the uncertainty is described by a suitable membership function. If only the y-variable is subject to error, as is generally assumed in calibration, a one dimensional symmetrical function of the form:

$$m(y) = (1 - |y - a| / b)_+ \tag{19.14}$$

could be specified. For the different observations of our example, it is represented in Fig. 19.5. The parameter of width, b, could be based on the standard deviation of the measurements, s. Suppose s is equal to 0.1 and b is equal to $2s = 0.2$. Consequently the support of the fuzzy observations is a line segment with length equal to 0.4. The meaning of this is the following: a point in the middle of the support is an absolute member of the observation and has a membership value equal to 1. A point at the extremes of the line or outside the line does not belong to the observation and has a membership value equal to zero. In-between the extremes, membership values between 0 and 1 are obtained.

To fit a straight line through these observations we look for the line which has the highest membership with the fuzzy observations. In Fig. 19.6 two regression lines are represented. Line 1 is the least squares line which does not give a good approximation to the fuzzy observations since it only intersects the membership

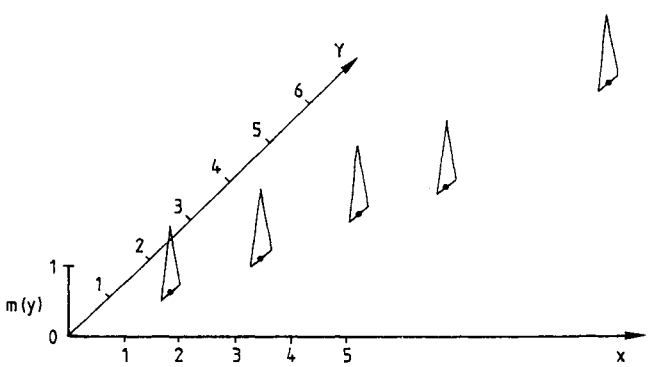


Fig. 19.5. Illustration of the triangular membership function (eq. 19.14) with a line as support for the fuzzy observations in regression.

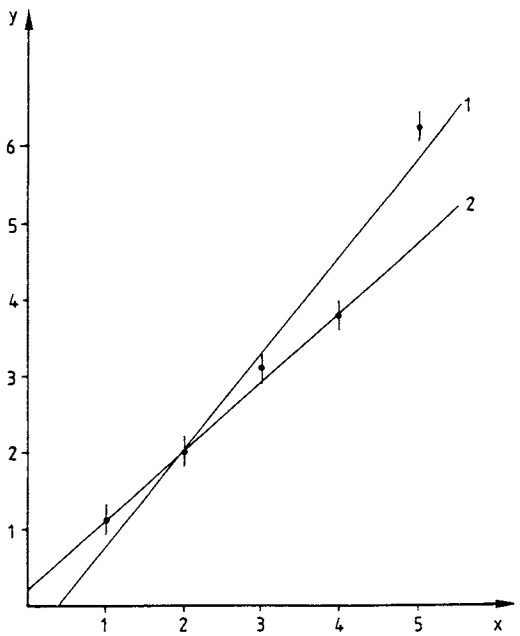


Fig. 19.6. Line 1 represents the least squares line ($\hat{y} = -0.48 + 1.26\,x$) while line 2 is the optimal line obtained with fuzzy regression ($\hat{y} = 0.19 + 0.91\,x$).

function of a single point. Line 2 obviously has a higher degree of approximation since it intersects the membership function of four of the data points. In fact the latter is the optimal line since it has the highest membership with the observations. The procedure to obtain this line can be summarized as follows:

1. Obtain the least squares line:

$$\hat{y} = -0.48 + 1.26\,x$$

2. At the different x_i values calculate for the corresponding fuzzy observation the membership values of the predicted y values, $\hat{y}_i$. This characterizes at each x_i the intersection of the actual straight line ($\hat{y} = b_0 + b_1 x$) with the membership function of the fuzzy observation:

$$m_i(b_0,b_1) = (1 - |\hat{y}_i - y_i| / 0.2)_+ \qquad (19.15)$$

For our example the $m_i(-0.48,1.26)$ thus obtained are given in Table 19.2.

3. Obtain a single value that characterizes the approximation of the whole line to the fuzzy observations which here will be denoted as $m(b_0,b_1)$. As observed in Section 19.3.1 this can be obtained as the mean of the $m_i(b_0,b_1)$ values:

$$m(-0.48,1.26) = \sum m_i(-0.48,1.26)/5$$
$$= 0.80/5 = 0.16$$

TABLE 19.2

Some calculations for the fuzzy regression of the data from Fig. 19.5

1. $\hat{y} = -0.48 + 1.26\,x$

x_i	y_i	$\hat{y}_i$	$m_i(-0.48,1.26)$
1	1.1	0.78	0
2	2.0	2.04	0.80
3	3.1	3.30	0
4	3.8	4.31	0
5	6.5	5.82	0

$m(-0.48,1.26) = 0.80/5 = 0.16$

2. $\hat{y} = 0.00 + 1.00\,x$

x_i	y_i	$\hat{y}_i$	$m_i(0.00,1.00)$
1	1.1	1	0.50
2	2.0	2	1.00
3	3.1	3	0.50
4	3.8	4	0
5	6.5	5	0

$m(0.00,1.00) = 2.00/5 = 0.40$

3. $\hat{y} = 0.19 + 0.91\,x$

x_i	y_i	$\hat{y}_i$	$m_i(0.19,0.91)$
1	1.1	1.10	1.00
2	2.0	2.01	0.95
3	3.1	2.92	0.10
4	3.8	3.83	0.85
5	6.5	4.74	0

$m(0.19,0.91) = 2.90/5 = 0.58$

4. Optimize $m(b_0,b_1)$ by varying b_0 and b_1 for example by means of a grid search in which different combinations of b_0 and b_1 values are considered [3] or by means of a Simplex optimization procedure [9]. The principle of the latter is described in Chapter 26.

5. The optimum regression parameters are those for which the highest value $m(b_0,b_1)$ is obtained.

Table 19.2 indicates that the line $\hat{y} = 0.00 + 1.00\,x$ better approximates the fuzzy observations than the least squares line since m (0.00, 1.00) = 2.00/5 = 0.40. The optimum is obtained for the line $\hat{y} = 0.19 + 0.91\,x$ which yields $m(0.19, 0.91) = 0.58$ (see Table 19.2). This line is very similar to the least squares line through the first four points. Notice that the line does not intersect the membership function for the fifth observation since $m_5(0.19, 0.91) = 0$ which means that this point is an outlier to the line. This forms the basis of a test to detect outliers in calibration proposed

by Hu et al. [9]. Since the selection of the size of the support is critical in the detection of outliers, the precision of the measurements has to be known. To account for the decreasing precision of the measurements with the concentration in heteroscedastic calibration data, the width of the membership function specified for the different data points, i.e. the size of the line support, can be varied accordingly.

By specifying a two-dimensional membership function of the form given in eq. (19.13) uncertainties in both the x and the y variable can be taken into account. A circle or an ellipse can be used as support for the fuzzy observations [9,10], describing a similar or a different error in both variables, respectively. This fuzzy approach could for example be applied in the comparison of two methods although it is not evident how to compare the slope and intercept with 1 and 0, respectively.

19.3.3 Other applications

Besides the above mentioned applications, fuzzy methods can also be applied to clustering data (see Chapter 30). A potential application of fuzzy theory is the use of fuzzy rules [11,12] in artificial intelligence (see Chapter 43). The incorporation of fuzzy rules in neural networks (see Chapter 44) has also been described [13].

References

1. L.A. Zadeh, Fuzzy sets. Inform. Control, 8 (1965) 338–353.
2. H. Bandemer and M. Otto, Fuzzy theory in analytical chemistry. Mikrochim. Acta (Wien), II (1986) 93–124.
3. M. Otto, Fuzzy theory explained. Chemom. Intell. Lab. Syst., 4 (1988) 101–120.
4. M. Otto and H. Bandemer, Pattern recognition based on fuzzy observations for spectroscopic quality control and chromatographic fingerprinting. Anal. Chim. Acta, 184 (1986) 21–31.
5. E. Stottmeister, H. Hermann, P. Hendel, D. Feeler, M. Nagel and H.-J. Dobberkau, Spurenbestimmung von Vergaser- und Dieselkraftstoffen in Wasser mittels Kapillarchromatographie/automatischer Mustererkennung. Fres. Z. Anal. Chemie, 327 (1987) 709–714.
6. T. Blaffert, Computer-assisted multicomponent spectral analysis with fuzzy data sets. Anal. Chim. Acta, 161 (1984) 135–148.
7. A.R. Goss and M.J. Adams, Spectral retrieval by fuzzy matching. Anal. Proc. incl. Anal. Commun., 31 (1994) 23–25.
8. M. Otto, W. Wegscheider and E.P. Lankmayr, A fuzzy approach to peak tracking in chromatographic separations. Anal. Chem., 60 (1988) 517–521.
9. Y. Hu, J. Smeyers-Verbeke and D.L. Massart, An algorithm for fuzzy linear calibration. Chemom. Intell. Lab. Syst., 8 (1990) 143–155.
10. M. Otto and H. Bandemer, Calibration with imprecise signals and concentrations based on fuzzy theory. Chemom. Intell. Lab. Syst., 1 (1986) 71–78.
11. M. Otto, Fuzzy theory, A promising tool for computerized chemistry. Anal. Chim. Acta, 235 (1990) 169–175.
12. M. Otto, Fuzzy sets, Applications to analytical chemistry. Anal. Chem., 62 (1990) 797A–802A.
13. B. Walczak, E. Bauer-Wolf and W. Wegscheider, A neuro-fuzzy system for X-ray spectra interpretation. Microchim. Acta, 113 (1994) 153–169.

Chapter 20

Process Modelling and Sampling

20.1 Introduction

Knowledge about the composition of a sample yields information on an object or a process, such as soil, a river, a chemical reactor. With this information a decision is made about, for instance, the necessity of cleaning the soil or performing a corrective action on the chemical process. Because the amount or concentration of the measured constituent may vary in time or with position, it is clear that making decisions on the composition of the sampled object based on a single sample is not possible. Instead, several samples should be taken according to a well-defined sampling scheme. Processes or objects may be sampled for three reasons: (i) to *describe* the composition or *state* of an object or a process, i.e. the concentration of a particular constituent at a specific position in the object or at a certain time; (ii) to *monitor* a state, e.g. when the state of a system varies in time, one may want to know whether the state of the system is drifting away from the target value and risks to cross a given threshold; (iii) to *control* a state. By control, which is a consecutive series of actions, one aims to diminish the system (e.g. process) fluctuations in order to manufacture a product within certain specifications. Monitoring and control lead to an action immediately after the measurement is finished. The aspect of time is thus important. When describing or modelling the fluctuations of a process, one may first collect all results. Because the modelling step is carried out afterwards, the time aspect is of less importance. It is obvious that the sampling plan together with the accuracy and precision of the measurement device define how well the process fluctuations are known. Monitoring and control of time-varying processes are daily practices in manufacturing. Therefore, adequate sampling strategies in combination with appropriate chemical or physical measurements are very important. They allow efficient process management and the production within narrow product specifications.

In Chapter 2 we discussed the influence of the measurement error on the ability to observe process fluctuations. Moreover, control charts were introduced in Chapter 7 to check whether a process is statistically in control (no drift, time invariant standard deviation). In this chapter we will discuss the situation in which we want not only to observe the process fluctuations but also to reduce those

fluctuations by control. In this case the time aspect is important. Control is ideally performed on the basis of the current value of the process. However, because it takes some time before the measurement is available and the control action is executed, the process value may be different from the value at the time the sample was taken. When the process is sampled at regular intervals, the process value between two samples is not known and needs to be estimated. In addition the imprecision of the measurement adds some uncertainty about the process value. As a result not all process fluctuations are removed by control causing some residual variation. The magnitude of these remaining variations depends on our knowledge of the true value (or state) of the process at the time that the control action is carried out. This knowledge of the process state is usually represented by a model, by which unobserved process states are estimated.

In this chapter we discuss a number of issues which influence our ability to estimate or to predict a system or process state. Such estimations are based on a number of observations (in time or space). Depending whether the process fluctuations are described or controlled with the model, unobserved process values are estimated either by (i) interpolation between two consecutive sampling points or by (ii) extrapolation from the last to the next analytical result. In the first case the analysis time is not important. Measured values are attributed to the sampling points. On the contrary, in process control the dead time, which is the time between the sampling and control action, and the sampling interval determine how far ahead a process value has to be forecast. With these estimates a plant manager or operator may intervene in the process.

20.2 Measurability and controllability

Uncertainty about process states is related to the difference between the true (but unknown) state and the estimated state (Fig. 20.1). If one controls a process according to the estimated state, the shaded area in Fig. 20.1 represents the prediction error, which are the fluctuations left after control. These residual fluctuations are the result of our imperfect or uncertain knowledge of the process fluctuations. The uncertainty before any observation has been done is expressed by the variance of the process fluctuations, s_0^2 and the uncertainty after a series of observations is the residual error s_e^2. If process states are perfectly forecast, $s_e^2 = 0$. If no information at all is obtained by the observations, the best prediction is the process mean and, therefore, $s_e^2 = s_0^2$. Van der Grinten [1–4] and Leemans [5] introduced the concept *measurability* (m), to express the ability of a measuring system to follow the process fluctuations:

$$m = \sqrt{(s_0^2 - s_e^2)/s_0^2} = \sqrt{(1 - s_e^2/s_0^2)} \qquad (20.1)$$

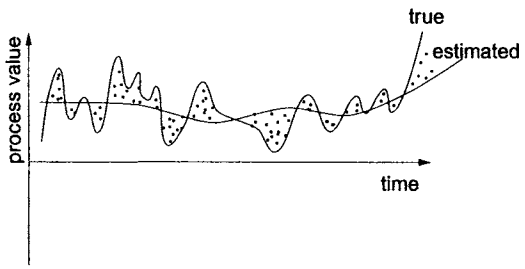


Fig. 20.1. True and estimated process values. The shaded area represents the unmodelled process fluctuations or prediction residuals.

A measurability equal to 1 means that the process value is perfectly known at any time. The measurability is zero when process values are not known at all. Measurability is thus a measure for the amount of information which is obtained by sampling a process and analyzing the samples. Perfect knowledge should allow a perfect monitoring or control. However, due to imperfections in the control system, not all process fluctuations will be removed and therefore *controllability* (r) is usually a fraction of the measurability:

$$r = fm \tag{20.2}$$

where $0 \leq f \leq 1$

When designing a control or measuring system one should choose the most cost-effective measuring system. A high measurability may require a costly measuring/control system, which should be paid back by higher revenues from a better product or fewer situations which are outside specification. This will be illustrated with a fictitious industrial process for the production of a nitrogen fertilizer with the following specifications: the long term average N-content over all batches (kegs or bags) should be at least 23.0%, and the amount in an individual keg should not be below 22.3%. Let us suppose that the standard deviation (s_0) of the naturally occurring fluctuations of the N content between the kegs is 1.2%. To meet the specifications in 99% of all products sold, an overdose (to 23.0%) of ($22.3 + 2.33 \times 1.2 - 23.0 =$) 2.1% N is required ($z_{0,01} = 2.33$). If by control the standard deviation (s_e) of the statistical fluctuations can be reduced to 0.7% N, this overdose can be decreased to ($22.3 + 2.33 \times 0.7 - 23.0$) = 0.9% N. The marginal return is the value of the 2.1 – 0.9 = 1.2% N which should be balanced against the marginal cost of putting a control system in place. From eq. (20.1) the required measurability of the measuring system can be calculated, which is:

$$m = \sqrt{1 - (0.7/1.2)^2} = 0.812$$

In order to meet the specifications without overdosing, the fluctuations should be further reduced to a standard deviation equal to (23.0 – 22.3)/2.33 = 0.30%, and the measurability of the measuring system should be:

$$m = \sqrt{1 - (0.3/1.2)^2} = 0.968$$

The next decision is to select a suitable analytical method with an appropriate sampling scheme. That this is not a trivial problem is reflected by the fact that at least seven candidate methods are available (see Table 20.1) for measuring the N-content in the mixing tank, each with a specific precision and analysis time. A possible approach is to select the method with the best balance between analysis time (T_d) and precision (s_a) by the Pareto optimization procedure of Fig. 20.2. The Pareto method is a multicriteria optimization method and is further explained in Chapter 26. In this particular instance methods 6 and 7 are the Pareto optimal choices. The approach described here does not take into account the characteristics of the object or process, nor does it provide an optimal sampling frequency based on cost considerations. Therefore, a more complex approach is appropriate, based on a process model by which system states can be predicted. This model should

TABLE 20.1

Methods for the analysis of nitrogen (adapted from Ref. [5])

Method	Analysis time (min) (T_d)	Precision (%N) (s_a)
1. Total N by distillation	75	0.17
2. Total N by automated distillation	12	0.25
3. NO_3 by autoanalyzer	15.5	0.51
4. NO_3 by ion selective electrode	10	0.76
5. $NH_4NO_3/CaCO_3$ by X-ray	8	0.80
6. Total N by neutron activation	5	0.17
7. γ-Ray absorption	1	0.64

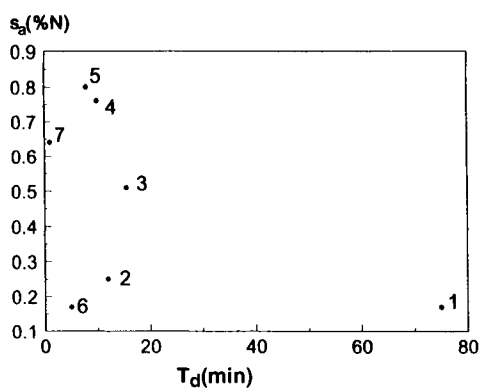


Fig. 20.2. Plot of the precision (s_a) vs. analysis time (T_d) of analytical methods for the analysis of N in a fertilizer (see Table 20.1). Methods 7 and 6 are Pareto optimal.

provide a relationship between the measurability and the characteristics of the measuring system: i.e. precision, sampling plan and measurement time (analysis time). Intuitively, we know that this will depend on some properties of the process fluctuations and also on the adequacy of the algorithm to estimate/predict process states.

20.3 Estimators of system states

Because physical or chemical measurements require time, the analytical result obtained for a sample does not necessarily represent the current state of the system or process at the moment that the result is obtained. Between sample taking and receipt of the analytical result, the system state may have changed (Fig. 20.3). The probability that the state is changed depends on the velocity of the process fluctuations. Therefore it is necessary to make assumptions on the evolution of the system state during that period. The simplest assumption is that the system state does not change between two analytical results. This situation is shown in Fig. 20.4. The points S and R on the time axis in this figure represent the sampling

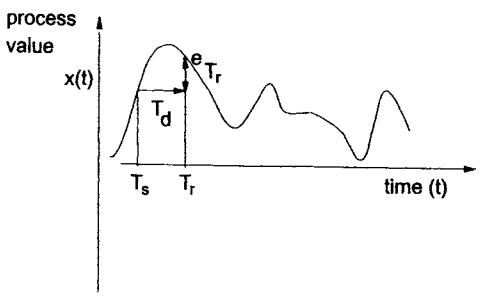


Fig. 20.3. Prediction error caused by the measurement (analysis) time (T_d). T_s is the sampling time, T_r is the time at which the result is available, e_{Tr} is the prediction error.

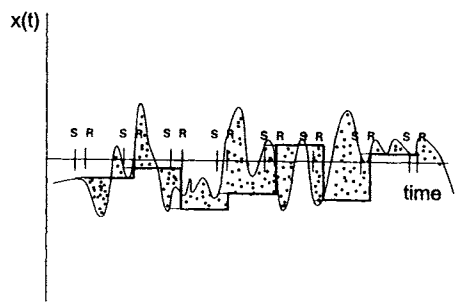


Fig. 20.4. True and estimated process states for a control system, where the estimated process value is the last analytical result (until the next one is obtained). The shaded area represents the remaining fluctuations after control.

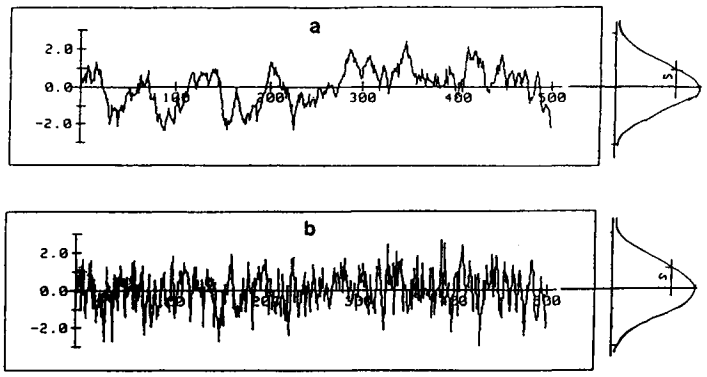


Fig. 20.5. A slow (a) and fast (b) process with equal probability distribution.

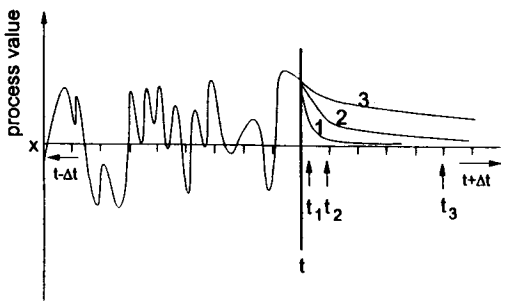


Fig. 20.6. The forecast of process states from the last observation at time t. Lines 1,2 and 3 are negative exponential functions.

points, respectively the reporting times. The horizontal lines represent the estimated states and the shaded area is the error of estimation, which represents the unobserved and therefore uncompensated fluctuations. The estimation can be improved by taking into account that the process fluctuations follow a certain pattern. They may be slow or fast (Fig. 20.5). Extreme values are less probable than the average process state. In any case and certainly after a long period without observation, the most probable process state is the mean process value and not the last measured value. One can imagine that the process forecast may evolve as shown in Fig. 20.6. Just after the last measurement has been obtained it is unlikely that the system has changed (point t_1 in Fig. 20.6). The longer we have to wait, however, the more likely it is that the process value is changed. As the mean is the most probable process value, this is the best prediction after a long time (t_3 in Fig. 20.6). Between t_1 and t_3 (t_2 in Fig. 20.6) one may assume process values between the last measured value and the mean process value. For fast fluctuating processes

the mean process value is the most probable already after a short time (line 1 in Fig. 20.6), whereas it takes some more time for slow processes (line 3 in Fig. 20.6). The exact curve can be derived by modelling these process fluctuations, as we explain in the next section.

20.4 Models for process fluctuations

Process fluctuations are the result of the dynamics of a system (in time or in space) and can be modelled by several approaches: directly, by applying a step or impulse response, or indirectly by analyzing the properties of the observed fluctuations. The application of a step or impulse is usually undesirable because it requires that the process is disturbed. Therefore, the indirect approach is more appropriate. We can choose between an analysis in the time domain and in the frequency domain (see Chapter 40). Although both types of analysis essentially provide the same information, the analysis in the time domain is preferred as it directly provides a model which allows to interpolate or extrapolate between two sampling points. Therefore, we need to introduce some new concepts, namely time series, autocorrelation and autoregression.

20.4.1 Time series

Any discrete process or sampled continuous process can be represented by a time series. For instance, the weekly averages of the Dow Jones index of the American Stock Exchange is a time series of a discrete process. The daily measurement of the temperature in a river is a time series of a sampled continuous process. Implicitly we have assumed that the process values are equidistant in time (or in space), which is a condition for a time series. A time series is said to be stationary when the statistical parameters, mean and standard deviation are time-invariant. Figure 20.5 illustrates that the dynamics of two time series with identical mean and standard deviation may be completely different. That difference will be quantified by the autocorrelation and autoregression function (Sections 20.4.2. and 20.4.3).

Let us first introduce the formalism which will be used in the next sections. Process values at discrete sampling points will be represented by $x(t)$, where the index t refers to t sampling intervals (Δt) since time $t(0)$. Thus $x(t)$ is the value of the series at a time $t(0) + t\Delta t$. In some instances we will need to indicate pairs of observations with the same distance τ, which is equal to an integer number of times the sampling interval. Thus $\tau = 2$ for the pair $\{x(2),x(4)\}$, which corresponds to a distance in time equal to $2\Delta t$.

20.4.2 Autoregressive models

In the same way one can fit a regression model through the measurements y as a function of x, or to use the formalism applied here, through pairs of observations $\{y(1),x(1)\}$, ... $\{y(n),x(n)\}$ one can also fit a model through the process values of a time series with equal time–distance, τ, between each other. For instance for $\tau = 1$, a regression model is fitted through the pairs $\{x(1),x(2)\}$, $\{x(2),x(3)\}$, $\{x(t-1), x(t)\}$, ..., $\{x(n-1),x(n)\}$ or in general between the pairs $\{x(1),x(1+\tau)\}$, $\{x(2),x(2+\tau)\}$, ..., $\{x(n-\tau),x(n)\}$.

Equation (8.65) derived in Chapter 8, expresses the regression line in terms of the correlation coefficient $r(x,y)$ and the standard deviation of the two variables x and y:

$$(y_i - \bar{y}) = r(x,y)\frac{s_y}{s_x}(x_i - \bar{x}) + e_i \tag{20.3}$$

Substituting the observation pair (x_i,y_i) by $\{x(t),x(t+1)\}$ and replacing s_y by s_x and $\bar{y}$ by $\bar{x}$ because there is only one mean and one standard deviation for stationary processes, we find:

$$[x(t+1) - \bar{x}] = r(1)\,[x(t) - \bar{x}] + e(t+1) \tag{20.4}$$

$r(1)$ is the correlation coefficient between all pairs $\{x(t+1),x(t)\}$ for $t = 1$ to $n-1$. Expressed in words, eq. (20.4) states that the deviation from the mean at time $(t+1)$ can be predicted from the observed deviation from the mean at a time t multiplied by the correlation coefficient $r(1)$ at $\tau = 1$. Equation (20.4) can be calculated for any τ. For example, for two sampling intervals we find

$$[x(t+2) - \bar{x}] = r(2)\,[x(t) - \bar{x}] + e(t+2)$$

and in general:

$$[x(t+\tau) - \bar{x}] = r(\tau)\,[x(t) - \bar{x}] + e(t+\tau) \tag{20.5}$$

Equation (20.5) represents a *first-order autoregressive model.* It is said to be first-order because the model contains only one parameter $r(\tau)$. The noise term $e(t+\tau)$ represents white noise (see Chapter 40) which is independent of all process values $x(t)$.

20.4.3 Autocorrelation function and time constant

The values $r(\tau)$ constitute an *autocorrelation function*, giving $r(\tau)$ as a function of τ. In the previous section we derived that the parameter of a first-order autoregressive model is the correlation coefficient $r(\tau)$. For any τ this correlation coefficient is calculated from the process values according to:

$$r(\tau) = \frac{1}{(n-\tau-1)}\sum_{t=1}^{n-\tau}\frac{(x(t) - \bar{x})\,(x(t+\tau) - \bar{x})}{s_x^2} \tag{20.6}$$

TABLE 20.2

Calculation of the autocorrelation function of x(t) at τ=1 and 2

t	$x(t)$	$x(t+1)$	$x(t)-\bar{x}$	$x(t+1)-\bar{x}$	$x(t+2)-\bar{x}$	$[(x(t)-\bar{x})][x(t+1)-\bar{x}]$	$[x(t)-\bar{x}][x(t+2)-\bar{x}]$
1	-4.5	3	-4.53	2.97	4.17	-13.45	-18.89
2	3	4.2	2.97	4.17	6.87	12.38	20.40
3	4.2	6.9	4.17	6.87	-6.83	28.65	-28.48
4	6.9	-6.8	6.87	-6.83	-11.03	-46.92	-75.78
5	-6.8	-11.0	-6.83	-11.03	8.37	75.33	-57.17
6	-11.0	8.4	-11.03	8.37	–	-92.32	–
7	8.4	–	8.37	–	–	–	–

$n = 7; \bar{x} = 0.03; s_x^2 = 55.38$ $\quad \Sigma = -36.33$ $\quad \Sigma = -159.98$

$$r(1) = \frac{1}{(7-1-1)} \cdot \frac{-36.33}{55.38} = -0.13$$

$$r(2) = \frac{1}{(7-2-1)} \cdot \frac{-159.92}{55.38} = -0.72$$

For $\tau = 1,2, ..., n-2$ this gives $r(1), r(2), ..., r(n-2)$. These values constitute the autocorrelation function. For $\tau = 0$, $r(\tau) = 1$. How to calculate an autocorrelation is illustrated with a hypothetical numerical example given in Table 20.2.

Autoregression and autocorrelation can be visualized by plotting all pairs $\{x(t+\tau), x(t)\}$ for various values of τ. The results obtained for the process shown in Fig. 20.5a are plotted in Fig. 20.7 for the τ-values 1,3,5, 7 and 10. In addition the best fitting regression line is plotted. As one can see, the correlation decreases with increasing distance between the pairs of observation. Moreover, the correlation values are statistically identical to the slopes of the regression lines as expected from the fact that the autocorrelation value is the slope parameter of the autoregression function (eq. (20.5)). A plot of the values of the autocorrelation at $\tau = 1, 3, 5, 7$ and 10 constitutes an *autocorrelogram* (Fig. 20.8).

Autocorrelation functions or autocorrelograms of the two processes given in Fig. 20.5 are shown in Fig. 20.9. The faster the process fluctuates, the quicker the autocorrelogram decays to zero. For a first-order process this decay can be expressed by a single parameter, the *time constant T*. T is small for fast fluctuating processes and large for slow processes. To derive an expression for T we need to model the autocorrelation function. Let us recall the autoregressive model which predicts the process value at a time $(t + 2)$ from the value measured at a time (t):

$$[x(t+2) - \bar{x}] = r(2)\,[x(t) - \bar{x}] + e(t+2)$$

with $e(t + 2)$ the residual between model and process value at a time $(t + 2)$.

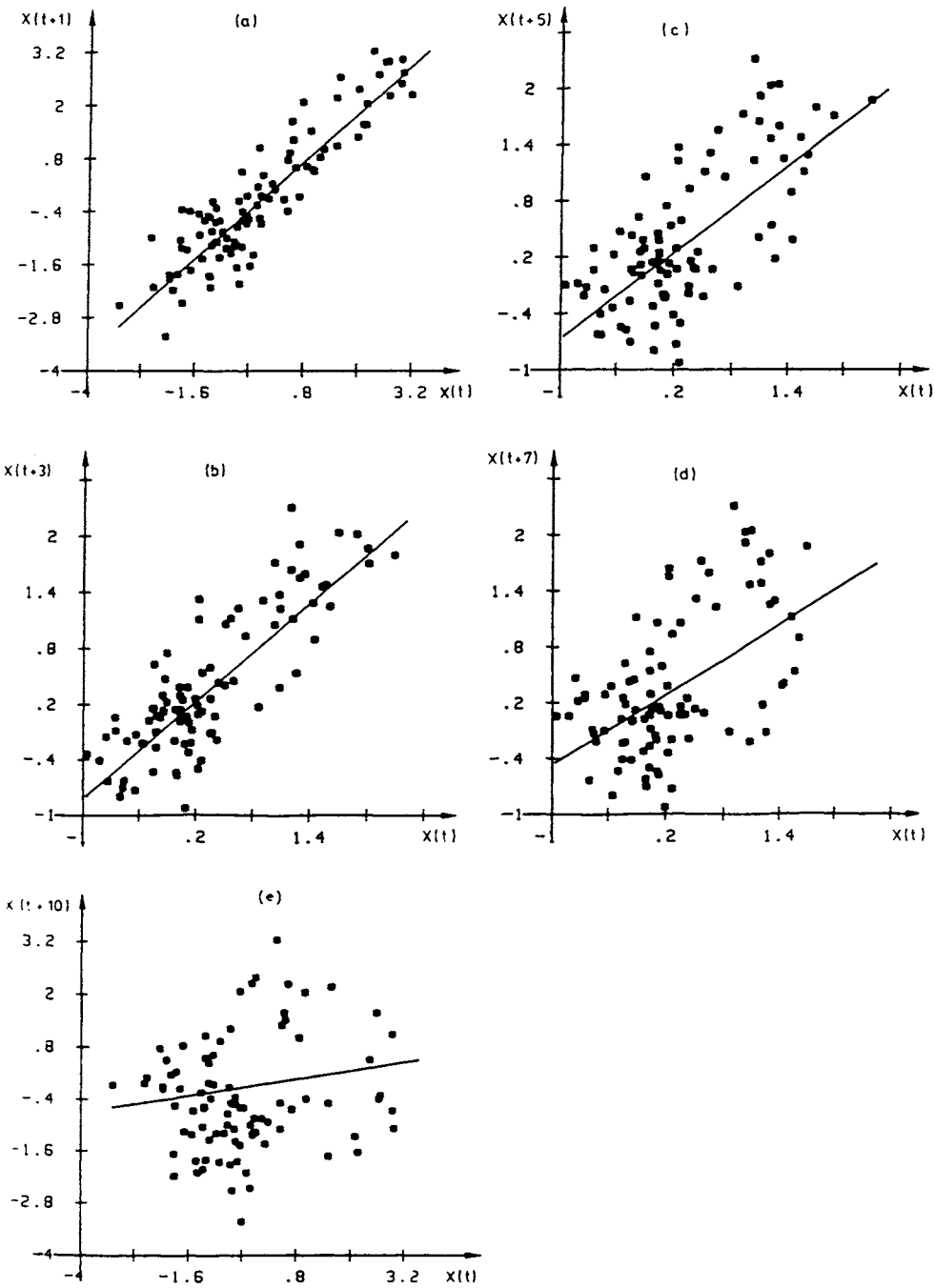


Fig. 20.7. Scatter plots of $x(t)$ vs $x(t + \tau)$ of the process in Fig. 20.5(a) for $\tau = 1, 3, 5, 7$ and 10. The line is the regression line $x(t + \tau) = b_0 + b_1 x(t)$. (a) $r(1) = 0.904$, $b_1 = 0.908$; (b) $r(3) = 0.812$, $b_1 = 0.831$; (c) $r(5) = 0.697$, $b_1 = 0.771$; (d) $r(7) = 0.591$, $b_1 = 0.611$; (e) $r(10) = 0.171$, $b_1 = 0.159$.

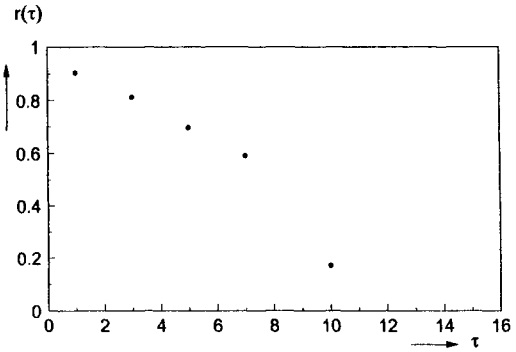


Fig. 20.8. Autocorrelation function ($r(\tau)$) of the process given in Fig. 20.5(a).

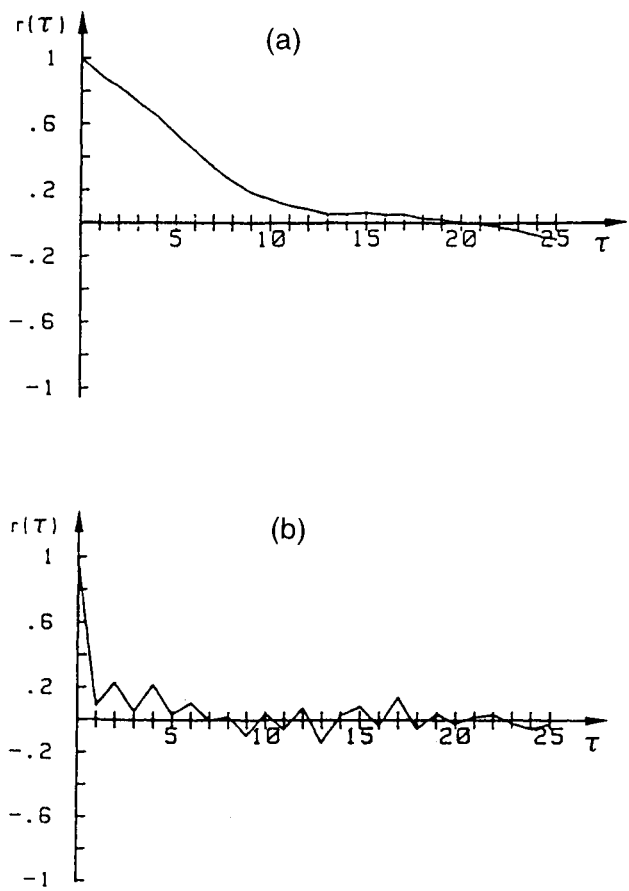


Fig. 20.9. Examples of autocorrelograms with a large (a) and small (b) time constant.

For a first-order process, $x(t + 2)$ can also be predicted by first predicting $x(t + 1)$ from $x(t)$, followed by the prediction of $x(t + 2)$ from $x(t + 1)$, which gives the following sequence of equations:

$$[x(t+2) - \bar{x}] = r(1)\,[x(t+1) - \bar{x}] + e(t+2)$$

$$[x(t+1) - \bar{x}] = r(1)\,[x(t) - \bar{x}] + e(t+1)$$

which gives

$$[x(t+2) - \bar{x}] = r(1)^2\,[x(t) - \bar{x}] + e(t+2)$$

We note that the residual $e(t + 2)$ in the above equations may have different values. Consequently, $r(2) = r(1)^2$ or in general $r(\tau) = r(1)^{\tau}$. This demonstrates that a first-order autoregressive function is fully defined by one parameter $r(1)$. The autocorrelation function of a first-order process is therefore an exponentially decaying function (because $r(1) < 1$), modelled by

$$\rho(\tau) = e^{-\frac{\tau}{T}} \tag{20.7}$$

containing only one parameter T, the time constant. In Fig. 20.10 an experimentally obtained autocorrelation function is fitted with an exponential function illustrating the validity of the exponential model. Because the autocorrelation $r(\tau)$ is calculated

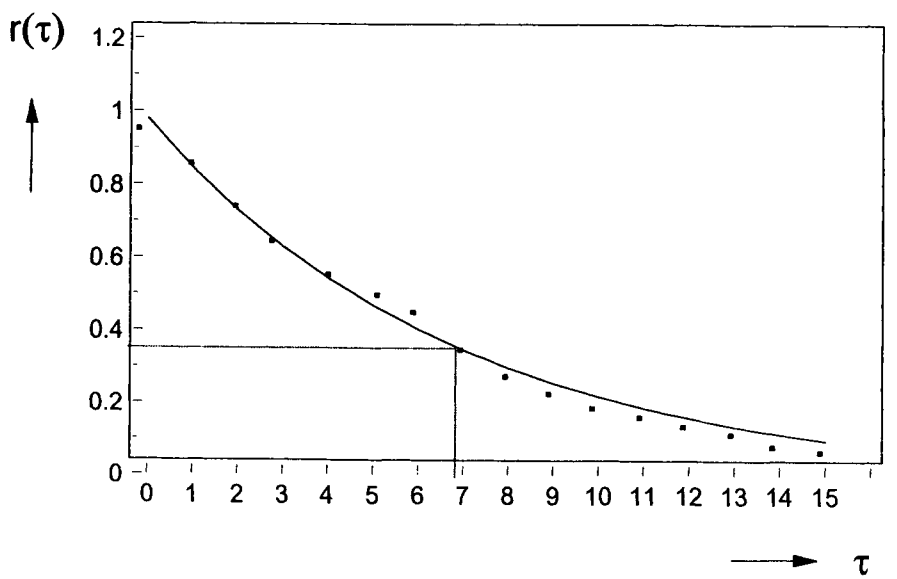


Fig. 20.10. Autocorrelogram of a first-order random process ($T = 7$), fitted with $\exp(-\tau/7)$. The dashed line gives the τ-value corresponding to $r = 0.37$, which is the time constant T.

from a finite number of observations equal to $(n - \tau)$, it is an estimate of the true value. A rule of thumb based on the exact equations derived by Bartlett [6] is that the standard deviation of the autocorrelation $r(\tau)$ is proportional to $k\sqrt{[T/(n-\tau)]}$ when T is not excessively small ($T > 5$ sampling intervals) where $k = 0$ for $\tau = 0$, $k \approx 0.3$ for $\tau = T/2$, $k = \approx 0.7$ for $\tau = T$ and $k = 1$ for $\tau > 1.5T$. Typically, 500 observations are needed to estimate $r(\tau)$ with a standard deviation equal to 0.1 at $\tau = T = 10$.

Before we discuss the time constant, two examples of an autocorrelation analysis are presented. In the first example, the fluctuations of the concentrations of NO_3^- and NH_4^+ in the river Rhine measured over a period of 468 weeks (Figs. 20.11a,b) were subjected to an autocorrelation analysis [7]. The autocorrelograms of the NO_3^- and NH_4^+ concentrations are given in Figs. 20.11c,d. In both autocorrelograms we see some deviations from the exponential shape which are explained

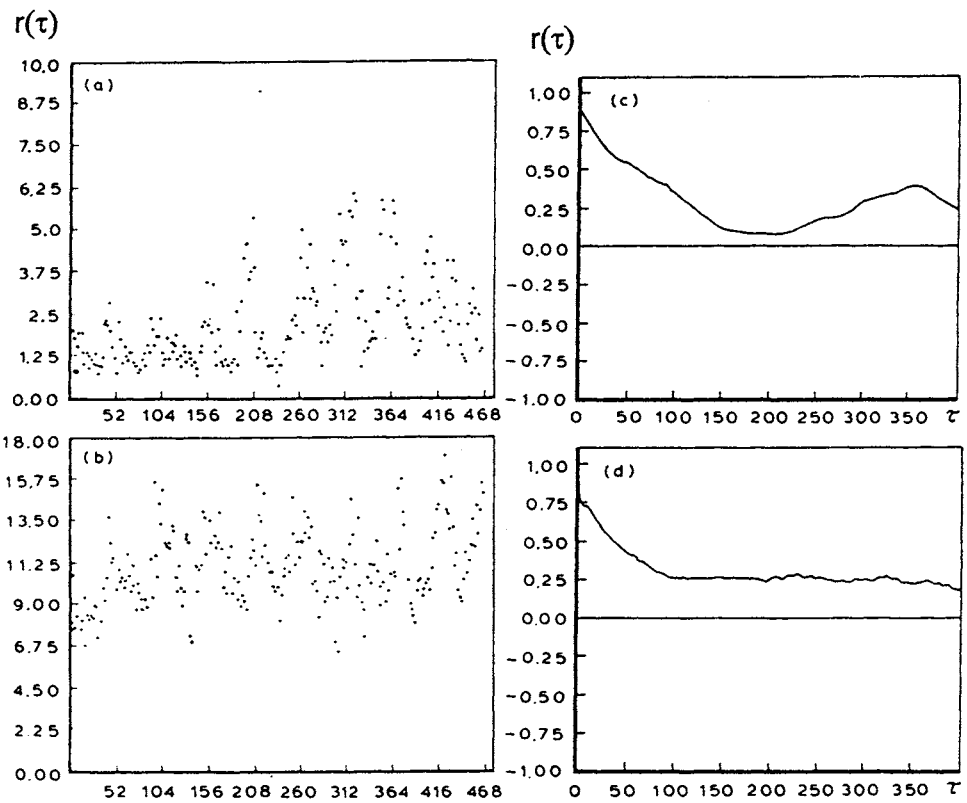


Fig. 20.11. Time series of the weekly NO_3^- (a) and NH_4^+ (b) concentrations in the river Rhine over a 468 weeks period and their respective autocorrelograms (c) and (d) (τ) is given in days. (Reprinted from Ref. [6]).

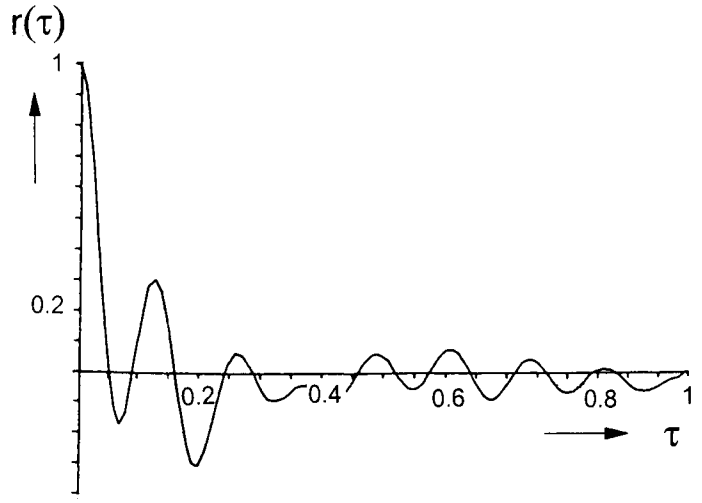


Fig. 20.12. Autocorrelogram of flame-ionization detector noise. (Reprinted from Ref. [7]).

below. The second example concerns a noisy base line obtained with a flame-ionization detector for GC [8]. The autocorrelogram is shown in Fig. 20.12. The detector contains a periodicity. From this example we intuitively see that autocorrelation functions and the Fourier transform are related as both provide information on the frequencies present in the signal (see e.g. the textbook by Priestly [9]).

The time constant can be calculated from the autocorrelation function in three ways:

(1) from $r(1) = e^{-1/T}$ it follows that $T = -1/\ln r(1)$

(2) for $\tau = T$, $r(T) = e^{-T/T} = e^{-1} = 0.347$. Thus the τ-value which corresponds to an autocorrelation value equal to 0.347 is equal to T (see Fig. 20.8).

(3) from a fit of eq. (20.7) through the autocorrelogram by non-linear regression

Although in principle the time constant of a process can be calculated from $r(1)$, it is advisable to calculate the autocorrelation over an extended τ-range (about 5 times the guessed time constant) in order to observe possible deviations from the exponential model. Such deviations may be informative as they indicate deviations from a first-order behaviour, e.g. a drift of the mean value, a periodicity, or several time constants:

(1) Drift of the mean: Drift of the mean introduces a positive (or negative) correlation with respect to the overall mean of the process (see Fig. 20.11d). This is reflected in the autocorrelogram by the fact that the values do not asymptotically approach the zero correlation line, but some other value. Because for large τ-values noise is uncorrelated, small drifts are more easily detected from the autocorrelogram than from the original data.

(2) Periodicity: Periodicity of process values introduces a periodic autocorrelogram. For the same reason as explained for drift, periodicities are more easily detected in the autocorrelogram than in the original process values (see Fig. 20.11c).

(3) Several time constants: The slopes of the autocorrelograms of Fig. 20.11 show a stepwise change at $\tau = 1$. The initially steep drop of the value of the autocorrelation from 1 at $\tau = 0$ to 0.8 at $\tau = 1$ is followed by a less steep exponential decay. This is a typical shape for processes with two time constants (second-order process). The first part of the autocorrelogram describes a fast process with a small time constant, whereas the second part of the autocorrelogram describes the slower part of the process. This situation may occur if the noise (or other source of variation) of the measuring device contributes substantially to the measured process variations. Because noise is a fast process compared to the signal two time constants are found. In some instances it is possible to derive the relative contribution of measurement noise (s_a^2) to the observed process fluctuations. For instance when the stepwise change in the autocorrelogram occurs at $r(\tau) = 0.8$, it means that the variance of the process (s_0^2) is 80% of the total observed variance which is equal to ($s_0^2 + s_a^2$), or $s_a^2 = 0.25 s_0^2$.

20.4.4 The autoregressive moving average model (ARMAX)

In Section 20.4.3 we introduced an autoregressive (AR) model for the prediction of the process variable, $x(t)$ (process output) at a time t from its past values ($x(t-1)$, $x(t-2)$, ...):

$$\hat{x}(t) = b(1)\, x(t-1) + b(2)\, x(t-2) + \ldots$$

We have also shown that for a first-order autoregressive model $b(2) = b(1)^2$. When $b(2)$ is independent of $b(1)$ the model is second-order, which provides satisfactory representations for a wide variety of applications in business and industry. In the AR model only past values of the output are employed to predict its present value $x(t)$.

If we assume that the process variable, $x(t)$, is regulated by a certain control variable $u(t)$ (see Fig. 20.13), the process output, $\hat{x}(t)$ can be predicted taking into account all settings of that control variable in the past at $(t-1)$, $(t-2)$,... giving

$$\hat{x}(t) = b(1)x(t-1) + b(2)x(t-2) + \ldots + d(1)u(t-1) + d(2)u(t-2) + \ldots$$

This is called a *controlled autoregressive model* ARX, where AR refers to the autoregressive part and X to the extra deterministic input, called the exogenous variable.

Each prediction is subjected to a prediction error $e(t)$. The moving average (see Chapter 7) of these prediction errors ($c(1)e(t-1) + c(2)e(t-2) + \ldots$), which are assumed to have a zero mean and to be uncorrelated can be included in the model to improve the process output predictions as follows:

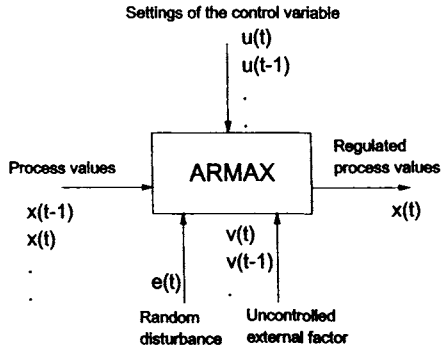


Fig. 20.13. The ARMAX process control model. The regulated process value $x(t)$ at a time t depends on: past values of the regulated process variable, past and current settings of the control variable, past and current values of any uncontrolled variable, and random noise.

$$\hat{x}(t) = b(1)x(t-1) + b(2)x(t-2) + \ldots + d(1)u(t-1) + d(2)u(t-2) + \ldots + c(1)e(t-1) + c(2)e(t-2) + \ldots.$$

A controlled autoregressive model combined with a moving average term is called a *controlled autoregressive moving average model* (ARMAX). Clearly this class of models includes the AR model (if all d() and $c()$ are zero), the moving average model, MA (if all $b()$ and $d()$ are zero), the ARX model (if all $c()$ are zero) and the autoregressive moving average model, ARMA (if all $d()$ are zero) as special cases. Assuming a first-order behaviour of the process (see Section 20.4.2), the ARMAX model reduces to:

$$\hat{x}(t) = b(1)x(t-1) + d(1)u(t-1) + c(1)e(t-1).$$

The problem of estimating the parameters, $b(1)$, $c(1)$ and $d(1)$ of an ARMAX model from n past observations of the output variable (x), control variable (u) and prediction error (e) can be transformed into the problem of solving $n-1$ linear equations with 3 unknowns, which can be solved by multiple linear regression (see Chapter 10).

$$\begin{aligned}
x(2) &= b(1)x(1) + d(1)u(1) + c(1)e(1)\\
x(3) &= b(1)x(2) + d(1)u(2) + c(1)e(2)\\
&= \ldots\ldots + \ldots\ldots + \ldots\ldots\\
x(t) &= b(1)x(t-1) + d(1)u(t-1) + c(1)e(t-1)\\
&= \ldots\ldots + \ldots\ldots + \ldots\\
x(n) &= b(1)x(n-1) + d(1)u(n-1) + c(1)e(n-1)
\end{aligned}$$

or in matrix notation:

$$\begin{pmatrix} x(2) \\ x(3) \\ . \\ x(N) \end{pmatrix} = \begin{pmatrix} x(1) & u(1) & e(1) \\ x(2) & u(2) & e(2) \\ . & . & . \\ x(N-1) & u(N-1) & e(N-1) \end{pmatrix} \begin{pmatrix} b(1) \\ d(1) \\ c(1) \end{pmatrix}$$

A numerical example given in Table 20.3 illustrates the way model parameters are estimated. These N process observations constitute a training set to identify the process model. In practice model parameters can be continually updated by adding a new observation and dropping the first one. This allows to catch up with changes in process behaviour. During this updating procedure the estimated model parameters may change. Such changes indicate which part of the process deviates from the

TABLE 20.3

Modelling of a time series with an ARX-model

t	$\mathbf{x}(t)$	$\mathbf{u}(t)$	$\mathbf{x}(t+1)$
1	1	−1	0.558
2	0.558	−1	0.280
3	0.280	−1	0.088
4	0.088	−1	−0.083
5	−0.083	+1	0.056
6	0.056	−1	−0.024
7	−0.024	+1	0.038
8	0.038	−1	−0.040
9	−0.040	+1	0.050
10	0.050	−1	−0.109
11	−0.109	+1	0.059
12	0.059	−1	−0.081
13	−0.081	+1	0.073
14	0.073	−1	−0.027
15	−0.027	+1	0.037
16	0.037	−1	−0.115
17	−0.115	+1	0.015
18	0.015	−1	−0.119
19	−0.119	+1	0.063

Process values are simulated with:
$x(t+1) = b(1)x(t) + d(1)c(t) + e(t+1)$ with $b(1) = 0.7$; $d(1) = 0.1$
$e(t)$ is uniformly distributed in the interval [−0.05 + 0.05]
$x(1) = 1$
The settings of the control variable $u(t)$ are:
if $x(t) > 0$ then $u(t) = -1$ else $u(t) = +1$
The parameters estimated by solving equation $\mathbf{x}(t+1) = b(1)\mathbf{x}(t) + d(1)\mathbf{u}(t)$ for $b(1)$ and $d(1)$ are:
$\hat{b}(1) = 0.662$; $\hat{d}(1) = 0.103$.

Normal Operating Conditions (NOC) of the process. For instance a change in the correlation structure of the process output is indicated by a change of the parameter $b(1)$ of the autoregressive part of the model. Such changes may be monitored by plotting the value of each model parameter in a Shewhart control chart, according to the rules explained in Chapter 7. Possible model deviations (e.g. from first-order to second order) are checked by inspecting the residuals $e(n + t)$ obtained by substituting a new process observation in the model equation derived from the training set and plotting this residual in a Shewhart control chart. The residual $e(n + t)$ of a new process value at time $n + t$ is given by:

$$e(n + t) = x(n + t) - \{b(1)x(n + t - 1) + d(1)u(n + t - 1) + c(1)e(n + t - 1)\}$$

Because the autoregressive parameter $b(1)$ is equal to the autocorrelation of $x(t)$ at $\tau = 1$, the residual plot, $e(t)$ is called the *autocorrelation control chart*. In this chart warning and action lines are constructed according to the rules of the Shewhart control chart (see Chapter 7). Crossing of the action line indicates that the process model is no longer valid. No indication, however, is obtained about which part of the process is disturbed.

ARMA models are suitable for fitting stationary time series. A steady increase of the prediction error indicates the presence of drift, which is a deviation from the stationary state. In this situation an *Autoregressive Integrated Moving Average* (ARIMA) model [6,10,11] should be used.

20.5 Measurability and measuring system

As explained before, the specifications of the measuring system define the ability to observe and to regulate a process. These specifications should be considered in relation to the process characteristics, e.g. the time constant T, which can be derived from the autocorrelation function of the process. There are four relevant specifications:

s_a^2/s_0^2: Variance of the measuring system relative to the variance of the process fluctuations.

T_d/T: Dead time (relative to the time constant), which is the time between the time when the sample is taken and the point when the correction is made (this includes the analysis time).

T_a/T: Time (relative to the time constant) between two consecutive sampling times (sampling interval).

T_m/T: Time (relative to the time constant) during which the sample is collected (sampling time).

Leemans [5] derived a relation between the measurability (m) (eq. (20.1)) and these four specifications of the measuring and control system, using the autocorrelation function as a predictor in the case of control or as an interpolator in the case of description (or reconstruction). For process control this relation is:

$$m = e^{-\frac{T_d}{T}} \; e^{-\frac{T_a}{2T}} \; e^{-\frac{T_m}{3T}} \left(1 - \frac{s_a}{s_0}\sqrt{\frac{T_a}{T}}\right) \tag{20.8}$$

Equation (20.8) shows that the dead time (T_d) has the largest effect on the measurability and the sample collection time (T_m) the smallest. For process reconstruction by modelling, there is no dead time and therefore eq. (20.8) reduces to:

$$m = e^{-\frac{T_a}{2T}} \; e^{-\frac{T_m}{3T}} \left(1 - \frac{s_a}{s_0}\sqrt{\frac{T_a}{T}}\right) \tag{20.9}$$

Figure 20.14 illustrates how the autocorrelation function is applied to interpolate between the sampling points [S1, S2, S3,...] (Fig. 20.14a) or to extrapolate from the last to the next reporting time [R1, R2,...] (Fig. 20.14b). The graphical representation (see Fig. 20.15a,b) of eq. (20.8) (with $T_m = 0$) can be applied to derive the best measuring strategy in practice, as we will now proceed to describe. Measurabilities

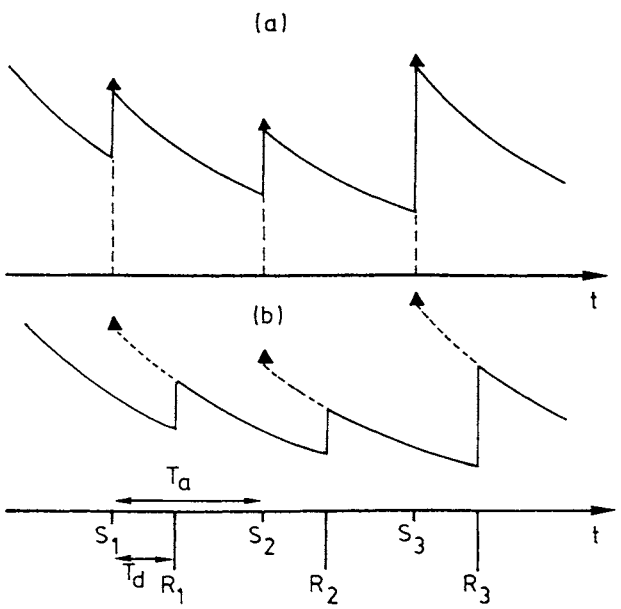


Fig. 20.14. Process interpolation (a), between sampling points and extrapolation (b) from one reporting time to another with an autocorrelation function. ▲ represents the measurements at time S, T_d is the dead time, T_a is the time between two samples, the solid line represents respectively (a) the interpolated process values and (b) the extrapolated process values.

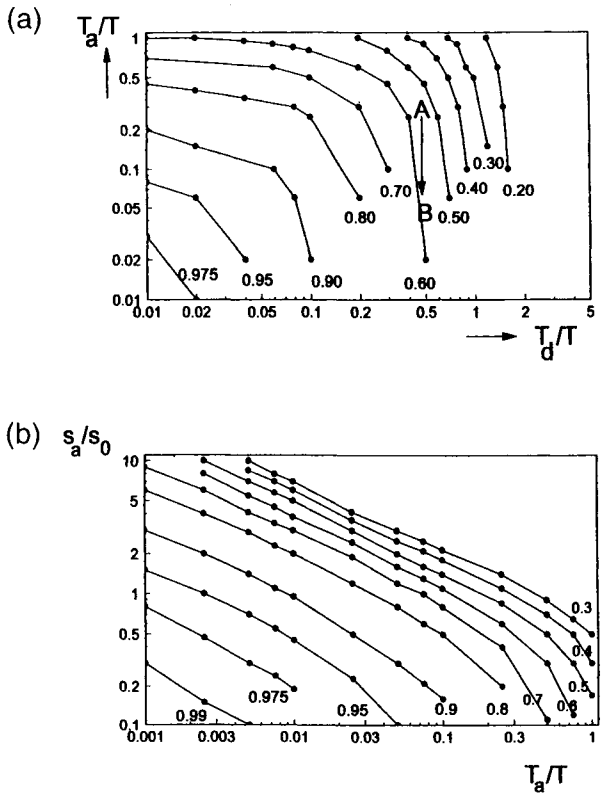


Fig. 20.15. Measurability as a function of the properties of the measuring system. (Adapted from Ref. [5]). (a) Measurability as a function of the analysis time (T_d) and the sampling interval (T_a), relative to the time constant T of the process. (b) Measurability as a function of the precision of the analysis (s_a) relative to the standard deviation of the process variations (s_0) and the sampling interval (T_a).

below 0.9 are not very practical. From eq. (20.1) it follows that for $m = 0.9$ the standard deviation of the process fluctuations left after control is still about half the value of the uncontrolled process ($s_e/s_0 = 0.43$). As one can see there is a large area in the space of the specifications of the measuring device where the measurements are ineffective ($m < 0.9$ in Fig. 20.15). One also observes that actions to improve or replace an ineffective measuring system may fail. For example in situation A, the decision to increase the sampling frequency (A → B) has no noticeable effect on the performance of the measuring system.

Equation (20.8) can also be used to estimate the effect of replacing a slow but precise method by a faster and less precise measurement device, which can be operated with a larger sample throughput, e.g. a Near Infrared Analyzer. Because of the shorter analysis time and the in-line continuous monitoring capabilities, high measurabilities are obtainable.

20.6 Choice of an optimal measuring system: cost considerations

When designing a control system two relationships have to be considered for all the analytical methods under consideration:

(1) the measurability of the measuring system as a function of the sampling frequency; and

(2) the cost of analysis as a function of sampling frequency.

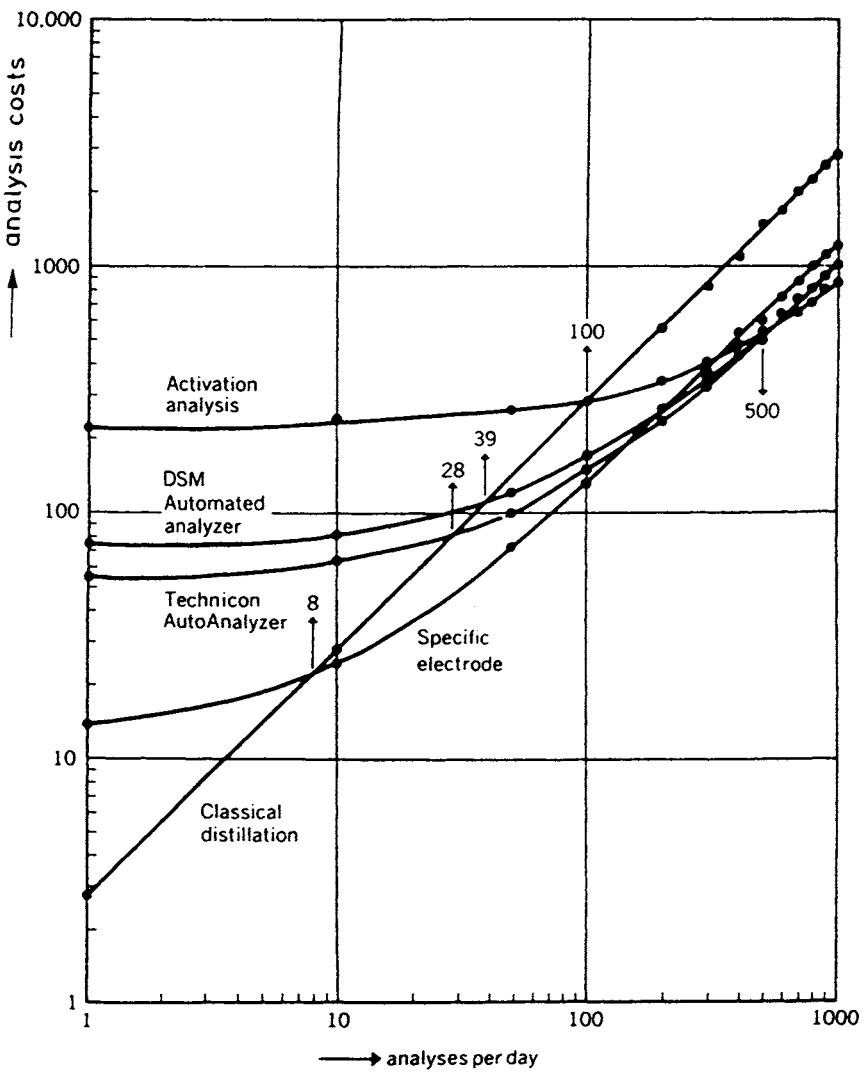


Fig. 20.16. Analysis cost of the methods listed in Table 20.1 as a function of sampling rate. (Adapted from Ref. [5]).

With these two relationships the measurability–cost relationship can be derived for each candidate method. Process and/or marketing engineers should provide the return obtained by reducing process variations (better specifications, less rework), by which the return can be calculated as a function of measurability. Cost of analysis as a function of sampling frequency should be provided by the laboratory. An example of such relationships is given (see Fig. 20.16) for labour intensive methods (e.g. classical distillation) and high-capacity analyzers (e.g. activation analysis). By plotting the effective return as a function of measurability, the most cost-effective control system can be selected.

Let us illustrate this procedure with the fertilizer plant, introduced in Section 20.2. Suppose that by an autocorrelation analysis a time constant has been found equal to 66 minutes. By substitution of the method specifications given in Table 20.1 into eq. (20.8) (with $T_m = 0$), and by varying T_a, the measurability is obtained as a function of the sampling rate. Sampling intervals which are smaller than the dead time are not considered as at that point the measurability reaches a constant value. From the curves obtained (see Fig. 20.17) one can observe that the methods 1 to 5 hardly reach measurabilities above 0.7, even at the highest possible sample throughput. Method 6 performs better, whereas large sample throughput are possible with method 7 giving a high (but still lower than 0.9) measurability. However in this particular case these high sample rates could only be achieved by replacing the manual sampling procedure by an at-line operation. The final decision should be based on the cost of running a control system and the return obtained by the better quality of the product. The plots which relate the measurability to the sampling rate (Fig. 20.17) and the analysis cost to the sampling rate (Fig. 20.16) can be combined to give the relationship between the analysis cost and measurability (Fig. 20.18). The solid lines in Fig. 20.18 are the cost for manual operation, whereas the arrows indicate the measurability/cost value for at-line operation. When

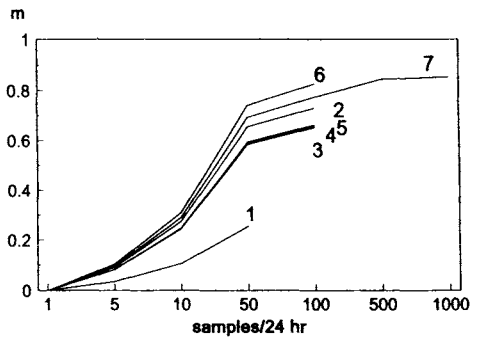


Fig. 20.17. Measurability of the methods listed in Table 20.1 as a function of the sampling rate to control a process with $T = 66$ min and $s_0 = 1.2\%$ N.

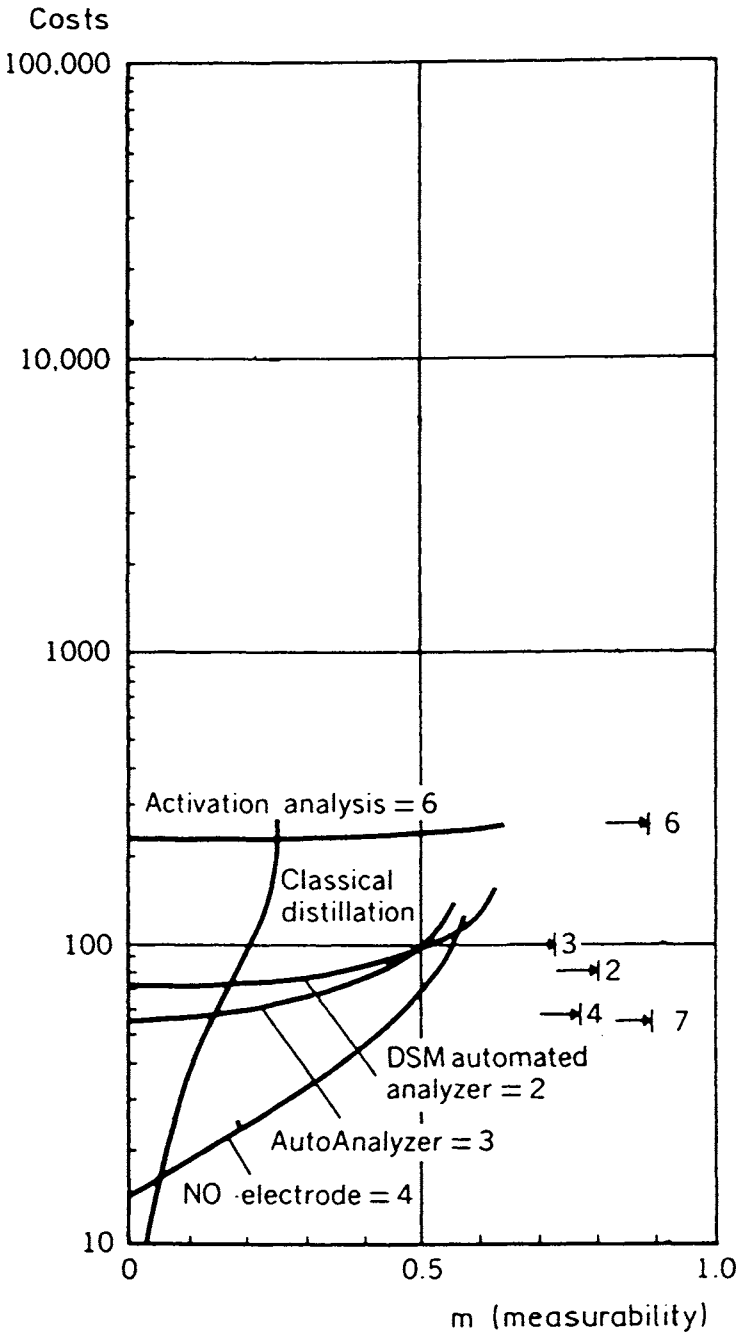


Fig. 20.18. Analysis cost of the methods listed in Table 20.1 as a function of the measurability. The arrows indicate the measurability and associated cost for at/in line analysis. (Reprinted from Ref. [5]).

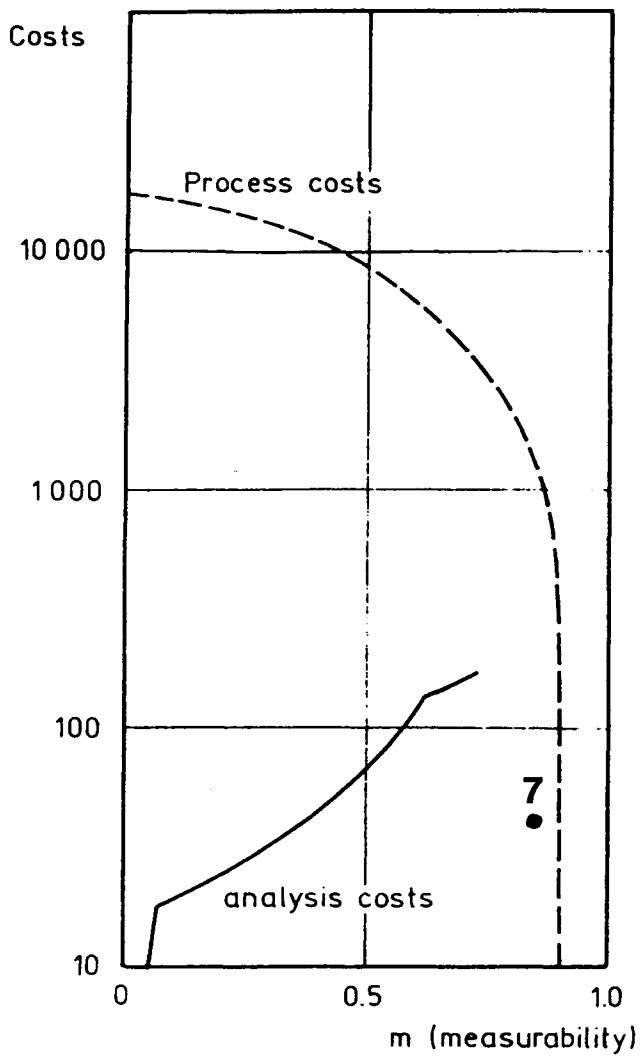


Fig. 20.19. Lower boundary line of the cost functions and the best at/in line point in Fig. 20.18. The dashed line represents the process costs. (Reprinted from Ref. [5]).

plotting the process cost and the analysis cost (boundary line in Fig. 20.18) in a single graph (Fig. 20.19), it is obvious that method 7, the at-line γ-ray absorption method, is the preferred method.

20.7 Multivariate statistical process control

When discussing measurability we assumed that the process is statistically in control, i.e. the mean and standard deviation of the regulated variable are time invariant. However, in practice deviations from the stationary state may be observed, for example, because the process average value drifts away from the target value. It can be considered as a slow source of variation with a large time constant. Drift can be detected by inspecting the autocorrelation function of a set of historical process observations (see Fig. 20.11). The correlation is positive at high τ-values, where one would expect a value near to zero. This indicates an up-slope drift. In that case the Shewhart control chart of the parameters of the ARMAX process model (see Section 20.4.4) will indicate that the autoregressive parameter, $b(1)$, is outside the normal operating conditions.

Real-time flagging of out-of-control situations is important in a manufacturing environment. The number of false positive alarms should be avoided as they lead to unnecessary inspection of the plant or rework of the product. For a conventional control chart this will happen in about three out of 1000 situations. Matters become different when several process values are monitored at the same time, e.g. the solids content of a margarine at six different temperatures. If these six values are independent and are monitored in separate control charts, the probability of a false alarm is 0.018 ($1–0.997^6$), which is unacceptably high. If all quality parameters were perfectly correlated, it would be sufficient to monitor only one, and the false alarm rate would remain three out of 1000. In reality process variables are neither perfectly dependent nor perfectly independent. Therefore, the false alarm rate will lie somewhere in between these two extremes. A more desirable approach is to fix the risk of a false alarm at a given level, e.g. five out of 1000 regardless of the correlation between the process variables. This can be achieved by defining a *multivariate control chart.* It is also possible that all individual control charts indicate an in-control situation, but that the small deviations from the target have a cumulative effect and may result in an overall out-of-control situation. Therefore, one should consider the joint probability of finding a certain combination of process values. For two process variables the joint iso-probability is given by the elliptical function (see Chapter 8) shown in Fig. 20.20. In terms of process control it would mean that the square area indicates an in-control situation when the two control charts are considered individually, whereas the smaller area inside the ellipse is the true in-control situation. The larger the correlation between the two variables the more the ellipse is stretched and deviates from the square. In process control the centre of the figure represents the combined target value of the two parameters, e.g. $[t_1,t_2]$, or $\mathbf{t}$. By multivariate control one considers the distance of any observation $[x_1,x_2]_i$ or $\mathbf{x}_i$ from the target value, thus the distance between $\mathbf{x}_i$ and $\mathbf{t}$. However this distance cannot simply be the Euclidean distance. For example the

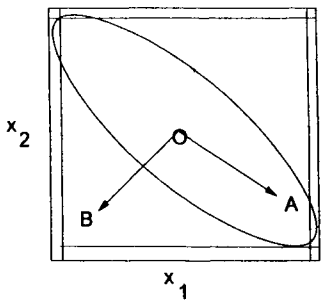


Fig. 20.20. A bivariate probability distribution of two variables (x_1,x_2). The in-control point A and out-of-control point B are at the same Euclidean distance from the target value **t**.

points A and B in Fig. 20.20 have the same Euclidean distance from the target whereas A is in-control and B is not. Therefore, a measure for distance is necessary, that takes into account the correlation between the variables. Such a measure is the Mahalanobis squared distance, which is defined as (see Chapters 9, 10 and 31):

$$d_{xt}^2 = (\mathbf{x}_i - \mathbf{t})^T \, \mathrm{Cov}^{-1} (\mathbf{x}_i - \mathbf{t}) \tag{20.10}$$

Cov is the variance–covariance matrix estimated from a sample of n past multivariate observations.

When d_{xt}^2 exceeds a defined critical value, the process is out of control. Because the Mahalanobis distance follows the Hotelling T^2-distribution [10], the critical value T_{UCL}^2 is defined by:

$$T_{UCL}^2 = \frac{(n-1)p}{n-p} F_{\alpha;p,n-p} \tag{20.11}$$

where n is the sample size used to calculate Cov(**x**), p is the number of process variables, F is the F-statistic with (p,$n - p$) degrees of freedom, and α is the accepted risk of false alarms.

By plotting T^2 in a control chart with an upper control limit (T_{UCL}^2) a multivariate control chart is obtained. Because the distance is always a positive number, the chart only contains an upper control limit. It has the same interpretation as a univariate control chart: a point outside the limit corresponds to an out-of-control situation. This indicates that the process has to be investigated for a possible cause (Fig. 20.21).

The calculation of the Mahalanobis distance from eq. (20.10) requires the inverse of the variance covariance matrix of n past multivariate observations. When the number of measured process variables is large, they may be highly correlated and lead to a nearly singular variance covariance matrix. In Chapter 17 we discussed principal components analysis (PCA) as a technique to decompose

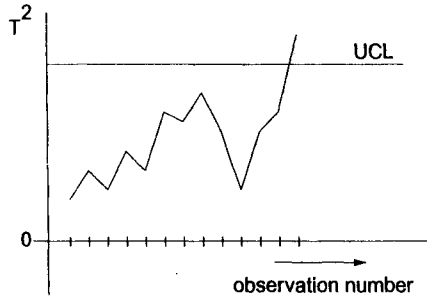


Fig. 20.21. A multivariate T^2 control chart. UCL is the upper control limit; T^2 is the Mahalanobis squared distance from the target value.

a nearly singular matrix into a product of a loading matrix (**V**), a score matrix (**U**) and a diagonal matrix $\mathbf{\Lambda}$ containing the singular values which are the square roots of the eigenvalues of $\mathbf{X}\,\mathbf{X}^T$, giving $\mathbf{X} = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{V}^T$.

As discussed in Chapter 17, the first a PCs are associated with structure in the data, whereas the remaining PCs are associated with random variations. Instead of calculating the Mahalanobis distances in the original data space we can calculate the distance between the scores of the target vector, $\mathbf{t}$ and the observation $\mathbf{x}_i$ in the space defined by the first a PCs. Usually the target value coincides with the mean value of the data. After mean centring of the data, the Hotelling T_i^2 of observation i with respect to the target value is then given by:

$$T_i^2 = \sum_{k=1}^{a} \frac{u_{ik}^2}{s_{u_k}^2}$$

where $s_{u_k}^2$ is the variance of the scores on PC_k (this variance is equal to $\lambda_k^2/(n-1)$ when the scores are defined as $\mathbf{U\Lambda}$).

For $a = 2$, T_{UCL}^2 describes an ellipse in the PC-space defined by PC_1 and PC_2. For $T^2 > T_{UCL}^2$ points are situated outside the ellipse and indicate an out-of-control situation. This procedure is equivalent to the standard Hotelling T^2 control chart with the exception that the distance between the data is not measured in the original data space but in the reduced PCA space. If all PCs are included in the process model both give exactly the same outcome. The optimal number of PCs to be included in the PCA model is determined by cross-validation as discussed in Chapters 10 and 31. Cross-validation is, however, less performable for nearly random data. Therefore we prefer to apply a randomization test described in [12]. The predictive error sum of squares (PRESS) obtained by the cross-validation procedure is a measure for the magnitude of the residuals. For each new observation a squared prediction error (SPE) can be calculated with respect to the PCA model:

$$SPE = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)^2$$

(n = number of observations, p = number of variables).

Normal Operating Conditions of a process with two significant PCs are now defined by a cylinder with its base in the plane defined by the principal components and with a height equal to SPE. The base of the cylinder is equal to T^2_{UCL} (eq. (20.11)) and the height of the cylinder is proportional to the residuals with respect to the process model (Fig. 20.22). We can therefore follow the evolution of the process by monitoring two control charts [13]. The first one is the ordinary T^2 chart based on a principal components and the second one is a chart which plots SPE for each new observation and compares it with an upper warning and upper control limit.

A number of different types of deviations from NOC can be observed.

– The observation falls outside T^2_{UCL}, but SPE is within the normal limits. In the case of $a = 2$ this means that the point falls outside the cylinder, but remains within the allowed distance from the PC-model (points 'x' in Fig. 20.22). The principal components model is thus still valid, implying that the model relation is unchanged. Because one or more scores of the observation must be the cause for $T^2 > T^2_{UCL}$, the process is apparently upset in one or more variables.

– The observation falls inside T^2_{UCL} and SPE is outside its normal limits (points '+' in Fig. 20.22). This is an interesting case as this situation would not be detected by using a T^2 chart alone. A too large SPE indicates that a process disturbance is introduced which makes the principal components model invalid (by the appearance of a new source of variation).

In order to be able to intervene in the process, we should be able to assign deviations from NOC to a possible cause (one or a combination of process

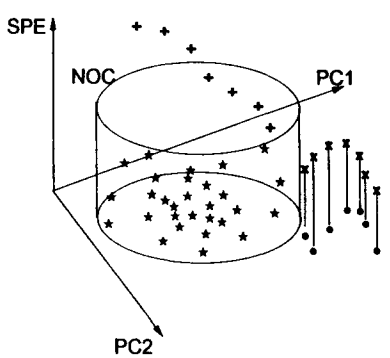


Fig. 20.22. A multivariate control chart in the plane defined by two principal components (PC-1 and PC-2) of the control data. SPE represents the squared prediction error, NOC is the normal operating condition. ★ are in-control points, x and + are out-of-control situations, ● are the projections of x on the PC1–PC2 plane. (Adapted from Ref. [13]).

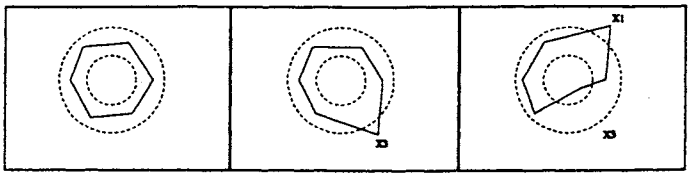


Fig. 20.23. Representation of the values of 6 process variables in a star. Dashed lines are the lower and upper univariate action limits. Each point of the star represents the value of one of the 6 process variables. See text for further explanation of the three situations.

variables). An approach which can be used with regular T^2 charts constructed in the space of the original variables is to combine the multivariate and univariate control charts in a so called *Star Chart* [14]. At each point of the control chart two concentric circles are constructed with a radius equal to the standardized univariate upper and lower control limits. A star is constructed with the corner points representing the values of the individual quality parameters. The distance of the corner (i) from the centre is equal to:

$$(x_i - \bar{x}_i)/s_i \tag{20.12}$$

In Fig. 20.23 three such stars are shown. In star I all six variables lie within their limits. In star II the third variable is out-of-control. In star III two variables are out-of-control; x_3 is too low and x_1 is too high. In a next step the stars are plotted in the multivariate control chart. The centre of the circles is at the Mahalanobis distance from the target, and the stars show the position of each variable with respect to their own control limits. Several situations may occur. Fig. 20.24a shows a process that is in-control in both the multivariate and univariate sense. The star chart in Fig. 20.24b indicates which variable may cause the multivariate out-of-control situation (point indicated $c10$). Fig. 20.25a illustrates the situation where the process is in-control although one univariate parameter ($n35$) is out-of-control. It may also occur that the process is out-of-control without any univariate parameter being out-of-control (Fig. 20.25b).

A second approach is to evaluate the variance of the PCA residuals for each variable j separately by [14]

$$s_j^2 = \sum_{k=a+1}^{p} v_{jk}^2 \lambda_k^2 / (n-1)$$

where v_{jk} is the loading of the jth variable in the kth principal component, λ_k is the singular value associated with this principal component, a is the number of PCs included in the PCA model. One should realize that the calculation of s_j^2 requires at least a window of a+2 new observations to retain 1 degree of freedom to test changes in s_j^2 by an F-test:

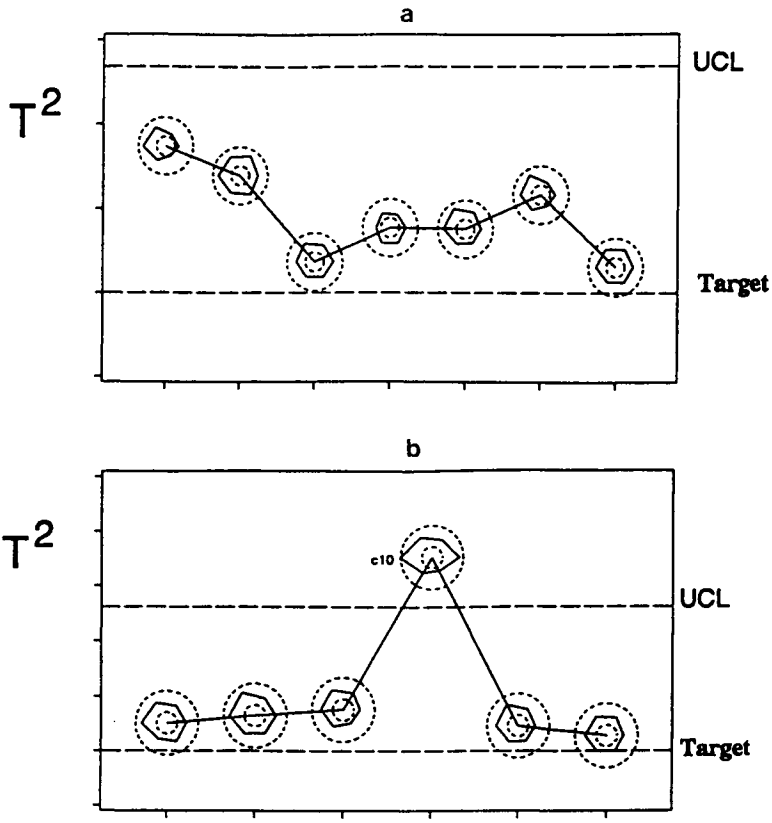


Fig. 20.24. Two Star Charts: (a) Univariate and multivariate in-control, (b) Univariate and multivariate out-of-control ($c10$).

$F = s_j^2(\text{new}) / s_j^2(\text{old}) \geq F_{\alpha;\nu_{new},\nu_{old}}$

with $\nu_{old} = n - a - 1$, ν_{new} = [window size] $- a - 1$; s_j^2(new) is the variance of the PCA residuals of variable j in the selected window of observations, and s_j^2(old) is the variance of the PCA residuals of variable j in all previous (in-control) observations.

Multivariate control charts are applicable to check whether an analytical method or instrument is statistically in control (see e.g. [15]). Let us for instance consider the example of the determination of two triglycerides of the type SOS and SSO by HPLC (S = Saturated fatty acid, O = oleic fatty acid). Figure 20.26 gives the two individual univariate control charts obtained by plotting the values of a check sample measured at regular intervals. Strictly speaking the independent use of these two control charts is only allowed when the variation in the SOS and SSO concentrations is uncorrelated. In this particular instance however significant correlations are found between SSO and SOS ($r = -0.4$), which indicates that the usage of a multivariate control chart shown in Fig. 20.27 is appropriate. In this example none of the univariate and multivariate

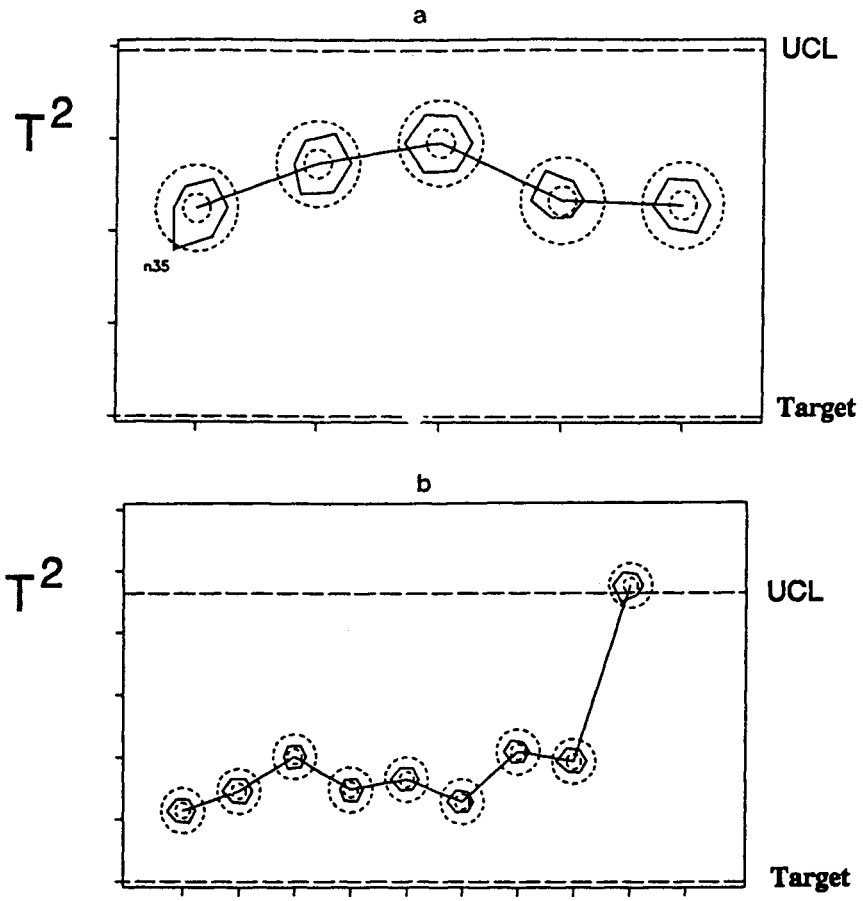


Fig. 20.25. Two Star Charts: (a) Multivariate in-control but n35 out-of-control, (b) Univariate in-control but multivariate out-of-control.

control charts exceeds the action limits. The two univariate control charts also stay within the warning limits. The multivariate charts warn at three occasions (days 16 and 17, and day 31). These are the situations where both SOS and SSO values are high, which is improbable due to the negative correlation.

When discussing ARMAX models (see Section 20.4.4) we related the value of a regulated variable to its own past values and the past values of a control variable. In case that several process variables are regulated by a number of control variables, we can model the relationship by a partial least squares (PLS) model (Chapter 35) and use that relationship to design a T^2 chart in the PLS factor space to monitor the stability of that relationship [17]. Additional reading on multivariate statistical process control can be found in the tutorial by Kourti and MacGregor [17], and in the papers by MacGregor and coworkers [13] and by Wise et al. [16].

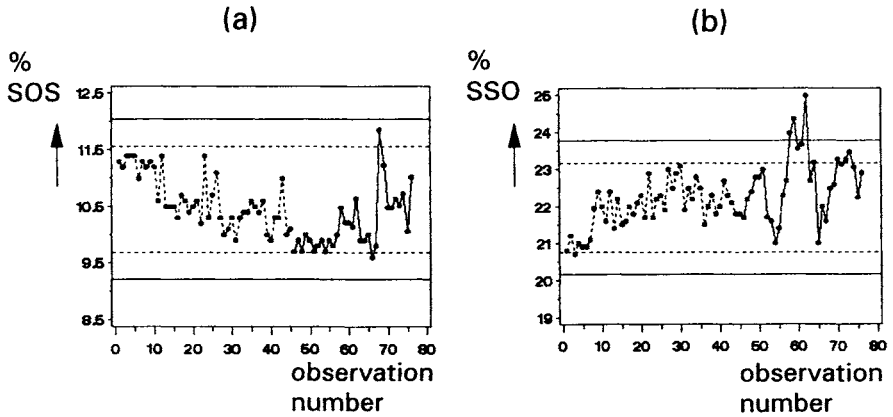


Fig. 20.26. Shewhart control charts of the relative concentrations of the SOS (a) and SSO (b) triglycerides in a check sample. The dotted line represents the learning data and the solid line represents the control data.

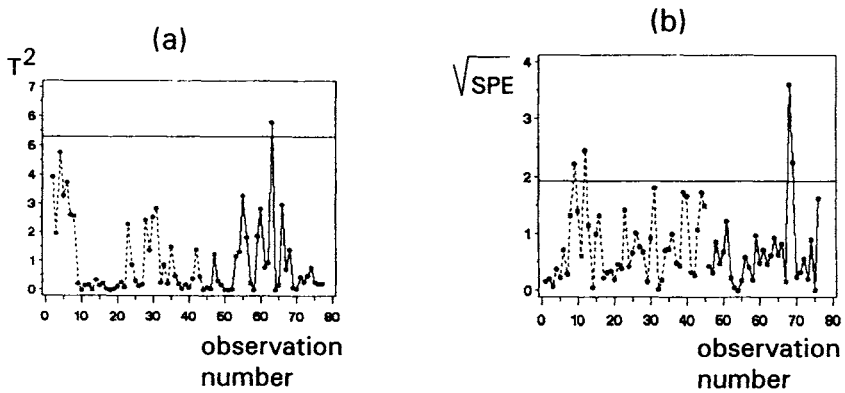


Fig. 20.27. T^2 chart and $\sqrt{SPE}$ chart of the relative concentrations of the SOS and SSO triglycerides (data from Fig. 20.26). The dotted line represents the learning data and the solid line represents the control data.

20.8 Sampling for spatial description

As discussed in Section 20.1, systems or objects can also be sampled to obtain a more or less detailed description of a certain property (e.g. the concentration of a given compound) in a static situation. The objectives are different from those described in earlier sections since no time variations are considered and henceforth different sampling and estimation strategies are necessary. This will be the subject of the remainder of this chapter. Two main kinds of requests may be involved: (i) global description in which the goal is to obtain an overall (global) estimate of a

property over the whole object and (ii) prediction or interpolation when a more precise description is required. It is important to pay proper attention to the required precision of the obtained information, since this will determine to a great extent the size of the sampling scheme. It is therefore necessary that the objectives of the study are clearly stated and translated in terms of required precision of the outcome.

It should be emphasized at the outset that many methods and strategies for spatial description contain a considerable empirical component. Much research is still necessary to improve these sampling strategies.

20.9 Sampling for global description

To obtain an overall estimate of a property an estimate of the population mean value is required. A statistical summary parameter such as the mean value, together with a confidence measure, is normally used. To obtain this estimate a sample that is representative for the whole system must be taken. Classical statistical sampling theory treats this problem extensively. A good textbook on this subject is Cochran [18]. Basically one distinguishes probability or random sampling, non-probability sampling and bulk sampling. The first two approaches assume that the object consists of distinct identifiable units (also called *sample units*) such as tablets, packaged items or persons. Bulk sampling involves taking samples of an object that does not consist of such identifiable units. They may consist of a single pile such as coal, fertilizer or it may be a large object such as a lake.

20.9.1 Probability sampling

When *probability sampling* is applied the sample units are selected according to statistical sampling principles and probability. In probability sampling each sample unit has equal chance to be selected. The appropriate number of sample units depends on the required precision of the estimate. The basic principles for this are given in Chapter 3. A key assumption in this approach is the normal distribution of the mean estimated from the sample. When the sample size is large enough this condition can be assumed to be fulfilled. The standard deviation of the estimated mean depends on the standard deviation of the property in the parent population (the object). Prior information on this population standard deviation is required to estimate the required sample size.

One can distinguish simple random sampling and stratified random sampling. In simple random sampling any sample unit of the whole object has an equal chance of being selected. In stratified random sampling the parent object is first subdivided into non-overlapping subpopulations of units, called *strata*. Simple

random sampling is then applied on each subpopulation. There may be many reasons for stratification. If, for example, different precisions are required for the different subpopulations, the stratification approach allows to treat the separate subpopulations in their own right. Suppose for example that one wants to assess the alcohol percentage in a batch of beer bottles. These bottles are meant for export to different countries. The legislation in the different export countries, however, requires different precision of the alcohol content. Subdividing the bottles into different groups is then advisable, according to their export destination. The simple random sampling strategy can then be applied on the separate groups or subpopulations and precision requirements can be taken into account.

20.9.2 Non-probability sampling

The selection of the units can also be done on a non-probabilistic basis because sometimes it is not feasible or desirable to apply probability sampling. Not all units may be readily accessible for sampling and the units are then selected according to their accessibility, or sampling cost. Examples of such strategies are systematic sampling, judgement sampling and convenience sampling.

In *systematic sampling* the units are selected on a systematic basis. A first unit is selected at random and thereafter e.g. every 5th or 10th unit is taken. Systematic sampling is often applied because it is easier to carry out without mistakes. The systematic selection may, however, cause bias in the result, especially if an unexpected periodicity is present in the object (see Fig. 20.28). For example, the measured values will be sytematically too high when the sampling strategy represented by the crosses (x) is applied. This effect is avoided when simple random sampling is used. Systematic sampling is also applied for predictive purposes (see Section 20.10).

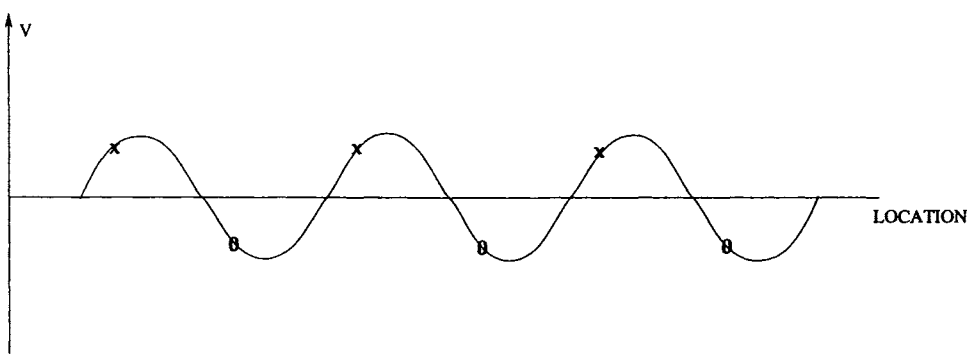


Fig. 20.28. Two systematic sampling strategies (0 and x) in one dimension that yield wrong results. *V* is the value of the property of interest.

In *judgement sampling* the units are selected based on the sampler's own experience on what representative units are. In *convenience sampling* the units are selected in a pragmatic way, based on their accessibility or sampling cost. Although the results of these selection strategies cannot be used to judge the object statistically, they can be quite useful, e.g., to guide future investigations. It is however dangerous to generalize the conclusions.

20.9.3 Bulk sampling

Bulk sampling involves the sampling of an object that does not consist of identifiable units. Each single portion (unit) that is taken from the bulk material is called an *increment*. The term increment may be confusing because it is not used in the meaning of 'addition'. It is, however, a well established term in sampling theory and it is used there as an incremental portion to be combined to form the eventual sample. An increment, taken from bulk material has the same meaning as a sample unit, taken from a packaged material. The terminology as recommended by IUPAC is used in this chapter [19].

When the object to be sampled is homogeneous, such as a bottle of whisky in which the alcohol has to be determined, it is clear that one sample of whatever size is representative. The determining factor is then the minimum amount that is needed for the analytical determination, e.g. the limit of determination or quantification of the analytical method.

When the object to be sampled is heterogeneous, additional aspects such as the size of the increment that is sampled, where they should be collected and how they should be reduced to a suitable laboratory size must be considered. The basis for this sampling research was laid by Baule and Benedetti-Fischler [20], Visman and colleagues [21,22] and Gy [23,24]. They considered objects consisting of two or more types of particles of equal size but with different composition. An increment is then a number of particles and the collection of a number of increments constitutes the sample. This situation can be compared with a bag containing red and white balls, from which a number of balls is taken. The difference between the composition of the increment and that of the whole object, i.e. the sampling error, depends on the size of the increment. It is possible to estimate the uncertainty (standard deviation) of the composition of the increment. This standard deviation depends on the overall composition of the object and on the size of the increment. For an object that consists of two types of particles, A and B, in a ratio P_A and P_B the standard deviation of the sample composition of the object, s, can be estimated from the binomial distribution (see Chapter 15):

$$s = (nP_AP_B)^{1/2} \tag{20.13}$$

$$s_g = \frac{100}{n} s = 100\left[\frac{P_AP_B}{n}\right]^{1/2} \%$$

where s_g = relative standard deviation (in %) of the composition of the increment, P_A = fraction of particle A in the whole object, $P_B = 1 - P_A$ = fraction of particle B in the whole object, and n = number of particles in the increment.

Suppose the object consists of 50% A particles and 50% B particles. Increments, consisting of 10 particles will have a composition whose mean will be 50% A with a standard deviation, s_g.

$$s_g = 100\left[\frac{0.5\ 0.5}{10}\right]^{1/2} = 16\%$$

This means that about one out of three of those increments will have a composition outside the interval (50 ± 16)%. The probability of finding an increment with only A particles is in this case about 0.1%. From eq. (20.13) it follows that the standard deviation decreases with increasing n. The increment is considered to be representative for the object when no distinction can be made between different increments. This is the case when the standard deviation due to the sampling, s_g is negligible compared to the standard deviation of the analytical method (s_a) or:

$$100\left(\frac{P_AP_B}{n}\right)^{1/2} < s_a$$

Therefore the number of particles required for a representative increment is:

$$n > \frac{P_AP_B\ 10^4}{s_a^2}$$

Different approaches to estimate the standard deviation and the required sample size are possible. Ingamells [25,26] proposed an experimental approach. Based on the knowledge that the uncertainty decreases with increasing sample size, he derived that in many cases the relation $WR^2 = K_s$ is valid. Here, W is the weight of the sample and R is the relative standard deviation (in percentage) of the sample composition. K_s is the sampling constant and can be interpreted as the weight of the sample, required for a relative standard deviation of 1%. The value of K_s may be determined from a diagram as in Fig. 20.29. Such a diagram is obtained by measuring a number of samples with different sample weights, W. For each series of measurements the mean and the standard deviation are determined. The two solid curves represent the upper and lower limits of the 68% confidence interval, i.e. the mean plus or minus one standard deviation. By extrapolation or interpolation K_s can then be estimated as the sample weight where the width of this confidence interval is about 2% of the mean measurement value.

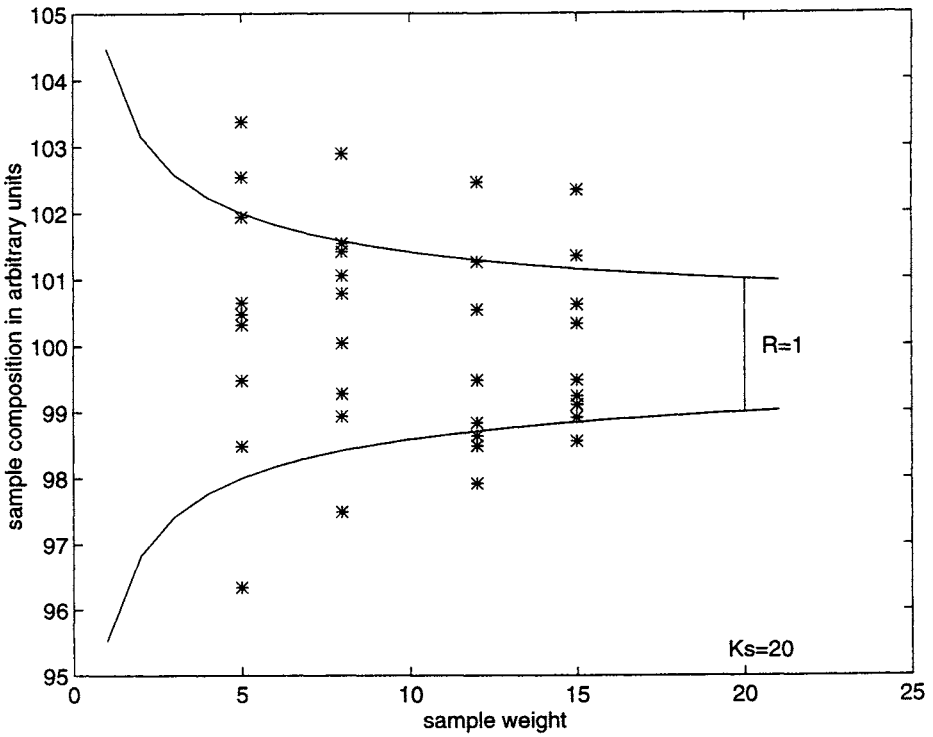


Fig. 20.29. Sampling diagram to determine K_s. See text for explanation.

Gy [23,24] described the most important approaches and derived experimentally an estimate of the standard deviation, that includes shape effects of the particles, especially useful in the mining area. Minkkinen [27,28] developed a computer program based on this approach.

20.10 Sampling for prediction

When investigating the quality of aquatic sediments in a lake it is important to have an estimate of the concentration of the substances at each location. This means that a prediction of the concentration must be possible at locations that are not sampled. The sampling itself is usually done along a systematic grid in contrast to the random selection of probability sampling. These kinds of problems are treated extensively in the field of geostatistics.

In fact there are some strong analogies between the approach for time-series analysis (Section 20.4) and sampling strategies for prediction:

– The starting point of both methods is the fact that no deterministic model can be formulated and a probabilistic approach must be used. For this purpose the so-called *random-function models* are used. In these models the sample data are viewed as being generated by a random process. It is clear that this is not a realistic view but the underlying processes that determine the value are not known and therefore, the property of interest appears as being randomly generated.

– Both methods assume that points that are situated close together (in time for time-series analysis or in space for sampling purposes) are more alike than points that are far away from each other. In sampling terminology this property is called *spatial dependency* or *spatial continuity*

– Another important underlying assumption of both strategies is the *stationarity* of the random function. This means that the relation between points separated by a certain distance (in time or in space) does not depend on their location. This assumption has been made implicitly when generating the time-lag scatter plots $x(t)$ versus $x(t+\tau)$. All pairs of points, separated at a time lag, τ, are included in the graph, regardless of their position.

– An important step in the sampling strategy, as it is in time series analysis, is to obtain a model for the spatial continuity. In time series analysis the autocorrelation function or autocorrelogram and time constant are commonly used. When developing a sampling strategy it is common to use besides the autocorrelogram also the covariogram and the variogram (see Section 20.10.1).

In sampling for prediction purposes the goal is to obtain estimates at non-sampled locations by means of interpolation. Several interpolation techniques are possible to obtain estimates at non-sampled locations. The better techniques take into account the knowledge of the spatial dependency (Sections 20.10.3 and 20.10.4). A good textbook on this subject is written by Isaaks and Srivastava [29].

20.10.1 h-Scatter plots, autocorrelogram, covariogram and variogram

The spatial location of every point can be represented by a vector and so can their differences, **h**, also called the lags. This is shown in Fig. 20.30. $\mathbf{h}_{ij}$ is the vector going from point i to point j. Sometimes it will be important to distinguish it from the opposite vector $\mathbf{h}_{ji}$ the vector going from point j to point i. In an *h-scatter plot* the values of the property of interest for pairs of points are plotted. The pairs consist of two data points that are located a lag, **h,** apart, thus at a certain distance in a particular direction. This is in contrast with the time lag in a time series. $\mathbf{h} = (a,b)$ means a distance a in the x-direction and a distance b in the y direction. In Fig. 20.31 some possibilities are shown of pairing the data in a certain direction at a certain distance.

In practice, it is often not possible to locate the sampling points so regularly, because of accessibility problems or because of uncertainties in sample point

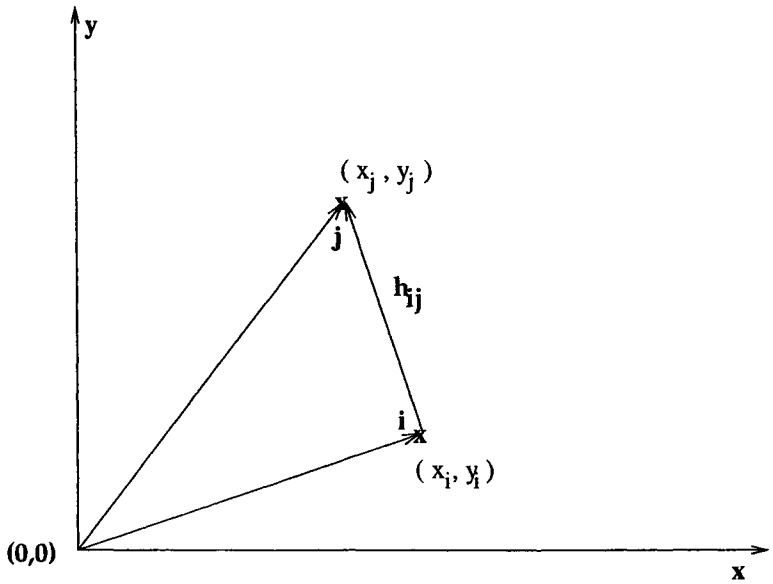


Fig. 20.30. The location vectors and the difference vector, **h**, of two points, *i* and *j*. *x* and *y* represent the 2-dimensional spatial coordinates.

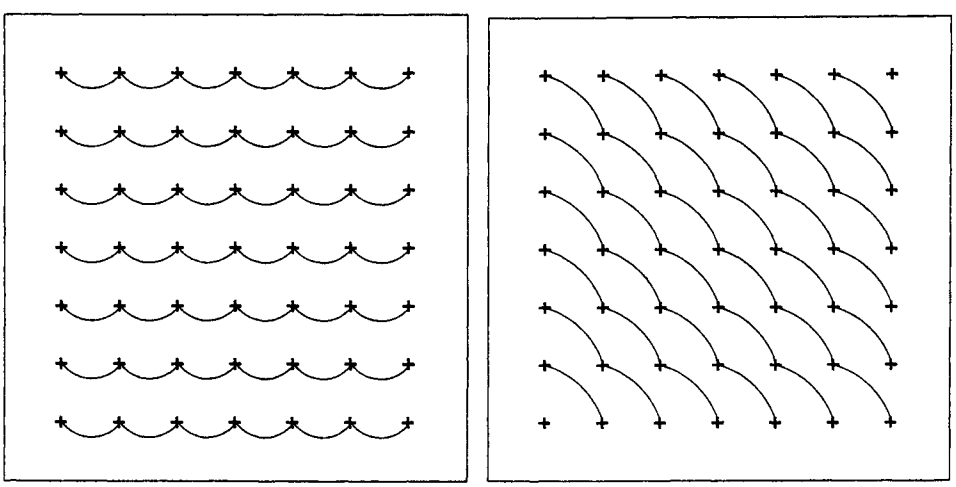


Fig. 20.31. Two ways of pairing sample points (+) in two dimensions for the construction of an **h**-scatter plot.

locations. It is therefore in practice not possible to use **h** as a crisp vector, with a well defined length and direction. It is often necessary to define tolerances on **h**. In Fig. 20.32 an illustration of a tolerance on **h** = (10,0) of 1 meter and 20 degrees is given.

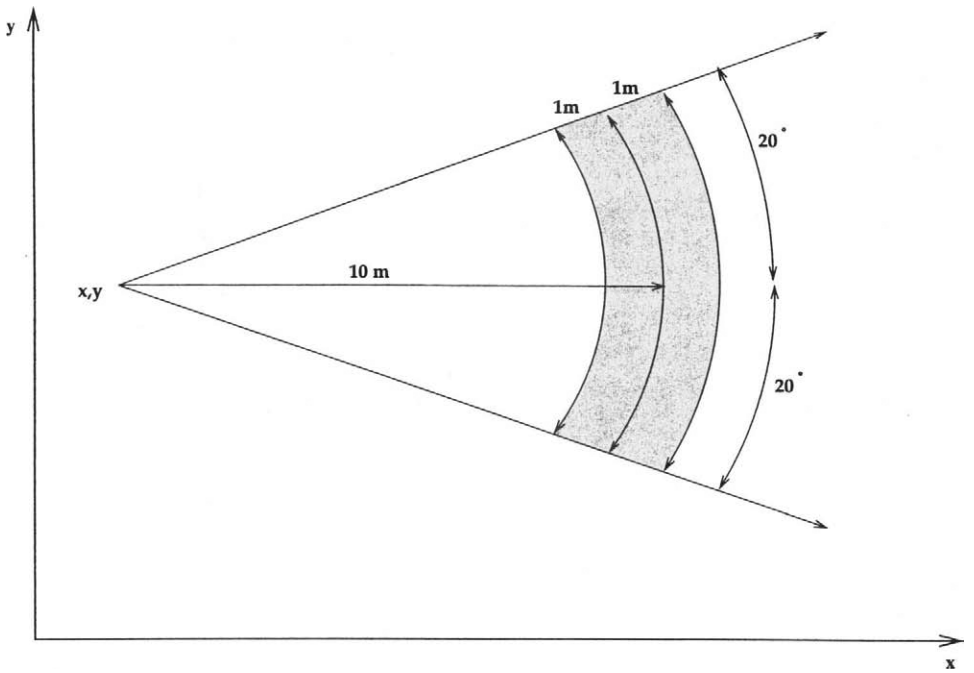


Fig. 20.32. The **h**-vector (10,0). The shaded area represents its tolerance zone (20 degrees and 1 m).

All points in the shaded area are considered to be separated by a lag **h** =(10,0).

The **h**-scatterplot (Fig. 20.33) is the analogue of the time-lag scatterplots in Fig. 20.8. In an **h**-scatter plot the x-axis corresponds to $V(\mathbf{l})$, the value of the property at position **l** and the y axis to $V(\mathbf{l+h})$. From the shape of the clouds of points in the **h**-scatter plot the spatial relationship at a lag **h** can be derived. The correlation coefficient between $V(\mathbf{l})$ and $V(\mathbf{l} + \mathbf{h})$ can be used to measure the relationship (see Fig. 20.33). It is then possible to derive a correlation function (an autocorrelogram) or a covariance function as a function of **h**. It is thus possible to display the autocorrelogram as a contour map that displays the value of the correlation coefficient as a function of the length and of the direction. This sort of display is, however, unusual. One prefers to display the correlogram as separate graphs of the correlation coefficient versus the length of **h** for various directions. This autocorrelogram can be modelled in the same way as the autocorrelogram of time series data (see Section 20.4).

Alternative measures of spatial relation that are often used in geostatistics are the covariance, $C(\mathbf{h})$ and the *semivariance*, $\gamma(\mathbf{h})$, which are defined in eq. (20.14):

$$C(\mathbf{h}) = \rho(\mathbf{h})\, \sigma_{v_i}\, \sigma_{v_{i+\mathbf{h}}} = \rho(\mathbf{h})\, \sigma^2$$

$$\gamma(\mathbf{h}) = \tfrac{1}{2} E[v_i - v_{i+\mathbf{h}}]^2 \qquad (20.14)$$

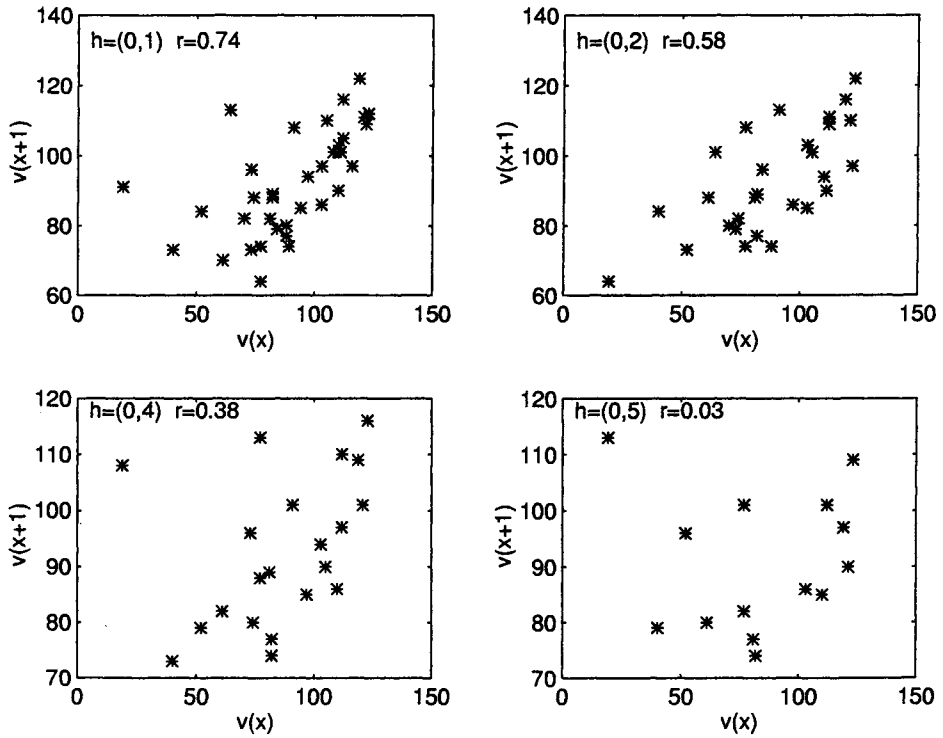


Fig. 20.33. Example of different **h**-scatterplots at increasing separation distances, |**h**| in the x-direction. As the separation distance increases, the similarity decreases and the correlation coefficient, r decreases.

$\rho(\mathbf{h})$ is the correlation coefficient for a lag, **h**, σ^2 is the variance of the property, v, E is the expectation value and v_i and $v_{i+\mathbf{h}}$ are the values of the property of interest of a pair of points at a lag **h**, apart. The semivariance can be estimated by:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2n_{\mathbf{h}}} \sum_{i=1}^{n_{\mathbf{h}}} [v_i - v_{i+\mathbf{h}}]^2$$

$n_{\mathbf{h}}$ is the number of pairs of data points a lag **h** apart. In this book we normally do not use Greek symbols for estimated values, but the Roman counterparts. However, to remain in accordance with the sampling terminology we use here the Greek letter with a hat to denote the estimated semivariance. The semivariance can be interpreted as the mean of the perpendicular squared distances, d^2, of the points in the scatter plot from the 45 degree line.

In Fig. 20.34 the three measures of spatial continuity: the autocorrelation coefficient, the covariance and the semivariance are plotted as a function of |**h**| for a typical stationary random function. The value of $\rho(0)$ in the autocorrelogram (Fig. 20.34a) equals 1. It is the correlation coefficient between $v(i)$ and $v(i + 0)$.

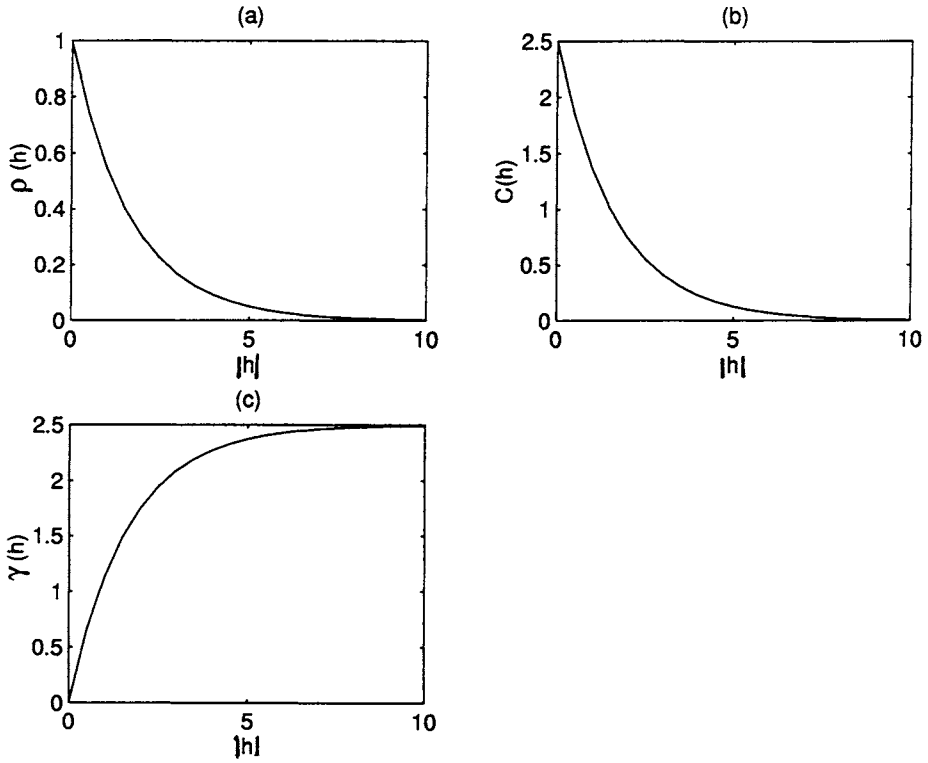


Fig. 20.34. (a) a theoretical autocorrelogram: the correlation coefficient as a function of **h**. (b) A theoretical covariogram: the covariance as a function of |**h**|.(c) A theoretical variogram: the semivariance as a function of **h**.

When the distance increases, the correlation coefficient decreases asymptotically to $\rho(\infty) = 0$. In the covariogram (Fig. 20.34b) $C(0)$ equals the variance of the property, σ^2, according to eq. (20.14). When the distance increases, the covariance decreases to $C(\infty) = 0$. In the semivariogram or more shortly the variogram the semivariance is plotted as a function of $|\mathbf{h}|$ (Fig. 20.34c). The 'semi' in the word semivariogram comes from the factor 1/2 in eq. (20.14). It is often omitted and the term variogram is used instead. When the stationarity condition holds it can be shown that the relation between the covariance and the semivariance can be written as:

$$\gamma(\mathbf{h}) = \sigma^2 - C(\mathbf{h}) = C(0) - C(\mathbf{h}) \tag{20.15}$$

The value of $\gamma(0)$ equals 0. Although this is theoretically true, the value of $\hat{\gamma}$ for a lag value approaching zero, is not necessarily zero in experimentally obtained data, due to different sources of error. This effect is called the *nugget* effect and is shown in Fig. 20.35. When the distance increases the semivariance reaches an asymptotic

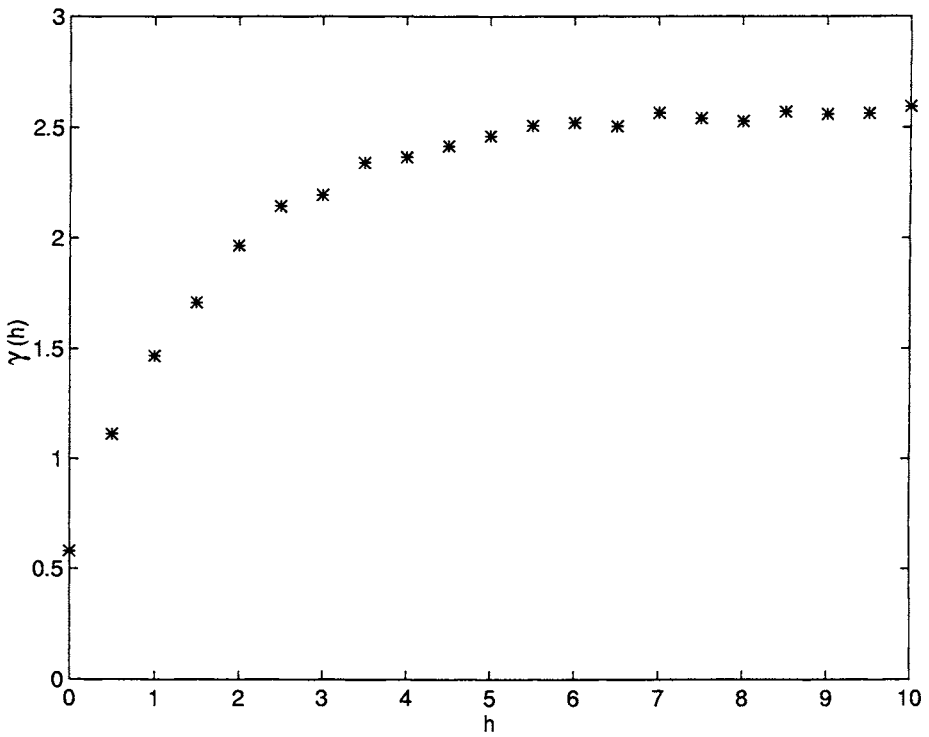


Fig. 20.35. An example of a variogram with a nugget effect. The nugget is the $\tau(0)$ value.

value, $\gamma(\infty)$, called the *sill*. According to eqs. (20.14 and 20.15), the value of $\gamma(\infty)$, the sill, equals σ^2,the variance of the property. The lag distance beyond which the value of the semivariance remains constant is called the *range*, *a*.

The variogram can be modelled in different ways. The models must however satisfy some restrictions because they will be used in the estimation step. Some often used models will be given in Section 20.10.5.

20.10.2 Interpolation methods

Estimation of the value of the property of interest at non-sampled locations requires a model about how the system behaves at the non-sampled locations. When the underlying mechanisms that cause the value of the property of interest are well understood, a deterministic model can be used. In geostatistics however this knowledge is rarely available. Most methods are based on a weighted linear combination of the available data:

$$\hat{v}_0 = \sum_{i=1}^{n} w_i v_i \tag{20.16}$$

with $\hat{v}_0$ the estimated property at a non-sampled location, v_i the values for the available sample points, and w_i the weights assigned to v_i.

When the stationarity assumption holds it is possible to derive the *unbiasedness condition* for the weights which is also intuitively acceptable [29]:

$$\sum_{i=1}^{n} w_i = 1$$

The unbiasedness condition is introduced to assure that the average prediction error, i.e. the difference between the predicted value and the true value, equals zero. The weights assigned to the data points may differ for estimates at different locations. Different approaches for assigning these weights give rise to different estimation methodologies. Basically two classes of methodologies can be distinguished: those that use criteria that are based solely on geometric information of the sample point locations (e.g. the inverse distance) and those that make use of the spatial continuity that is described in the previous section. The Kriging method (see Section 20.10.4) is the most popular of the latter category.

20.10.3 Interpolation methods using only location information

The simplest interpolation method is to use the value of the closest available sample point as an estimate for the value. This method is called the polygon method. It leads to discontinuities in the estimated values and this is usually not desirable. The *triangulation method* makes use of the three nearest neighbours. First the plane is calculated that fits the values of these three neighbours. The equation of the plane is then used to estimate the value at the unknown location.

Example:

A sampling scheme has been carried out on the points with coordinates as given in Table 20.4. The values of a property of interest, v, are also given in the table. Suppose that one wants to estimate $v(1.8,1.8)$, the value of v at the unsampled location with coordinates $x = 1.8$ and $y = 1.8$. The three nearest neighbours of this

TABLE 20.4

Example of triangulation interpolation method

Sample point	x	y	$v(x,y)$
1	1.0	1.0	26.2
2	1.0	2.0	39.2
3	2.0	1.0	27.4
4	2.0	2.0	40.4

point are the sampled locations 2, 3 and 4. The equation of the plane through these three points is given by $v(x,y) = 1.2\ x + 13\ y + 12$. From this equation it can be readily derived that $v(1.8,1.8) = 37.56$.

Other techniques use the information in more data points. The method of smoothing splines discussed in Chapter 11 belongs also to this category. Still other methods assign the weights according to the inverse distance to all data points, that are situated in a certain window around the unknown location.

$$\hat{v} = \frac{\sum_{i=1}^{n} \frac{1}{d_i} v_i}{\sum_{i=1}^{n} \frac{1}{d_i}}$$

d_i is the Euclidean distance between the ith sample location and the point where an estimate is wanted. The numerator in the equation is introduced to satisfy the unbiasedness condition of the weights. Variants of this method use a power of the inverse distance. In general when the data points are evenly distributed, the estimate improves when more data points are included. When one includes many points the danger of violating the underlying first order stationary model increases. It is indeed more likely that in a smaller window the stationary model is locally valid. Defining the search window for neighbourhood is one of the steps of the procedure.

20.10.4 Kriging

None of the previous methods makes use of the spatial continuity present in the dataset. Krige proposed a method, called *Kriging* that became the most popular method that does take information on spatial continuity into account [29,30]. Kriging is often called the BLUE method (Best Linear Unbiased Estimation). It is designed to minimize the variance of the estimation error (eq. (20.17)), by means of a weighted linear combination of the available sample points. The minimization of the error variance distinguishes Kriging from the previous methods. Three different variants can be distinguished: ordinary Kriging, block Kriging and co-Kriging. We will discuss only ordinary Kriging, since it is the most used variant.

The estimation error, e_0, at a non-sampled location is defined as:

$$e_0 = \hat{v}_0 - v_0$$

combined with eq. (20.16), this becomes:

$$e_0 = \sum_{i=1}^{n} w_i\, v_i - v_0$$

Using the properties of the variance of the weighted sum of variables the following equation for the error variance at the unsampled location $(\sigma_e^2)_0$ can be derived:

$$(\sigma_e^2)_0 = C_{00} + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j C_{ij} - 2 \sum_{i=1}^{n} w_i C_{i0} \qquad (20.17)$$

C_{ij} is the covariance between v_i and v_j. C_{i0} is the covariance between the ith sample and the unknown location and C_{00} is therefore the variance of the value at the non-sampled location. It is assumed to be the same as the variance at any location, σ^2 in eq. (20.15). The covariances for the different i- and j-locations must be established first. They can be derived from the variogram or covariance function, usually for different directions. To make this possible it is necessary to model the variogram or covariance function in order to be able to derive covariance values for all possible distances (see Section 20.10.5). The next step is to minimize the error variance with respect to the weights. Since the unbiasedness condition requires the sum of the weights to be one, a constrained minimization must be applied. This can be done with the Lagrange technique. It can be shown that this minimization procedure leads to the following equation:

$$\sum_{j=1}^{n} w_j C_{ij} + \lambda = C_{i0} \quad \text{for } i = 1 \ldots n$$

$$\sum_{i=1}^{n} w_i = 1$$

where λ is the Lagrange parameter. This system of equations is called the ordinary Kriging system. In matrix notation this equation becomes:

$$\mathbf{C} \qquad \mathbf{w} = \mathbf{C}_0$$

$$\begin{bmatrix} C_{11} & \ldots & C_{1n} & 1 \\ . & \ldots & . & . \\ . & \ldots & . & . \\ C_{n1} & \ldots & C_{nn} & 1 \\ 1 & \ldots & 1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ . \\ . \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} C_{10} \\ . \\ . \\ C_{n0} \\ 1 \end{bmatrix}$$

where $\mathbf{C}$ is the matrix containing the covariances, augmented with a column and a row of ones, $\mathbf{w}$ is the vector that contains the weights and λ, and $\mathbf{C_0}$ is the vector containing the covariances between the sampling points and the unsampled location.

Solving this equation for the weights gives:

$$\mathbf{w} = \mathbf{C}^{-1} \mathbf{C}_0$$

To obtain all coefficients of the $\mathbf{C}$-matrix, also those that are not measured, and the $\mathbf{C_0}$-vector the covariance function must be modelled. It is clear that the matrix

C must be non-singular. When the weights are obtained eq. (20.17) can be used to estimate the actual prediction error variance, $(\sigma_e^2)_0$.

20.10.5 Modelling the variogram

From the previous section it is clear that the covariances at non-sampled distances must be estimated in order to solve the Kriging system. These covariances can be obtained by modelling the covariogram. In this way any C_{ij} can be derived using the model at the appropriate **h** value. In geostatistics however it is common use to model the variogram instead of the covariogram. The use of eq. (20.15) immediately yields the conversion between the semivariance and the covariance:

$$C(\mathbf{h}) = \sigma^2 = \gamma(\mathbf{h}) = \gamma(\infty) - \gamma(\mathbf{h})$$

Thus one only has to subtract the value of the estimated semivariance at a lag **h** from the sill value of the variogram to obtain the corresponding estimated covariance value (see also Fig. 20.36).

In theory any model can be used to fit the variogram, but since the resulting **C** matrix must be non-singular it is common practice to use one of the basic models given below, which are assured to yield a non-singular **C** matrix

– *The linear model with sill (see Fig. 20.37a)*

$\gamma(|\mathbf{h}|) = c_0 + c/a|\mathbf{h}|$ for $|\mathbf{h}| < a$

$\gamma(|\mathbf{h}|) = b = c_0 + c$ for $|\mathbf{h}| \geq a$

c_0 is the nugget, b the sill, c is the difference between the nugget and the sill, a is the range and c/a the slope of the linear part.

– *The spherical model (Fig. 20.37b):*

$\gamma(|\mathbf{h}|) = c_0 + c[(1.5\ |\mathbf{h}|/a) - 0.5(|\mathbf{h}|/a)^3]$ for $|\mathbf{h}| < a$

$\gamma(|\mathbf{h}|) = c_0 + c$ for $|\mathbf{h}| \geq a$

The spherical model is the most commonly used variogram model in geostatistics.

– *The exponential model (Fig. 20.37c):*

$\gamma(|\mathbf{h}|) = c_0 + c[1 - \exp(-3\ |\mathbf{h}|/a)]$

This model reaches the sill asymptotically.

– *The Gaussian model (Fig. 20.37d):*

$\gamma(|\mathbf{h}|) = c_0 + c[1 - \exp(-3\ |\mathbf{h}|^2/a^2)]$

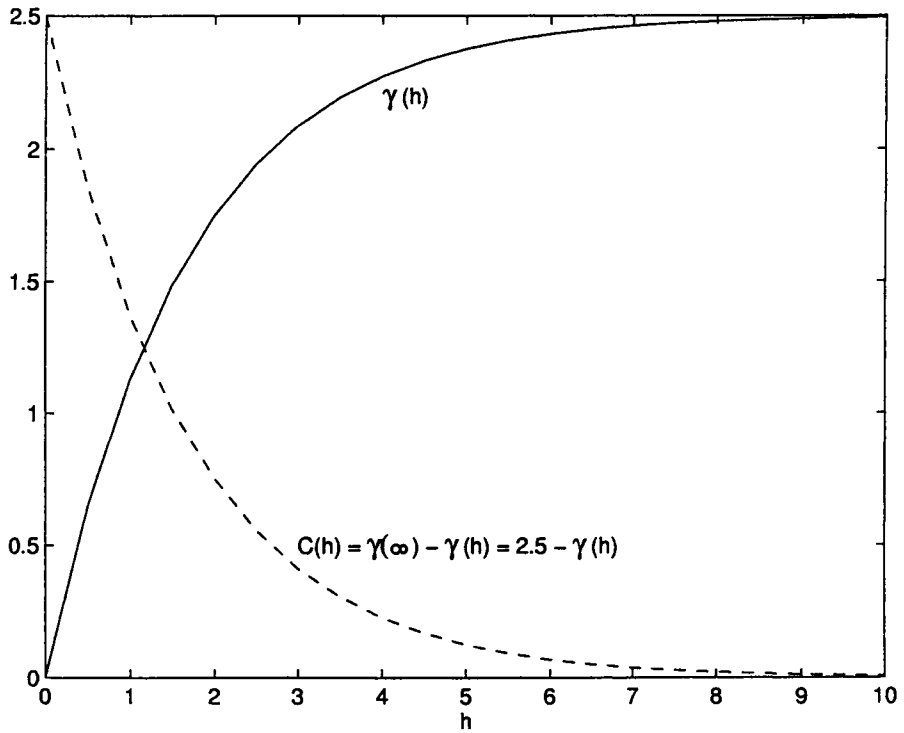


Fig. 20.36. Estimation of the covariance from the variogram. The solid line represents the modelled variogram; the dashed line is the covariogram derived from the variogram.

These are the basic models. More elaborate models have been developed which take into account anisotropic effects, i.e. the effect that the variogram is not the same in all directions. These models are however beyond the scope of this book (see Isaaks [29] for further details). Examples of the use of variograms can be found in [31,32].

20.10.6 Assessing the uncertainty of the prediction

Assessing the uncertainty of the predictions involves basically two aspects: judging whether the prediction method that is used is acceptable and deriving estimates of the uncertainty of the predicted value, e.g. a standard deviation. Evaluation of the prediction method that is used can be done using cross-validation techniques (see Chapter 10). The second aspect involves again the basic assumptions. The variance of the predictions, that are weighted sums of sample data points is given by eq. (20.17). We assume that the variance, C_{00}, is equal for all locations and equals σ^2, the variance of the property of interest. The first term thus accounts

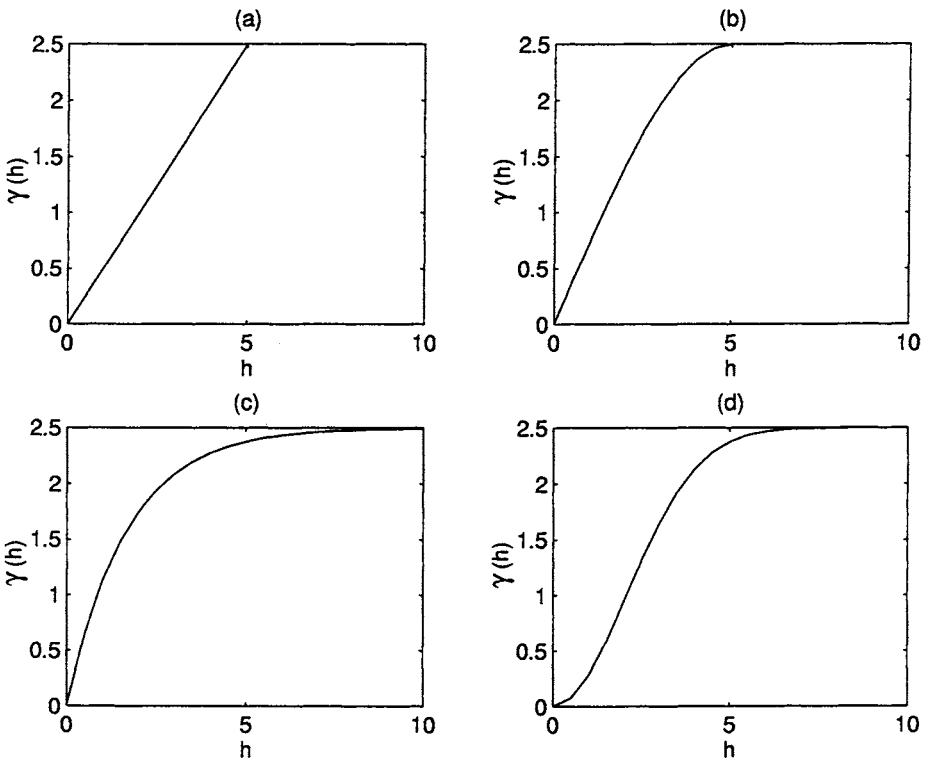


Fig. 20.37. Some commonly used transition models: (a) linear, (b) spherical, (c) exponential and (d) Gaussian.

for the errors present in the data. In eq. (20.17) C_{ij} in the second term is the covariance between v_i and v_j. It is a weighted sum of the covariances between the various sample pairs. It increases when the selected points i and j are close to each other, i.e they are clustered. This term thus takes the negative effect on the precision of clustered data points into account. A regular sampling scheme minimizes this term. C_{i0} in the last term of eq. (20.17) is the covariance between the ith sample and the unknown location. This term accounts for the spatial continuity of the data. As a consequence the precision increases when large covariances are present.

The estimate of the precision given by eq. (20.17) is valid for all prediction methods, based on a weighted sum of sample points. It can thus be understood that when the spatial continuity is not taken into account for the derivation of the weights (e.g. for the methods of the first category) the predictions are less precise.

Ordinary Kriging is in many aspects the preferred method, since it takes many aspects into account. It is however not always straightforward to apply. The basic problem is the estimation of the spatial continuity and selecting an appropriate

model to describe it. Since in practice the values of the semivariances for the available sample points are noisy this is a difficult task. This is certainly the case in environmental sampling problems. Several studies have been carried out in this field, e.g. by Einax et al. [35], by Domburg [31,33] and by Brus [34]. Wehrens et al. [32] described a sampling strategy that includes a preliminary sampling scheme from which the variogram is estimated. Using this variogram the final sampling scheme is determined based on the required precision of the estimates. The effort of doing additional sampling may be worthwhile, considering the influence of a correct model on the predictions. If from the predictions decisions must be taken on e.g. the parts of lakes that must be cleaned, then a decrease in prediction errors may yield a considerable saving. The extra sampling effort is in that case negligible compared to the overall cost. Moreover, when enough information is available on the spatial continuity it is possible to estimate beforehand the prediction error of the intended sampling scheme. In that way it is possible to avoid selecting sampling schemes that are too small or too large [32].

20.11 Acceptance sampling

An important issue in quality control is the acceptance or rejection of a lot or batch of products. As it is impossible to analyze all units of the batch one has to base ones decision on the inspection of a few units of the lot. A process which is under statistical control produces a certain constant rate of defectives. As a result each lot delivered to the customer will also contain a number of defectives. When the number of defectives in a lot is within the specifications, the level of nonconformances is considered to be acceptable. A sampling plan that will assure both the buyer and the producer of a well founded decision for delivery or acceptance is called *acceptance sampling*.

Acceptance sampling is generally used to protect against accepting lots that may have levels of nonconformances that are too large [36]. Therefore the percentage of defective (P_d) units in a lot is estimated by taking a sample of a number of units (n) and determining the number of units (n_d) of this sample which are defective. Let us consider a sampling plan where a sample of 5 units is taken from a lot which contains 5% defectives. The probability of finding no defectives (n_d =0) is given by the binomial distribution (Chapter 15) $0.95^5 * 0.05^0 = 0.77$. In general the probability ($P(nd)$) of finding n_d defectives in a sample of n units taken from a lot with a defective rate equal to P_d is given by

$$P(nd) = C_n^{n_d} (1 - P_d)^{n-n_d} (P_d)^{n_d}$$

with

$$C_n^{n_d} = \frac{n!}{n_d!(n - n_d)!}$$

The probability (P_a) of finding no more than an accepted number of defectives equal to n_c in a sample of n units taken from a batch with P_d defectives is therefore given by:

$$P_n = C_n^0 (1 - P_d)^n P_d^0 + C_n^1 (1 - P_d)^{n-1} P_d^1 + \ldots + C_n^{n_d} (1 - P_d)^{n-n_d} P_d^{n_d}$$

20.11.1 Operating characteristic curve

A curve, which relates the probability P_a of finding no more than an accepted number of defectives equal to n_c in a sample of n units taken from a lot, to its defective rate P_d is called the *operating characteristic curve* (OC). This term comes from statistical decision theory and was already introduced in a similar context in Chapter 4. Figure 20.38 shows a family of operating characteristic curves when the batch is accepted if no defectives are found in a sample of respectively 1, 2, 4 and 10 units. From that figure one observes that under a sampling plan which includes acceptance of a lot only when no units out of ten are found defective, the consumer or buyer still has about a chance of 43% of accepting lots with 8% defectives instead of zero defectives as he would wish. On the other hand the producer runs the risk that when a true defective rate of 5% is acceptable to the consumer, he will under the same sampling plan reject such a lot with a probability of 25%! The true defective rate which is acceptable to the consumer is called the *acceptable quality level* (AQL). AQL is defined as the maximum percentage of defective units in a lot or batch, that can be considered satisfactory on the long run as a process average [36]. This means that a lot with a defective rate equal to the AQL should have a high probability (usually about 95%) of being accepted. Lots with a higher defective rate should have a very low probability (e.g. 5%) of acceptance. The poorest quality in an individual lot which has such a desired low probability of acceptance is called the *lot tolerance percent defective* (LTPD).

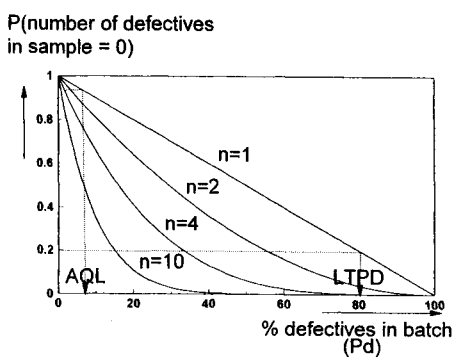


Fig. 20.38. Family of Operating Characteristic curves when a batch is accepted if no defectives (n_c = 0) are found in a sample of n units. AQL is the accepted quality level obtained for $P(n_d$ in sample = 0) = 0.95, LTPD is the lot tolerance percent defective obtained for $P(n_d$ in sample = 0) = 0.20.

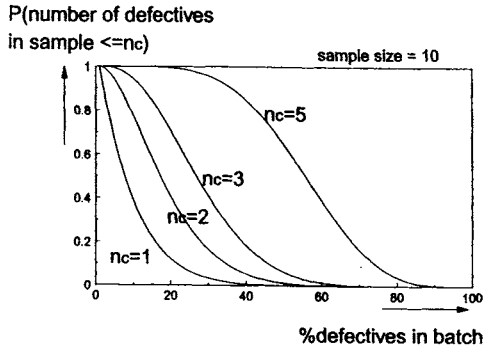


Fig. 20.39. Family of Operating Characteristic curves when n_c = 1, 2, 3, 5 defectives are accepted in a sample of 10 units.

The producer's risk and consumer's risk in acceptance sampling correspond with Type I (α) and Type II (β) errors in hypothesis testing (see Chapter 4). The producer's risk for any given sampling plan is the probability (α) of rejecting a lot that is within the AQL. Consider, for instance, the AQL = 2.2% situation. When one sample out of twenty is found to be defective on inspection, a lot with a true defective rate of 2.2%, which is the accepted quality level, has a probability of acceptance of 95% or, in other words the producer runs the risk that 5 times out of 100, a lot with an accepted quality level of 2.2% is rejected. On the other hand, the consumer has a 5% probability of accepting a lot with somewhat more than 20% true defectives. This consumer's risk for any given sampling plan is the probability (β) of accepting a lot which is below the LTPD. Both risks depend on the sampling plan which is completely described by the sample size, the acceptance number and the lot size. As can be seen from Fig. 20.38, the probability of acceptance of the lot changes considerably with the sample size. When sampling 10 units, the probability of acceptance of a lot with 10% defectives is 35% whereas for a sample of 2 units, 82% of such lots are accepted. All above conclusions are independent of the lot size itself. This is important to realize as this means that a large fraction of small lots has to be sampled to obtain the same probability of acceptance as for large lots. For instance when the OC curve for n = 10 and n_c = 0 is used, it means that 10% of a lot containing 100 units has to be inspected, whereas for a lot of 250 units only 4% of the lot should be inspected. The effect of the acceptance number (n_c) (see Fig. 20.39) for a given sample size is of about the same order of magnitude as the effect of the sample size.

20.11.2 Sequential sampling plans

When the tests are destructive or costly, one will not decide to sample large fractions of a lot. Instead a sequential approach is applied until the lot can be

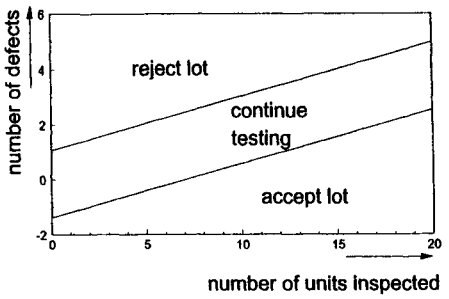


Fig. 20.40. Sequential sampling plan. $\alpha = 0.05$, $\beta = 0.1$, AQL is 5% defectives, RQL is 30% defectives.

accepted or should be rejected. A sequential sampling plan is based on the producer's risk (α), the consumer's risk (β), the acceptable quality level (AQL) and the rejectable (or unacceptable) quality level (RQL). It is then possible to design a graph in which we plot the number of nonconforming units detected versus the number of units inspected. This graph is divided in three zones (Fig. 20.40), a *rejection zone*, a continue testing zone and an *acceptance zone* The two lines which separate the zones have the same slope (s) which is given by:

$$s = \frac{\log[(1 - \text{AQL})/(1 - \text{RQL})]}{\log \dfrac{\text{RQL}(1 - \text{AQL})}{\text{AQL}(1 - \text{RQL})}}$$

The acceptance zone has a negative abscissa which is equal to:

$$\frac{\log \dfrac{(1 - \alpha)}{\beta}}{\log \dfrac{\text{RQL}(1 - \text{AQL})}{\text{AQL}(1 - \text{RQL})}}$$

The abscissa of the rejection zone is equal to:

$$\frac{\log \dfrac{(1 - \beta)}{\alpha}}{\log \dfrac{\text{RQL}(1 - \text{AQL})}{\text{AQL}(1 - \text{RQL})}}$$

Consider now a sampling plan that will have a 5% probability of rejecting a lot with 5% true defectives and a 10% probability of accepting a lot with 30% true defectives. In this case $\alpha = 0.05$, $\beta = 0.1$, AQL = 0.05 and RQL = 0.3.

Substituting these values in the above equations gives the following lines (see Fig. 20.40):

Reject line: number of nonconforming units = $0.1960\,n + 1.073$

Accept line: number of nonconforming units = $0.1960\,n - 1.378$

In this particular case we will accept the lot if e.g. no nonconforming units have been found in the first 8 units ($> 1.378/0.1960 = 7.03$). Otherwise we continue testing, unless the number of nonconforming units exceeds the reject line, which would already happen when the two first sampled units are nonconforming ($2 > 0.1960 * 2 + 1.073 = 1.465$).

References

1. P.M.E.M. van der Grinten, Finding optimum controller settings. Control Eng., 10 (1963) 51.
2. P.M.E.M. van der Grinten, Determining plant controllability. Control Eng., 10 (1963) 87.
3. P.M.E.M. van der Grinten, Control effects of instrument accuracy and measuring speed, I. J. Instr. Soc. Am., 12 (1965) 48.
4. P.M.E.M. van der Grinten, Control effects of instrument accuracy and measuring speed, II. J. Instr. Soc. Am., 13 (1966) 58.
5. F.A. Leemans, The selection of an optimum analytical method. Anal. Chem., 43 (1) (1971) 36A–49A.
6. G.E.P. Box and G.M. Jenkins, Time Series Analysis, Forecasting, and Control, 2nd Edn. Holden-Day, San Francisco, 1976
7. P.J.W.M. Müskens and W.G.J. Hensgens, Time series analysis on ammonia concentration and load values of the river Rhine. Water Res., 11 (1977) 509–575.
8. H.C. Smit and H.L. Walg, Baseline noise and detection limits in signal-integrating analytical methods, application to chromatography, Chromatographia, 8 (1975) 311–323.
9. M.B. Priestly, Spectral Analysis and Time Series. Springer-Verlag, New York, 1981.
10. T.P. Ryan, Statistical Methods for Quality Improvement. Wiley, New York, 1989.
11. D.C. Montgomery, Introduction to Statistical Quality Control, 2nd Edn. Wiley, New York, 1991.
12. G.B. Dijksterhuis and W.J. Heiser, The role of permutation tests in exploratory multivariate data analysis. Food Qual. Preference, 6 (1995) 263–270.
13. B. Skagerberg, J.F. MacGregor and C. Kisparisides, Multivariate data analysis applied to low-density polyethylene reactors. Chemom. Intell. Lab. Syst., 14 (1992) 341–356.
14. SAS Institute Inc., SAS/QC software: usage and reference, version 6, 1st edition, Vol. 2, Chapter 42, SAS Institute Inc., Cary, 1995.
15. S.J. Smith, G.P. Candill, J.L. Pirkle and D.L. Ashley, Composite multivariate quality control using a system of univariate, bivariate and multivariate quality control rules. Anal. Chem., 63 (1991) 1419–1425.
16. B.M. Wise, N.L. Ricker, D.F. Veltkamp and B.R. Kowalski, A theoretical basis for the use of principal component models for monitoring multivariate processes. Process Control Qual., 1 (1990) 41–51.
17. Th. Kourti and J.F. MacGregor, Process analysis, monitoring and diagnosis using multivariate

projection methods. Chemom. Intell. Lab. Syst., 28 (1995) 3–21.
18. W.G. Cochran, Sampling Techniques, 3rd Edn. Wiley, New York, 1977.
19. W. Horwits, Nomenclature for sampling in analytical chemistry. Pure Appl. Chem., 62 (1990) 1193–1208.
20. B. Baule and A. Benedetti-Pischler, Sampling of granular materials. Fres. Z. Anal. Chem., 74 (1928) 442–449.
21. J. Visman, A.J. Duncan and M. Lerner, Mater. Res. Stand., 8 (1971) 32–41.
22. J. Visman, A general theory of sampling. J. Mater., 7 (1972) 345–354.
23. P.M. Gy, The analytical and economic importance of correctness in sampling. Anal. Chim. Acta, 190 (1986) 13–23.
24. P.M. Gy, Sampling of Particulate Materials: Theory and Rractice. Elsevier, Amsterdam, 1982.
25. C.O. Ingamells and P. Switzer, A proposed sampling constant for use in geochemical analysis. Talanta, 20 (1973) 547–568.
26. C.O. Ingamells, New approaches to geochemical analysis and sampling. Talanta, 23 (1974) 141–155.
27. P. Minkkinen, Evaluation of the fundamental sampling of particulate solids. Anal. Chim. Acta, 196 (1987) 237–245.
28. P. Minkinnen, SAMPEX — A computer program for solving sampling problems. Chemom. Intell. Lab. Syst., 7 (1989) 189–194.
29. E.H. Isaaks and R.M. Srivastava, An Introduction to Applied Geostatistics. Oxford University Press, Oxford, 1989.
30. D.G. Krige, Lognormal–de Wijsian geostatistics for ore evaluation. South African Institute of Mining and Metallurgy Monograph Series, Geostatistics 1, Johannesburg, 1985.
31. P. Domburg, A knowledge-based system to assist in the design of soil survey schemes, PhD thesis at the DLO Winand Staring Centre for Integrated Land, Soil and Water Reserach (Sc-DLO), Wageningen, the Netherlands, 1994.
32. R. Wehrens, P.van Hoof, L. M.C. Buydens, G. Kateman, M. Vossen, W.H. Mulder and T. Bakker, Sampling of aquatic sediments, the design of a decision-support system and a case study. Anal. Chim. Acta, 271 (1993) 11–24.
33. P. Domburg, J.J. de Gruijter and D.J. Brus, A structured approach to designing soil survey schemes with prediction of sampling error from variograms. Geoderma, 62 (1994) 151–164.
34. D.J. Brus, J.J. de Gruijter, B.A. Marsman, R. Visschers, A.K. Bregt, A. Breeuwsma and J. Bourma, The performance of spatial Interpolation methods and chlorppleth maps to estimate properties at points: a soil survey case study. Environmetrics, 7 (1996) 1–16.
35. J. Einax, B. Machelett, S. Geiss and K. Danzer, Chemometric investigation on the representativity of soil sampling. Fres. Z. Anal. Chem., 342 (1992) 267–272.
36. T. Pyzdek and R.W. Berger Eds, Quality Engineering Handbook. Dekker, New York, 1992.

Recommended reading

G. Kateman and L. Buydens, Quality Control in Analytical Chemistry, 2nd Edn. Wiley, New York, 1993.
F.S. Budnick, Applied Mathematics for Business Economics and the Social Sciences, 2nd Edn. McGraw-Hill, New York, 1986.
E.G. Schilling, Acceptance Sampling in Quality Sontrol. Dekker, New York, 1982.
A.J. Duncan, Quality Control and Industrial Statistics, 5th Edn., Irwin, Homewood, IL, 1986.

Chapter 21

An Introduction to Experimental Design

21.1 Definition and terminology

This chapter introduces some terminology and concepts of experimental design. A more detailed description of the methods is given in Chapters 22 to 26. Many books and reviews have been written about experimental design; the books by Box, Hunter and Hunter [1] and Daniel [2] are classics. In the chemometrical literature there are books by Morgan and Deming [3], Carlson [4], Goupy [5,6] and Atkinson and Donev [7]. Some of the following chapters are, to a great extent, based on courses prepared by Phan-Tan-Luu [8]. Unfortunately, these courses are written in French and available only as course notes. A book [9] and a review [10] by Morgan provide simple introductions to the subject. Another useful introductory review is by Grize [11]. Many other books and reviews will be cited in Chapters 22–26.

The term *experimental design* is used in two contexts. The first is to describe the set of experiments that is carried out with the intention of developing a model, e.g. a regression model or an ANOVA model. We have seen in Chapters 8 and 10 that the selection of the experiments, i.e. the experimental design, has an influence on the quality of the model (e.g. on the precision with which the regression coefficients are determined). The term is also used in the context of optimization of products or processes: experimental design is applied to determine in an efficient way the set of conditions that are required to obtain a product or process with desirable, often optimal, characteristics. Chapters 22–26 are written essentially with the second context in mind. However, modelling is one of the main techniques applied to optimize, so that many of the concepts, e.g. the D-optimality principle of Chapter 24.4.1, can also be applied to obtain optimal models. Since the second context is the more important in what follows, let us analyze its definition.

We would like to determine a set of conditions, or values of factors. *Factor* is the name given in this field to variables that are changed in a controlled way to study their effect on the process or product and which have (or may have) an influence on the characteristics studied. The fact that the definition refers to a set of conditions means that one will nearly always be interested in several factors.

Typically, experimental design is multivariate in its approach. The characteristics of the product or processes to be optimized are often called the *response(s)*. They can be considered as variables describing the performance.

We can see that there are two types of variable: the responses and the factors; the responses are the dependent and the factors the independent variables (see Chapter 10). Responses will be symbolized by the letter y and factors by the letter x:

$$(y_1, y_2, \ldots, y_p) = f(x_1, x_2, \ldots, x_n) \tag{21.1}$$

In most of the cases described in the literature each of the responses are treated separately. Equation (21.1) then reduces to:

$$y = f(x_1, x_2, \ldots, x_n) \tag{21.2}$$

The *model* relating the response to the effect of the factors is called the *response function* or, because of its multivariate character, the *response surface*. Figure 21.1 shows some typical response surfaces.

The models are obtained from experiments. The word *design* means that these experiments are carried out not in a haphazard way, but in a carefully considered and *planned* way.

21.2 Aims of experimental design

Experimental design is used to obtain a product or process with desirable characteristics in an efficient way. This means one aims to

- understand the effect of the factors and/or
- model the relationship between y and x

with a minimum of experiments. This requires an orderly and efficient mapping of the experimental space. Experimental design, well applied, is therefore cost-saving.

In many applications one combines these aims (Fig. 21.2). One starts by determining which factors influence the response and to what extent. The next step often is to obtain a model describing in a quantitative way the effect of these factors on the responses. Eventually, one wants to define the optimum settings of the factor levels, i.e. the combination of factor values yielding the best characteristics of the product, process or procedure investigated. One then uses experimental design to optimize responses. The optimum may for instance be the highest or lowest value of the responses, but, as we will show later, there are other possibilities. It is also possible that one is not interested in the optimal result, but in a region where the results are sufficiently good.

Applications of experimental design are found in many areas of the chemical or related sciences. In analytical chemistry one could maximize for instance the absorption of a colorimetric procedure with as factors the amount of reagent, pH

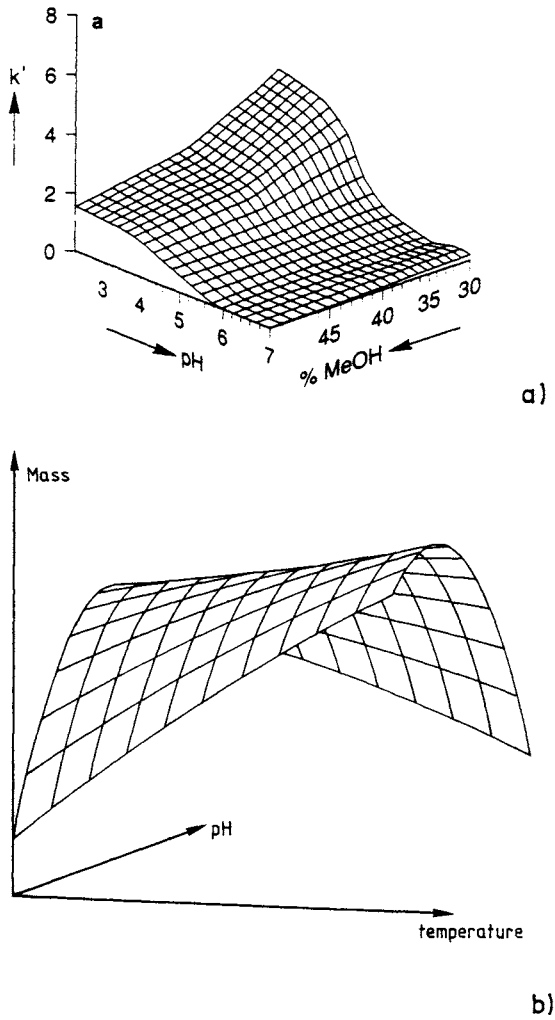


Fig. 21.1. Typical response surfaces. (a) Retention (k') as a function of pH and % methanol [12]; (b) mass of ribonuclease fixed to a silica support as a function of pH and temperature (adapted from Ref. [13]).

and type of buffer and, in organic chemistry, the yield of organic syntheses while minimizing the amount of certain byproducts. In food science one could optimize the sensory characteristics of food products according to their composition; in pharmaceutical technology one minimizes the friability of a granulate as a function of composition. In industrial chemistry one may optimize the rheological properties of a plastic as a function of factors related to the preparation process or the smoothness and soil release of a bedsheet in function of the type and

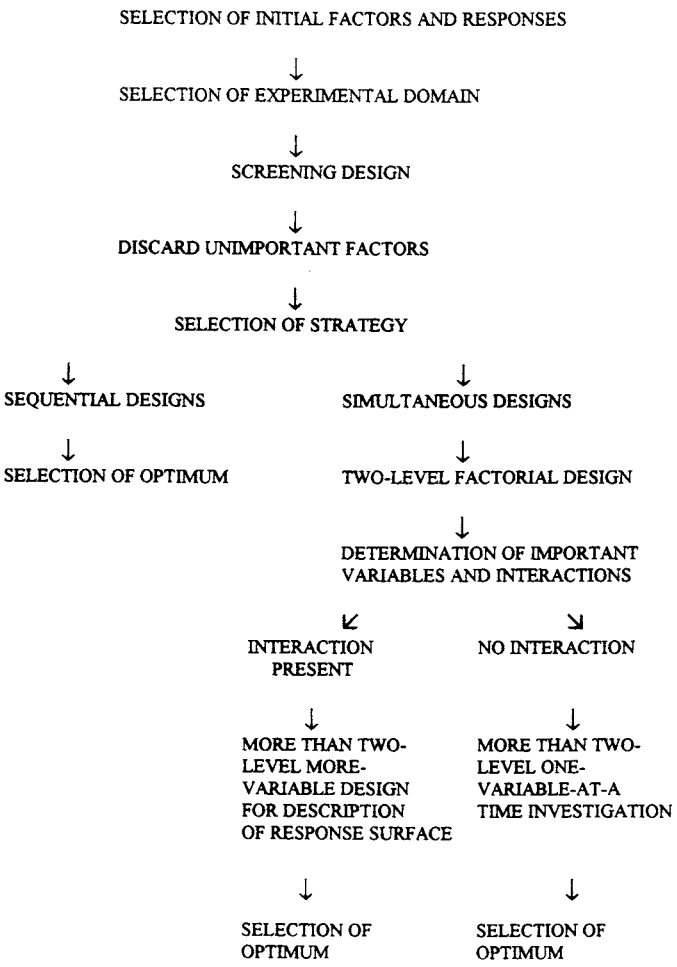


Fig. 21.2. The main steps in experimental design.

the amount of resin additives. These and other examples will be used to illustrate the following chapters.

As mentioned in the preceding section, experimental design is not only applied to obtain optimal responses but also optimal models. We have seen in Chapter 8 that the selection of the points to be used for the construction of a calibration line is important (Section 8.2.4.3) and allows us to optimize the quality of the prediction of concentrations. In Chapter 10.4 we have seen that the confidence intervals of the model parameters largely depend on the experimental design used. In Chapter 36 on multivariate calibration, we will see that typical experimental designs, such as those described in Chapters 24 and 25, are also applied to obtain multivariate calibration models.

21.3 The experimental factors

Factors in experimental design can be *qualitative* or *quantitative*. If one is interested to know whether the use of different catalysts or the type of a solvent has an effect on the yield obtained by a certain organic synthesis, then the factors are qualitative. If the factor is pH, then it is quantitative. The different values one gives to the factor are called *levels*. If the experimental design requires experiments at pH values of 5 and 9, then there are two levels of pH. The term level, which has a quantitative connotation in everyday use (high or low level) is also used for qualitative factors. When investigating the effect of a solvent on a certain process, one would for instance indicate that two levels were investigated, e.g. methanol and acetonitrile. Mixed situations are also possible. For instance, one can investigate the effect of solvent (qualitative) and pH (quantitative) on a chromatographic response.

The selection of the factors is generally the very first step in an experimental design application (see Fig. 21.2). Sometimes one knows which factors have an effect, but frequently one does not have this information. In this case, one starts by writing down all the factors that might have an effect and then carries out a screening. Screening designs can be applied for this purpose (see Section 21.7 and Chapter 23).

Once the variables have been selected, one needs to define the boundaries of the *experimental domain*, i.e. the extreme levels at which the factors will be studied. The experimental domain is *bounded* by the levels taken by a certain factor. Consider the simplest possible experimental design: a response is measured at two levels of one factor (see Fig. 21.3a). This design defines a one-dimensional space bounded by the levels at which the experiments are carried out. A two-dimensional design is shown in Fig. 21.3b. It is a two-level factorial design in two dimensions.

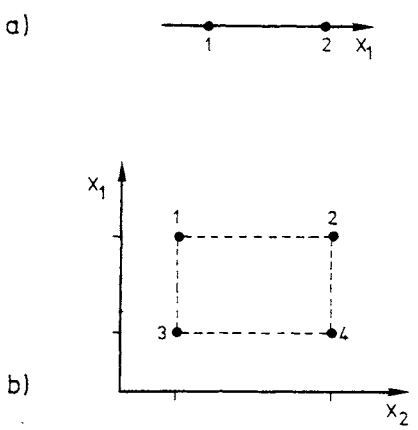


Fig. 21.3. Mapping the multivariate space. (a) One-dimensional space bounded by two levels of the factor; (b) two-dimensional rectangular space bounded by the two levels of two factors. (*Continued on next page.*)

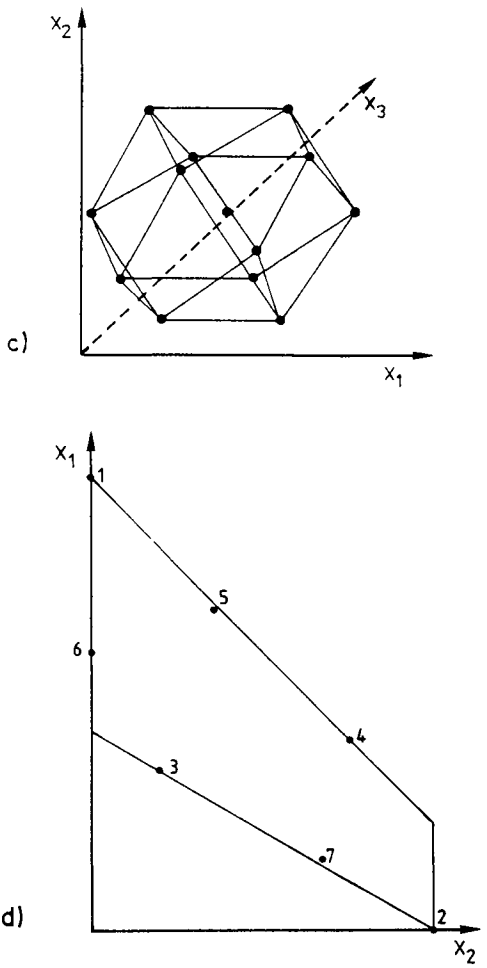


Fig 21.3 *continued.* (c) Three-dimensional spherical space bounded by 12 of the 13 experimental points of the Doehlert design; (d) an irregular experimental domain for the optimization of pH and % methanol of the HPLC separation of chlorophenols [12].

Experiments are carried out at two levels for each of the two factors. This type of design will be discussed in more detail in Chapter 22. One notes that 4 experiments are carried out. Together they define a rectangular two-dimensional space, while the 13 experiments of the Doehlert design (Fig. 21.3c, discussed further in Chapter 24) define a three-dimensional spherical space. The examples define symmetrical regions. However, as will be shown for the D-optimal designs of Chapter 24 and the mixture designs of Chapter 25, this is not always possible or even recommended. Figure 21.3d gives as an example the irregular experimental domain for the optimization of a separation of some dichlorophenols [12].

A correct definition of the boundaries and the levels is necessary. Boundaries that are too wide often require us to carry out experiments in useless conditions and lead to insufficient precision of the model in the area investigated. Boundaries that are too narrow can miss the optimum. Aids to defining boundaries for useful experiments exist in some domains; for instance, gradient elution methods for chromatographic optimization [14,15] allow selecting of boundary conditions at which solutes are eluted in an acceptable retention domain.

21.4 Selection of responses

A very important step in experimental design is the selection of the response(s) to be investigated. Usually, one models each response separately and tries to find the factor values yielding the highest or lowest response.

In real life it is common that there is more than one response and that the results obtained are conflicting. In HPLC, separation usually becomes better when the retention increases. However, it also means that the time required for the analysis becomes longer and, for many chromatographers, time is an important criterion which should be minimized. One observes that quality of separation (response 1) may be opposed to analysis time required (response 2). One does not need to find the optima of the two responses separately, but rather an adequate compromise. Techniques for treating conflicting responses, *multicriteria methods*, are described in Chapter 26.

A frequent error is to try and model composite responses. Consider for instance Fig. 21.4. Figure 21.4b describes the separation coefficient $\alpha = k'_1/k'_2$ (for $k'_1 > k'_2$) or k'_2/k'_1 (for $k'_1 < k'_2$), where k'_1 and k'_2 measure the retention of two substances, as a function of solvent strength. Figure 21.4a does this for log k'. While Fig. 21.4a can be modelled by a (quadratic) function, the function for the composite criterion α can be modelled less easily. Therefore, if at all possible, one should model the basic responses (in this case retention, (log) k') and obtain in a second step the complex response function (here α) from the model(s) derived in the first step.

Optimization in its classical sense consists in finding the factor values for which the highest response (e.g., highest yield of main product) or lowest response (e.g., lowest yield of an undesirable byproduct) is obtained. However, this is not necessarily the best choice. Consider, for example, the response surface of Fig. 21.5. The highest point is situated on a narrow ridge. Small changes in the value of a factor will lead to an immediate change (and in this case, decrease) of the response. It may be much better to choose point B. The response is not as high, but it is much more robust. Small changes in the factors do not lead to big changes in the response. Methods that optimize at the same time the magnitude of a response and its robustness were first proposed by Taguchi and will be described in Section 26.5.

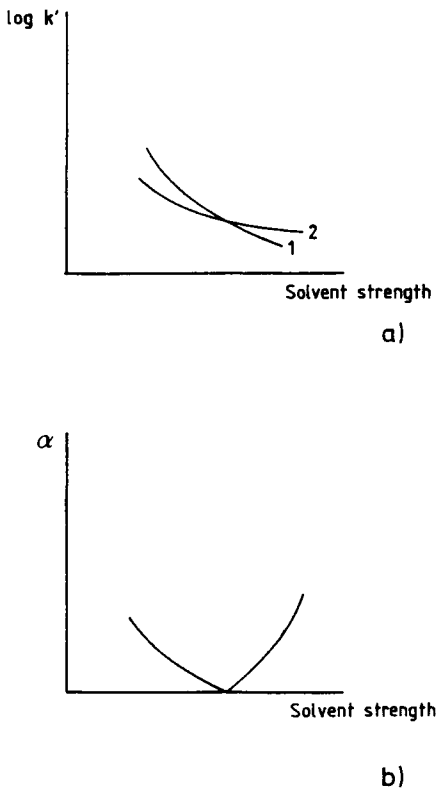


Fig. 21.4. (a) Capacity coefficient k' for two substances (1 and 2) as a function of solvent strength; (b) separation coefficient α as a function of solvent strength.

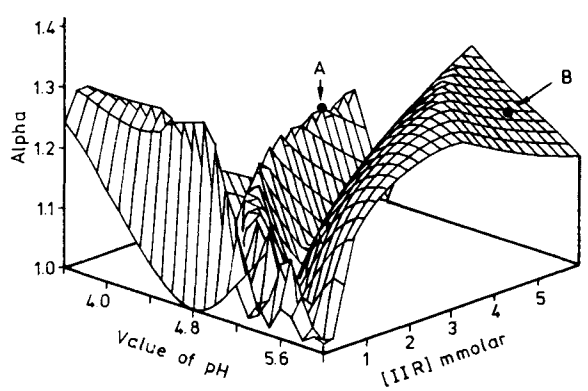


Fig. 21.5. Robustness. Point A has a high response but is not robust; point B has a response that is somewhat lower but is more robust (adapted from Ref. [16]).

21.5 Optimization strategies

For simplicity, we will suppose in the following sections that it is the highest value of the response that is needed. This is given by the maximum in the response surface. Instead of using a response surface graph, one often uses a *contour plot*. This translates the response surface in the same way as a geographical map for a mountainous area. The *isoresponse lines* can be viewed as the contour lines in the map.

As stated earlier, experimental design is usually multivariate in nature. Let us first compare a multivariate approach with the classical one-factor-at-a-time (univariate) strategy. Consider for instance Fig. 21.6a. This is an optimization with two factors. The starting point is A and the optimum to be reached is O. As a first step, one would carry out a certain number of experiments to optimize factor x_1 at constant value of x_2 and obtain B. Subsequent optimization of x_2 at the value of B for x_1 would yield O, as desired. This would not be the case in Fig. 21.6b. The univariate approach would yield here C as the assumed optimum. However, clearly C is far from optimal. The optimum could have been obtained by repeating this procedure (i.e. again keeping x_2 constant and optimizing x_1 anew, which would yield D and so on), but this would be inefficient because of the large number of experiments required. If more than two factors had been involved the situation would have been even worse. The inefficiency of this procedure is due to interaction between factors x_1 and x_2. The existence of interaction is demonstrated in Fig. 21.6c. Suppose one wants to investigate the effect of x_2 at the levels a and b of x_1 and at the levels c and d of x_2. One would then measure the response at locations A, B, C and D. The difference in response between the experiment at locations A and B would indicate that, when x_2 increases, the response decrease. However, when we would do the same for experiments C and D, we would observe an increase when x_2 increases. In other words, the effect of x_2 on the response depends on the value of x_1: this is what was defined in Chapter 6 as an interaction. This situation would not occur with the response surface of Fig. 21.6a: there is no interaction in that case. We also conclude that the univariate optimization strategy is efficient when there is no interaction and that it is not when there is interaction. In practical cases, interactions occur more often than not, so that multivariate approaches are usually more efficient than univariate ones.

There are two main multivariate optimization strategies. These are often called *sequential* and *simultaneous* optimization strategies. Some mixed approaches can also be used. The simultaneous strategies entail carrying out a relatively large number of experiments according to a pre-arranged plan. The factorial and mixture designs described in Chapters 22 to 25 belong to this group of designs and, as they are the more important ones, we will introduce them further in Sections 6 and 7 of this chapter. The experimental results are used to obtain models, such as those described in eqs. (21.3 to 5) and from these models the optima, i.e. the values for which y is highest can be derived.

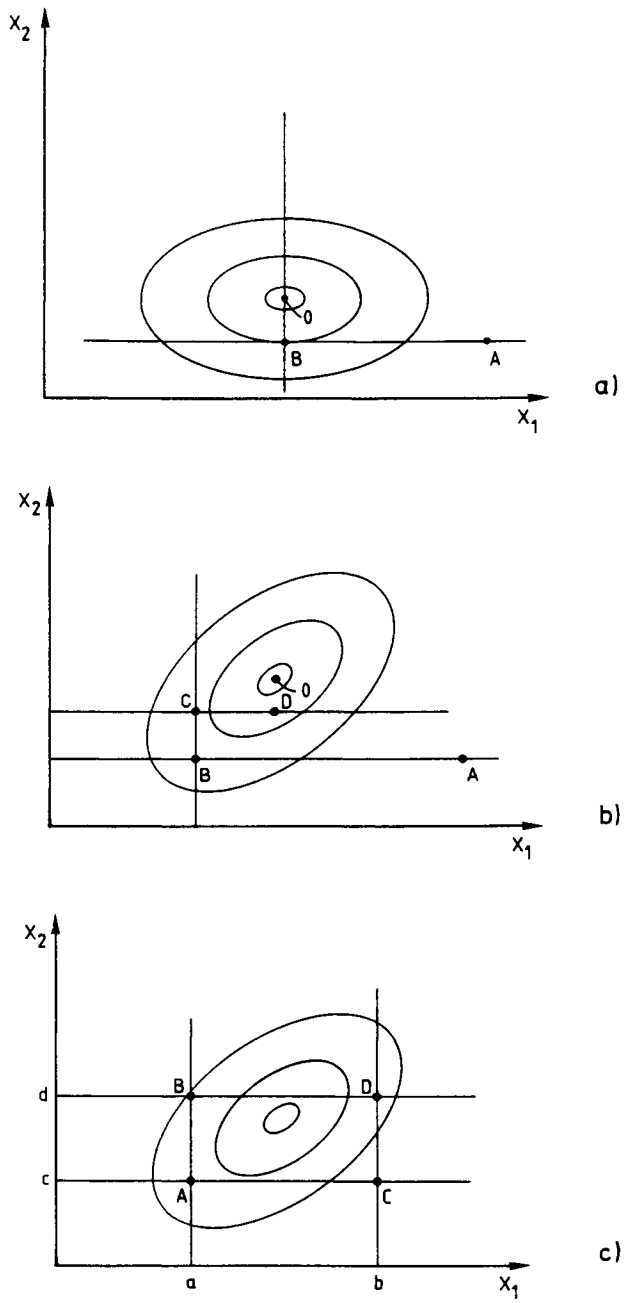


Fig. 21.6. Optimization of one factor at a time in a two factor experiment; (a) when there is no interaction; (b) when there is interaction; c) the interaction: the effect of x_2 is different at levels a and b of x_1.

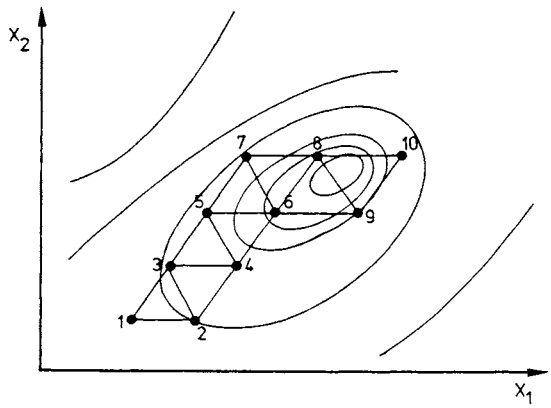


Fig. 21.7. Simplex optimization for two factors x_1 and x_2.

A sequential strategy consists in carrying out only a few experiments at a time and using these to determine the experiment to be carried out next. The best known sequential method is called the *Simplex* method. Figure 21.7 explains the principle of the simplest such method, the fixed size Simplex method, for the optimization of a response as a function of two factors, x_1 and x_2. In this example one would start with three experiments (1,2,3) situated in a triangle. Experiment 1 yields the worst result and, therefore, it is concluded that this point is situated in the wrong direction and that one should move in the opposite one. A new triangle is therefore constructed consisting of experiments 2,3 and the new one, 4. In this triangle the process is repeated: 2 is worst, it is replaced by 5. The Simplex methodology is described further in Chapter 26. Its application would have led to the selection of experiments 6 to 10, successively. The decision on the experiments to be carried out was made in a sequential way: only 1, 2 and 3 were decided simultaneously at the beginning.

A more complete description of the sequential and some mixed approaches is given in Chapter 26. Sequential designs are chosen when the optimum of a single response is the only information desired, i.e. when a model relating response and variables is not desired. Thanks to their hill-climbing algorithms, the sequential methods are usually very efficient in this respect. When not only the optimum is required, but also a model of the response surface, simultaneous methods would be preferable. Although the results obtained during the sequential design would allow us to map part of the response surface, that model would not necessarily be a good one. It can be shown that the quality of a model depends on the design chosen (see Chapters 10 and 24). In the sequential method the design depends on the route followed to reach the optimum and it would therefore be a matter of luck to have a good design. There are several situations where sequential methods are not

indicated. For instance, they cannot be used easily when there is more than one response, as will be explained in Chapter 26. For all these reasons, more attention is paid in this book to simultaneous than to sequential methods.

Experimental design is usually applied to multiple factor situations. In some cases, experimental design can be useful for the optimization of a single factor. Strategies for such situations also exist: one — *window programming* — is described in Chapter 26.

A special case of optimization is *numerical optimization*. This was applied in, e.g., Chapter 11 where hill-climbing methods such as steepest ascent and Simplex were used to estimate the coefficients in a non-linear equation. This is an optimization problem, since it requires to find the combination of coefficients that yields the minimum of the squared residuals to the model (least squares). In more complex cases, special techniques such as genetic algorithms or simulated annealing are required. These are described in Chapter 27.

21.6 Response functions: the model

Experimental design is used to develop empirical models. This means that it is used in situations in which one is not able to derive the response function from theory. In practice, this is nearly always the case. Theory often allows us to predict that there should be a relationship between a certain factor and a response, it sometimes permits us to derive what type of function (linear, quadratic, etc.) should be obtained, but is it rarely able to give the coefficients in that function.

Typical response functions are given below for two factors, x_1 and x_2, and the single response y:

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 \tag{21.3}$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2 \tag{21.4}$$

One notes that they consist of the following.

– A constant term b_0, which is the value of y when x_1 and x_2 are zero. Very often one works with *coded factors*. Suppose that one of the factors is a concentration of a reagent and that the two levels at which experiments are carried out are 0.1 M and 0.3 M. These are then coded, for instance, as –1 and +1. The 0 level is in between those –1 and +1 levels and is therefore the centroid. In our simple example the 0 level would be 0.2 M; b_0 then describes the value of y at that location.

– First-order (eq. (21.3)) and second-order (eq. (21.4)) terms for x_1 and x_2

– The last term in both eqs. (21.3) and (21.4) is the interaction term (see also Chapter 6).

This type of model is valid only for so-called *process variables*. The term is best understood in contrast with the term *mixture variables* or factors. Mixture factors are components of a mixture and are characterized by the fact that they add up to 1. This is not the case for process variables. Temperature, pH, type of machine used are typical process variables and the fractions of acetonitrile, methanol and water in a ternary mixture of those three solvents are mixture variables. It should be noted here that one uses *mixture designs* (Chapter 25) for mixture variables and *factorial designs* (Chapter 22 to 24) for process variables.

One notes that the model for process variables is second order: it contains squared terms and binary interactions. In principle, one could think of third and higher order polynomials, but this is rarely necessary. Ternary interactions are rarely relevant and third-order models or non-linear models (in the statistical sense of the term non-linear — see Chapter 10) do not often occur. In practice, nature can often be approximated, at least locally, by smooth functions such as second-order equations. Exceptions exist; for example, pH often leads to sigmoid curves when the measured response is due to the cumulated response of the ionized and the non-ionized species of the same substance. Responses that are bounded between 0 and 1, e.g., a sieve fraction in pharmaceutical technology, also often lead to sigmoid relationships that have a lower and upper plateau. Figure 21.8 gives an

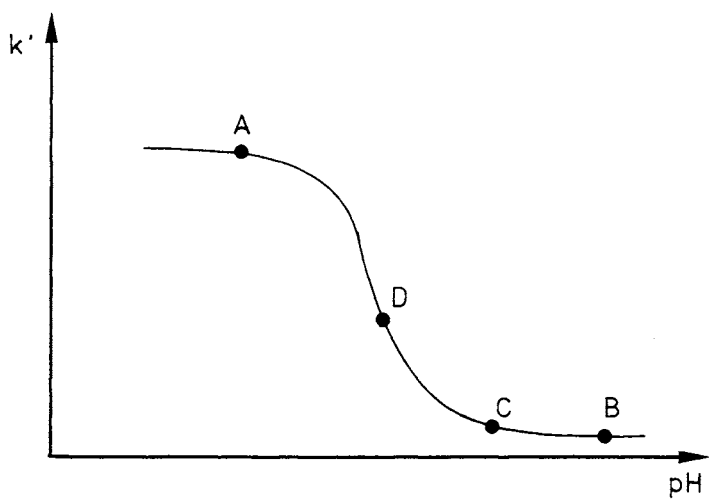


Fig. 21.8. Example of a response that would be difficult to map with a quadratic function. The response is k' as a function of pH in reversed phase partition chromatography. While it would not be possible to model the function over the region A–B, the more restricted region C–D can be approximated with a quadratic function.

example of a response that would be difficult to model over the whole experimental domain with a second order equation: as stated in Chapter 11.3.1, polynomial functions are unsuitable for fitting curves that have horizontal asymptotes (or plateaus). Quite often, one will not be interested in the whole domain, but in a more restricted region such as indicated in the same figure. In that case it becomes feasible again to model the response by a quadratic function of the independent variable.

The model for mixture variables does not include the quadratic terms for the individual factors such as x_1^2, nor the constant b_0. An example is given below for three factors

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 \; (+ \; b_{123}x_1x_2x_3) \qquad (21.5)$$

Although eq. (21.5) does not contain the squared terms, it is a second-order model. In Chapter 25, it is explained why the squared terms have disappeared. In contrast with process factors, mixture factors are sometimes modelled with cubic models. In the example of eq. (21.5) this would require us to add the term between brackets.

The models described by eqs. (21.3) to (21.5) are regression models. The b-coefficients are obtained by multiple regression, at least when the number of experiments is higher than or equal to the number of coefficients. When multiple regression is required the knowledge gained from Chapter 10 should be applied. For instance, one could:

- check whether all terms in the model are required using the techniques of Section 10.3.3;
- validate the model;
- try to obtain the most precise possible estimate of the coefficients; and
- try to obtain the smallest prediction error of the optimum or other regions of interest.

In some cases more advanced regression techniques may be useful such as PLS which can use the correlation structure between the variables or between the factors on the one hand and the responses on the other. This type of application will be discussed in Chapter 35.

21.7 An overview of simultaneous (factorial) designs

Simultaneous designs are often collectively called factorial designs. The following main classes of designs can be distinguished.

– Designs in which the emphasis is on detecting which factors have an influence or on estimating that influence. The basic design used here is the full *two-level factorial design*. In such a design one considers each factor at two levels and the

experiments are carried out at each possible combination of the two levels. These designs permit the estimation of the effect of all the factors and their interactions and the making of a first-order model including binary interactions (eq. (21.3)). The two levels are the boundaries of the experimental domain and because of the linear nature of eq. (21.3) the optimum response will necessarily be found along that boundary. Two-level factorial designs are described in Chapter 22.

When the number of factors increases, so does the number of experiments. For instance, for 10 variables, one would require 1024 experiments. In such cases one carries out only a fraction of these experiments, for instance 1/2, 1/4, 1/8. These designs are called *fractional factorial designs*. Because one carries out fewer experiments, one also loses some information about some or all of the interactions. This is described more in detail in Chapter 23.

In some cases, one is not at all interested in interactions. This typically occurs when the only aim is to determine which factors are relevant. When studying a new process, it can be the case that it is not known which of many possible factors affect the results. One wants to *screen* the candidate factors to select those that do. In such cases the smallest possible fraction of a two-level design (a *saturated fractional factorial design*) or the related *Plackett–Burman* designs can be used. They are described in Chapter 23.7. Such designs are also used to determine the collective influence of a large number of factors on the variance in the results of a process without necessarily trying to distinguish which factors have most effect. This is what is used in the determination of robustness (Chapter 13) or in Taguchi-type designs (Chapter 26).

– Designs in which the emphasis is on modelling. This requires that one is able to describe curved relationships and therefore one needs the second-order model of eq. (21.4). Therefore at least three levels of each factor have to be considered. Typical designs are the *central composite design* and the *Doehlert uniform network*. These designs are described in Chapter 24. The main reason for using them is to be able to derive models such as those given by eq. (21.4) and the corresponding response surfaces. Usually two-level designs will be applied first to decide which factors are important and only these important factors will be studied with more than two level designs.

There are two special cases. One is the case of mixture factors, as explained already in Section 6. These do not only yield specific models, but also require specific designs which are described in Chapter 25. The other special case is that in which process factors are used for which it is not possible to control the levels, because nature does. Suppose one wants to study the influence of functional groups situated at two locations of a lead molecule used as the substrate for some organic or biological reaction. The two factors are considered to be the hydrophobicities of the substituents at the two locations. It will not be possible to synthesize molecules with exactly the levels of hydrophobicity at the two sites required by, for instance,

the central composite design. Rather, one will have available or will be able to synthetize a certain number of molecules from which one will pick a few with convenient hydrophobicity values of the two substituents to derive the model describing reactivity in function of hydrophobicity. The best set of substituents is selected by a *D-optimal design* or a mapping design. This is described in Section 24.4.

References

1. G.E.P. Box, W.G. Hunter and J.S. Hunter, Statistics for Experimenters, An Introduction to Design, Data Analysis and Model Building. Wiley, New York, 1978.
2. C. Daniel, Applications of Statistics to Industrial Experimentation. Wiley, New York, 1976.
3. S.N. Deming and S.L. Morgan, Experimental Design: A Chemometric Approach, 2nd Edn. Elsevier, Amsterdam, 1993.
4. R. Carlson, Design and Optimization in Organic Synthesis. Elsevier, Amsterdam, 1982.
5. J.L. Goupy, La méthode des plans d'expériences. Optimisation du choix des essais et de l'interprétation des résultats. Dunod, Paris, 1988.
6. J.L. Goupy, Methods for Experimental Design, Principles and Applications for Physicists and Chemists. Elsevier, Amsterdam, 1993.
7. A.C. Atkinson and A.N. Donev, Optimum Experimental Designs. Clarendon Press, Oxford, 1992.
8. R. Phan-Tan-Luu, Course notes, Université d'Aix Marseille III, 1991.
9. E. Morgan, Chemometrics: Experimental Design. Wiley, Chichester, 1991.
10. E. Morgan, K.W. Burton and P.A. Church, Practical exploratory experimental designs. Chemom. Intell. Lab. Syst., 5 (1989) 283–302.
11. Y.L. Grize, A review of robust process design approaches. J. Chemom. 9 (1995) 239–262.
12. B. Bourguignon, P.F. de Aguiar, M.S. Khots and D.L. Massart, Optimization in irregularly shaped regions: pH and solvent strength in reversed-phase high-performance liquid chromatography separations. Anal. Chem., 66 (1994) 893–904.
13. J.L. Marty in ref.[8].
14. J.W. Dolan, D.C. Lommen and L.R. Snyder, Drylab computer simulation for high-performance liquid chromatographic method development. II Gradient elution. J. Chromatogr., 485 (1989) 91–112.
15. P.J. Schoenmakers, A. Bartha and H. Billiet, Gradient elution method for predicting isocratic conditions. J. Chromatogr., 550 (1991) 425–447.
16. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, Chemometrics: a Textbook. Elsevier, Amsterdam, 1988.

Chapter 22

Two-level Factorial Designs

22.1 Terminology: a pharmaceutical technology example

Full two-level factorial designs are carried out to determine whether certain factors or interactions between two or more factors have an effect on the response and to estimate the magnitude of that effect. This is the object of this chapter. These designs require an experiment to be carried out at all possible combinations of the two levels of each of the k factors considered. The experiments are sometimes called *runs* or *treatments*. The latter term comes from agronomy, a science in which much of the experimental design methodology was originally developed. One would for instance investigate the effect of applying phosphorus and nitrogen to the yield of a crop. In its simplest form this requires that the crop be treated with four different treatments (low P–low N, low P–high N, high P–low N, high P–high N).

Two-level two-factor experiments can be represented as shown in Fig. 22.1. Of course, more factors can be included. This is shown in Fig. 22.2 for three factors. The number of experiments is equal to 4 ($= 2^2$) in Fig. 22.1 and to 8 ($= 2^3$) in Fig. 22.2. In general, the number of experiments required is 2^k, where k is the number of factors. Consequently, a two-level k-factor design is called a (full) 2^k *factorial* design.

The levels can be represented in different ways. A much used method is to indicate one level as + (or + 1) and the other as – (or – 1). When the factors are quantitative the + level indicates the higher value, the – level the lower value and 0 then indicates the centre, the value in between. This 0 value will not be required in this section, but there is a use for it in Sections 22.6.3 and 22.9.3. The notation is also applied to qualitative factors. The + level is then not higher than the – level, but simply different from it and there is usually no 0 level. A commonly used method is also to indicate one level as 0 and the other as 1. The combinations of + and – or 0 and 1 identify an experiment. For instance, experiment + – means that factor A was at the + level and factor B at the – level. However, it is useful to identify them in a shorthand way, particularly in view of the application of Yates' method (see Section 22.3) and fractional factorials (see Chapter 23). This is explained by an example adapted from Malinowski and Smith [1]. They studied the effects of four spheronization process variables on tablet hardness. The design is therefore a 2^4 design and requires 16 experiments. The factors are given in Table 22.1. One

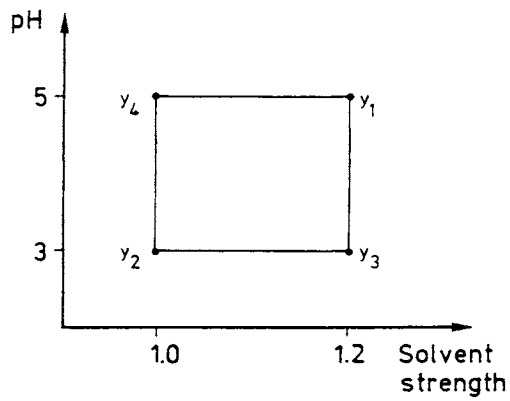


Fig. 22.1. A two-level two-factor design. The factors are solvent strength (levels 1.0 and 1.2) and pH (levels 3 and 5). Four experiments are carried out, yielding the responses y_1 to y_4 (retention of a chromatographed compound).

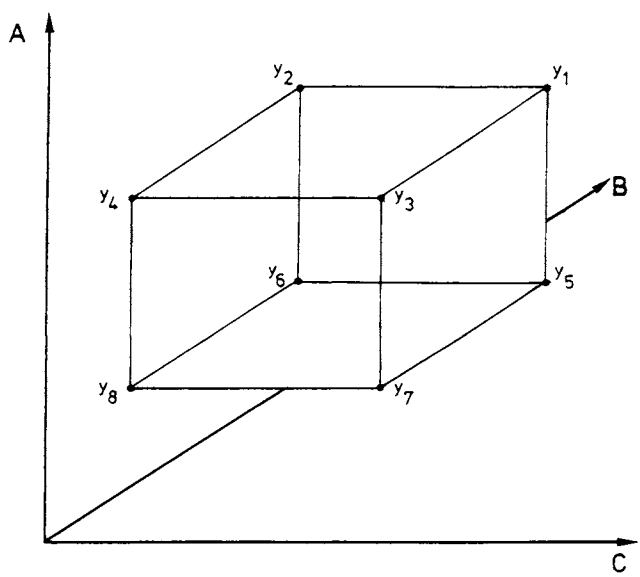


Fig. 22.2. A two-level three-factor design with factors A, B and C.

should note that, in keeping with the usage of the field, a factor is shown with a capital letter. The results are shown in Table 22.2. The experiments are shown in the order they were executed. The order is random to prevent artefacts in the conclusions (see further Section 22.9.2), so that it is not so easy to verify that indeed all combinations of + and – levels are represented. This is easier to verify in Table 22.3, which contains the same information, but grouped in another way.

TABLE 22.1

Factors and factor levels studied in connection with a spheronization process (adapted from Ref. [1])

Factor		Level	
		Low (–)	High (+)
Water content (ml)	(A)	250	325
Extruder speed (rpm)	(B)	39	59
Screen size (mm)	(C)	0.8	1.5
Spheronizer speed (rpm)	(D)	700	1010

TABLE 22.2

Experimental results for a full factorial design with the factors of Table 22.1. The response is tablet hardness.

Experiment	A	B	C	D	Response
a	+	–	–	–	4.2
ac	+	–	+	–	4.8
(1)	–	–	–	–	6.1
acd	+	–	+	+	3.7
b	–	+	–	–	6.4
d	–	–	–	+	4.7
abcd	+	+	+	+	3.7
ab	+	+	–	–	4.4
c	–	–	+	–	6.5
bcd	–	+	+	+	6.6
abd	+	+	–	+	3.4
ad	+	–	–	+	3.9
cd	–	–	+	+	6.7
abc	+	+	+	–	5.4
bc	–	+	+	–	8.3
bd	–	+	–	+	6.3

TABLE 22.3

The results of Table 22.2 presented in an ordered format

		A –				A +			
		B –		B +		B –		B +	
C–	D –	6.1	(1)	6.4	b	4.2	a	4.4	ab
	D +	4.7	d	6.3	bd	3.9	ad	3.4	abd
C+	D –	6.5	c	8.3	bc	4.8	ac	5.4	abc
	D +	6.7	cd	6.6	bcd	3.7	acd	3.7	abcd

In Tables 22.2 and 22.3 the experiments are identified with small letters. For instance, the first experiment in Table 22.2 is identified as a. This label is obtained by writing down in small letters all factors for which the level is + in that experiment. The level of A is + and it is – for B, C, D. The last experiment in Table 22.2 has positive levels for B and D and consequently is labelled bd. The experiment, where all levels are negative, is labelled (1).

22.2 Direct estimation of effects

We will return to the example of Tables 22.1 to 22.3 later, but to explain how to determine the effect of a factor we will first consider a somewhat simpler example. Suppose therefore that we want to estimate the effect of 3 factors (A, B and C) on a response y. The *design matrix* is given in Table 22.4. The order of these experiments clearly is not random. Randomization would be needed for practical applications, but for ease of explaining we will not take this into account.

Consider, for example, the effect of factor A. If one compares experiments 1 and 5, one observes that in both experiments the levels at which B and C are measured are the same but that A is once at level + and once at level –. The difference between the results y_1 and y_5 is therefore an estimate of the effect of A when B and C are both at the + level. The difference between the results y_2 and y_6 constitutes another estimate of the effect of A, this time at the + level for B and the – level for C. In total, four estimates for the effect of A can be obtained and the average effect of A can therefore be estimated as:

$$\text{Effect A} = [(y_1 - y_5) + (y_2 - y_6) + (y_3 - y_7) + (y_4 - y_8)]/4 \text{ or}$$
$$\text{Effect A} = (y_1 + y_2 + y_3 + y_4 - y_5 - y_6 - y_7 - y_8)/4 \qquad (22.1)$$

It is important here to note that one has averaged estimates. By performing more than one estimate statistical evaluation becomes possible (see further Sections 22.6

TABLE 22.4

Design matrix for a 2^3 factorial experiment

Run	A	B	C	Response
1	+	+	+	y_1
2	+	+	–	y_2
3	+	–	+	y_3
4	+	–	–	y_4
5	–	+	+	y_5
6	–	+	–	y_6
7	–	–	+	y_7
8	–	–	–	y_8

to 22.8). Moreover, the experiments are chosen so that they map the experimental domain in an efficient way. The average estimate is as representative for the whole domain as is possible without further knowledge of the process.

There is an easier way to write eq. (22.1). One merely has to sum the results of all the experiments carried out at the + level and to subtract those carried out at the – level of the factor considered.

$$\text{Effect} = (\sum \text{positive level runs} - \sum \text{negative level runs}) / 4 \tag{22.2}$$

In other words, the *main effect* of a factor is the difference between the average responses at the high level and the average responses at the low level. For instance, the effect of C is given by:

$$\text{Effect C} = (y_1 + y_3 + y_5 + y_7 - y_2 - y_4 - y_6 - y_8)/4 \tag{22.3}$$

Indeed the experiments 1, 3, 5 and 7 are carried out at the + level of C and the others at the – level. In general, for k factors, one can write:

$$\text{Effect} = (\sum \text{positive level runs} - \sum \text{negative level runs}) / (2^{k-1}) \tag{22.4}$$

or

Effect = mean of positive runs – mean of negative runs

When the effect is not described as the mean difference between + and – levels, but rather as the difference between one of these levels (sometimes called *extreme levels*) and the intermediate 0 level (sometimes called the *nominal level*) as is the case when one describes the results with regression models (see Section 22.8), then the effect as defined in eq. (22.4) must be divided by 2 or the nominator in eq. (22.4) must become 2^k instead of 2^{k-1}.

Equation (22.4) can be applied to any table, where the + and – levels are identified, as is the case in Table 22.3. Indeed one can estimate the effect of, for instance A, by taking all the results of experiments performed at a positive level for A minus the results obtained at the negative level and dividing this by 8.

$$\begin{aligned} A = [&(4.2 + 4.4 + 3.9 + 3.4 + 4.8 + 5.4 + 3.7 + 3.7) - \\ &(6.1 + 6.4 + 4.7 + 6.3 + 6.5 + 8.3 + 6.7 + 6.6)]/ 8 \\ = &- 2.26 \end{aligned}$$

In other words, one estimates that by changing the content of water from 250 ml to 325 ml the hardness decreases on average with 2.26 units.

The *interaction effects* can be estimated in exactly the same way, but some additional computations on the design matrix are required. Let us compare the evaluation of the effect of A from $y_1 - y_5$ and from $y_3 - y_7$ in Table 22.4. Both differences estimate the effect of A at the same (+) level of C. However the former

does this at the + level of B and the latter at the – level. By subtracting one from the other and dividing by 2, one estimates to what extent the effect of A is affected by this difference in B value. In other words, one estimates the interaction effect of B on A:

$$\text{Interaction of B on A} = [(y_1 - y_5) - (y_3 - y_7)]/2 = [(y_1 + y_7) - (y_3 + y_5)]/2.$$

The interaction effect of A on B can also be estimated. At the high level of C and A, the effect of B is estimated as $y_1 - y_3$ and at the same high level of C but at the low level of A by $y_5 - y_7$, so that:

$$\text{Interaction of A on B} = [(y_1 - y_3) - (y_5 - y_7)]/2 = [(y_1 + y_7) - (y_3 + y_5)]/2.$$

One finds that the estimates of the interactions of A on B and B on A are really the same, so that one can state that, at the high level of C, the interaction between A and B, written as AB or A × B is:

$$[(y_1 + y_7) - (y_3 + y_5)]/2$$

It can be verified that, at the lower value of C, a second estimate of the A × B interaction can be obtained by

$$[(y_2 + y_8) - (y_6 + y_4)]/2$$

Averaging the two estimates, one obtains:

$$A \times B = (y_1 + y_2 + y_7 + y_8 - y_3 - y_4 - y_5 - y_6)/4 \qquad (22.5)$$

In principle, it is not difficult to write down all the interactions in this way, but it requires attention and becomes very tedious when one needs to write down triple interactions or has more than three factors to consider. Fortunately, there is a much easier method, similar to the easy way of deriving main effects with eq. (22.4). How to do this is explained in Table 22.5.

TABLE 22.5

Computation of interaction levels for 3 variables

A	B	C	AB	AC	BC	ABC	Run
+	+	+	+	+	+	+	1
+	+	–	+	–	–	–	2
+	–	+	–	+	–	–	3
+	–	–	–	–	+	+	4
–	+	+	–	–	+	–	5
–	+	–	–	+	–	+	6
–	–	+	+	–	–	+	7
–	–	–	+	+	+	–	8

TABLE 22.6

Computation of interaction levels for 4 variables (only some interactions are given as an example)

A	B	C	D	AB	ABC	ABCD	Run
+	–	–	–	–	+	–	1
+	–	+	–	–	–	+	2
–	–	–	–	+	–	+	3
+	–	+	+	–	–	–	4
–	+	–	–	–	+	–	5
–	–	–	+	+	–	–	6
+	+	+	+	+	+	+	7
+	+	–	–	+	–	+	8
–	–	+	–	+	+	–	9
–	+	+	+	–	–	–	10
+	+	–	+	+	–	–	11
+	–	–	+	–	+	+	12
–	–	+	+	+	+	+	13
+	+	+	–	+	+	–	14
–	+	+	–	–	–	+	15
–	+	–	+	–	+	+	16

The three first columns of Table 22.5 are the same as those of Table 22.4. They constitute the design matrix; together they completely describe the experiment. The other columns (the interaction matrix) are needed only for computational purposes. They are obtained by simple multiplication. For instance the sign of the first row of AB = sign A × sign B = (+) × (+) = + and for row 6: (–) × (+) = –. In the same way the sign of row 8 for ABC is given by: (–) × (–) × (–) = –.

The computation of the interaction effect is now simple: one uses eq. (22.4). It can be verified that this leads to eq. (22.5) for the interaction AB. In Table 22.5 runs 1, 2, 7 and 8 have positive signs for this interaction and the other runs have negative signs. In Table 22.6 the computations are performed for some of the interactions that can be computed from Table 22.2. From Table 22.6 one concludes for instance that interaction ABC is computed as follows:

$$\text{Effect ABC} = [(y_1 + y_5 + y_7 + y_9 + y_{12} + y_{13} + y_{14} + y_{16}) - (y_2 + y_3 + y_4 + y_6 + y_8 + y_{10} + y_{11} + y_{15})]/8 = [(4.2 + 6.4 + 3.7 + 6.5 + 3.9 + 6.7 + 5.4 + 6.3) - (4.8 + 6.1 + 3.7 + 4.7 + 4.4 + 6.6 + 3.4 + 8.3)] / 8 = 0.1375$$

22.3 Yates' method of estimating effects

Although the calculation of effects as described in Section 22.2 is straightforward, it does require many manipulations and therefore a simpler computing scheme can be valuable for the user who does not have software available. Such a scheme was proposed by Yates [2]; it is explained in Table 22.7. It requires the

TABLE 22.7

Computation of the effects for the design of Table 22.2 with Yates' method

(1) Run	(2) y	(3)	(4)	(5)	(6)	(7) Effects	(8) Factor
(1)	6.1	10.3	21.1	46.1	85.1	5.32	T
a	4.2	10.8	25.0	39.0	−18.1	−2.26	A
b	6.4	11.3	18.3	−8.5	3.9	0.49	B
ab	4.4	13.7	20.7	−9.6	−3.3	−0.41	AB
c	.6.5	8.6	−3.9	2.9	6.3	0.79	C
ac	4.8	9.7	−4.6	1.0	−2.9	−0.36	AC
bc	8.3	10.4	−3.7	−1.3	0.7	0.09	BC
abc	5.4	10.3	−5.9	−2.0	1.1	0.14	ABC
d	4.7	−1.9	0.5	3.9	−7.1	−0.89	D
ad	3.9	−2.0	2.4	2.4	−1.1	−0.14	AD
bd	6.3	−1.7	1.1	−0.7	−1.9	−0.24	BD
abd	3.4	−2.9	−0.1	−2.2	−0.7	−0.09	ABD
cd	6.7	−0.8	−0.1	1.9	−1.5	−0.19	CD
acd	3.7	−2.9	−1.2	−1.2	−1.5	−0.19	ACD
bcd	6.6	−3.0	−2.1	−1.1	−3.1	−0.39	BCD
abcd	3.7	−2.9	0.1	2.2	3.3	0.41	ABCD

experiments to be first written down in what is called the *standard order*. This order is obtained as follows. One first writes down experiment (1), then a, b and ab. Then one adds experiments that include c by writing down all the experiments that were already included in the standard order and multiplying them by c. The fifth experiment is therefore (1) × c = c, then follow ac, bc and abc. The standard order for the experiments with d is then obtained by multiplying the 8 experiments already written in the required order and multiplying this by d. The ninth experiment is therefore (1) × d = d and the last one abc × d = abcd.

The experiments are written down in their standard order in column (1) and column (2) contains the responses. It must be stressed that this order is only used for the computation and not for the experimentation, where the order of the experiments must be randomized. Then for the computations one prepares a number of columns equal to the number of factors. For the example of Table 22.2 this requires 4 columns, i.e. columns (3) – (6). Column (3) is obtained from column (2) in the following way. One first adds the numbers in column (2) two by two and writes those down in column (3). In the example, one first adds 6.1 and 4.2 and writes down the resulting 10.3 in column (3). Then one sums 6.4 and 4.4 and writes down 10.8. This is continued until one has added together 6.6 and 3.7 and has written the resulting 10.3 in row 8 of column (3). One then subtracts the first number of column (2) from the second. The result (4.2 – 6.1 = −1.9) is used to continue filling up column (3). Then one does the same for the two following

numbers in column (2) (4.4 – 6.4 = – 2.0), etc. Once column (3) has been filled up, one repeats the whole operation to obtain column (4) from column (3), column (5) from column (4) and eventually column (6) from column (5). One can verify that this yields the (Σ positive level runs – Σ negative level runs) of eq. (22.4) for each of the effects. One then still needs to divide by 2^{k-1}, in this case by 8 (except for the first row, see further). The result is written down in column (7) and is the effect of the factor, with the same description as the experiment in that row. For instance the result in row 2 (–2.26) is the effect of A and the number in row 16 (0.41) is the effect of ABCD. These factors were written down in column (8). One notes that in row 1, one has obtained in column (6) the sum of all responses and in column (7) the average value. The symbol T is customary and is derived from total. The same table can be used to compute sums of squares as a first step towards analysis of variance of the data. This is described further in Section 22.7.

22.4 An example from analytical chemistry

A second example comes from analytical chemistry. It was described in a book published by the AOAC [3], which describes several applications of experimental design applied to analytical method development. It concerns the determination of acetone in cellulose acetate in water by distilling the acetone and determining it in the distillate after reaction with hydroxylamine hydrochloride by acid–base titration of the liberated HCl. One wanted to know whether the distillation could be avoided by disintegrating the cellulose acetate and performing the determination in that suspension. As factors one considered the components of the disintegration solvent (A–: acid, A+: basic, B–: water, B+: methanol) and the disintegration time (C–: 3 minutes, C+: 6 minutes). Each of the determinations was replicated. The results are shown in Table 22.8. The original distillation method was also carried out and yielded the result 4.10. This can be considered as the correct result. One would therefore probably conclude that it is worthwhile investigating further the A–, B–, C– experiment yielding an average value of 4.05 in Table 22.8 or the A–, B–, C+ (4.06) or A–; B+, C– (4.14) experiments. Experimental design is often used merely to map in an efficient way the experimental domain with the aim of selecting suitable conditions directly from the experimental results. In the present case, one might stop here. However, we will not do so and try to quantify the effects of the factors.

Yates' method applied to the totals in column y_T yields the results of Table 22.9. To obtain the effects in column (7) from the results in column (5), one divides by 16 (for T) or 8 because one used the sum of two observations in column (2) and the effects are obtained by $2^{k-1} = 4$ comparisons. We decided to use the sum of the two replicates, but it would have been possible to obtain the same result with the mean of the replicates instead. In this case, the divisors in column (6) would be 8 and 4.

TABLE 22.8

The determination of acetone in cellulose acetate: the analytical example (adapted from Ref. [2])

Run	y_1	y_2	y_T	$\bar{y}$
(1)	4.04	4.06	8.10	4.05
a	7.02	6.82	13.84	6.92
b	4.16	4.12	8.28	4.14
ab	5.68	5.80	11.48	5.74
c	4.08	4.04	8.12	4.06
ac	7.23	7.20	14.43	7.21
bc	4.26	4.20	8.46	4.23
abc	5.72	5.86	11.58	5.79

TABLE 22.9

Yates' method applied to y_T of Table 22.8

(1) Run	(2)	(3)	(4)	(5)	(6)	(7)	(8) Effect
(1)	8.10	21.94	41.70	84.29	16	5.27	T
a	13.84	19.76	42.59	18.37	8	2.30	A
b	8.28	22.55	8.94	−4.69	8	−0.59	B
ab	11.48	20.04	9.43	−5.73	8	−0.72	AB
c	8.12	5.74	−2.18	0.89	8	0.11	C
ac	14.43	3.20	−2.51	0.49	8	0.06	AC
bc	8.46	6.31	−2.54	−0.33	8	−0.04	BC
abc	11.58	3.12	−3.19	−0.65	8	−0.08	ABC

22.5 Significance of the estimated effects: visual interpretation

22.5.1 Factor plots

In Fig. 22.3 the results are given in a visual form for the analytical example and in Fig. 22.4 the results are given in a form similar for the pharmaceutical example. In these examples, one describes the value of the response at the different locations in multivariate space at which the experiments were carried out.

Figure 22.3 immediately shows that, for C, all + levels yield a higher result than their corresponding − levels (5.79 > 5.74; 7.21 > 6.92; 4.23 > 4.14; 4.06 > 4.05): there probably is an effect of C. The same is true for A and the effect is larger than for C (6.92 > 4.05; 5.74 > 4.14; 7.21 > 4.06; 5.79 > 4.23). A is clearly significant and its effect is large compared to C.

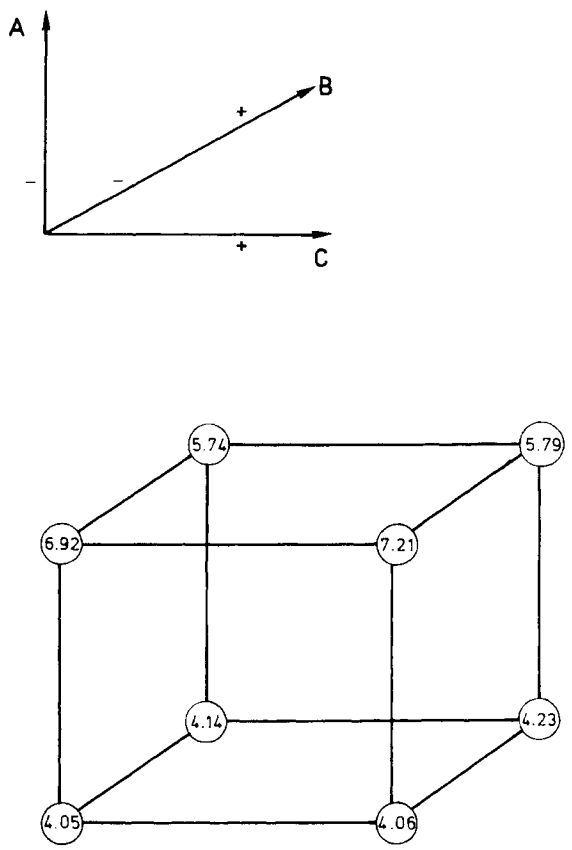


Fig. 22.3. Factor plot for the analytical example.

The interpretation is less evident for B. At the lower level of A, there might be a small effect since 4.14 > 4.05 and 4.23 > 4.06, but one might hesitate to make this conclusion on the basis of only two comparisons for this small effect. On the other hand, there is a large effect at the higher level of A. There the results are clearly higher at the B– level than at the B+ level. One concludes that there is an effect of B that depends on the level of A. This yields, in this case, both a significant effect of B and of the interaction AB.

It is not necessarily the case that one must find an effect of both main factors when the interaction is significant: if the effect of B at the lower level of A had the opposite direction and the same magnitude as that at the higher level of A, one would have found a significant effect of AB and none of B itself. Usually, however, at least one of the main factors is found to be significant.

This example also shows that when there is an interaction, one should not try to interpret a factor by its main effect alone. Conclusions about the effect of the

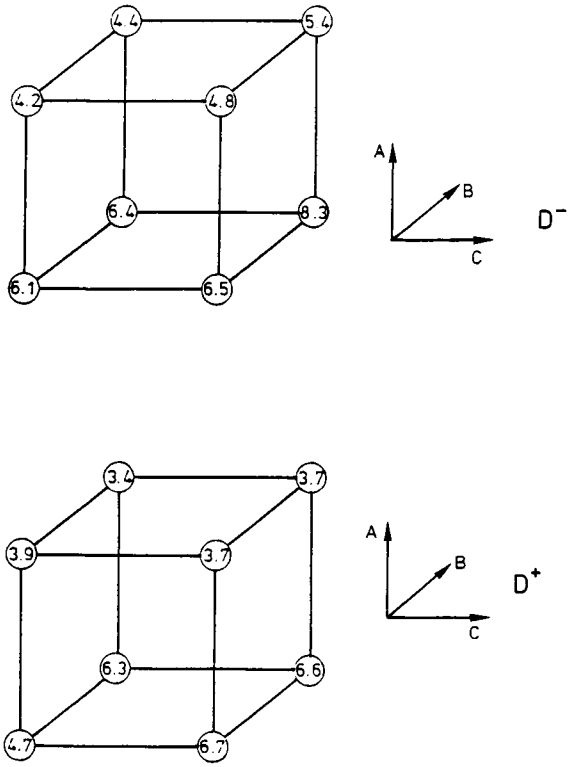


Fig. 22.4. Factor plot for the pharmaceutical example. The upper figure shows the responses in function of A, B and C at the D– level, the lower does the same at the D+ level.

solvent alone would be quite erroneous, as it would lead in this case to the conclusion that the solvent has an effect in all experiments, while clearly it has none or only a small one in acidic medium.

In the pharmaceutical example of Section 22.1 it is clear that there is a large influence of A. Its value at the lower (–) level is always higher than the corresponding one at the higher (+) level. There also seems to be an effect of C. In 7 out of 8 cases, the higher level yields a higher result than the lower level. For D too there is an effect as the (–) level (the upper cube) yields in 7 out of 8 cases a higher result than the (+) level. For B the situation is less clear. On the basis of the figure, one would probably not be able to come to a conclusion, but decide to carry out a more complete statistical analysis.

22.5.2 Normal probability plots

The effects obtained in Sections 22.2 to 22.4 are estimates. They are the average differences of 2^{k-1} pairs of observations and if the effects were not real, they would

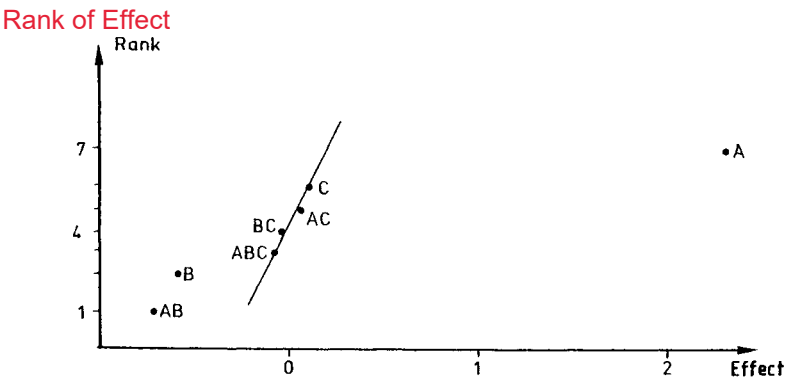


Fig. 22.5. Normal probability plot of the effects for the analytical example.

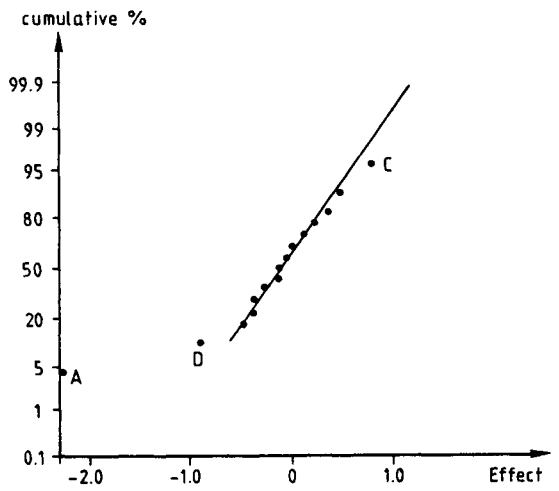


Fig. 22.6. Normal probability plot of the effects for the pharmaceutical example.

be dispersed around zero according to a normal distribution. Turning this argument around, the effects that are found to be part of this normal distribution cannot be considered to be significant. One ranks the effects from the most negative to the most positive and then proceeds to visualize the distribution, as described in Chapter 3.

In the analytical example the ranking is –0.72, –0.59, –0.08, –0.04, 0.06, 0.11, 2.30. The resulting normal probability plot is shown in Fig. 22.5. Very clearly, AB and B deviate at the negative end and A at the positive end of the straight line on which fall the points that are part of the normal distribution. These effects are therefore considered significant. The analytical example is based on observations at only 8 locations in space and one must therefore be careful about the conclusions made.

Figure 22.6 gives the results for the pharmaceutical example. At the negative end A (–2.26) is identified very clearly as an effect and D (– 0.89) too is considered significant. One would be hesitant, however, about C.

22.6 Significance of the estimated effects: by using the standard deviation of the effects

The computations in the preceding sections give an estimate of the effects. An estimate is subject to error and one can then wonder whether these estimates are statistically significant. In many practical examples, effects will either be so large that the experimenter will be able to decide that the observed effect does indicate that the factor in question is important in determining the response or it will be so small that he can decide he does not want to bother considering it further, even though it may be statistically significant. However, in many other cases this is not possible and the decision about whether a factor is really significant or not must be based on more statistical considerations.

The effects determined by the procedures described in the foregoing sections are estimates of the true effects. If one is able to determine the standard deviation of that estimate (the standard deviation of the effect), then one can also decide whether or not it is significant. One can apply a *t*-test to compare it to 0, using either the confidence interval approach or the critical level approach as described in Chapters 4 and 5.

One should be aware that there are two standard deviations that can be computed. The first is the standard deviation of an experimental measurement, i.e. the experimental uncertainty, and the other is the standard deviation of an effect. It is the latter one needs, but it can be computed from the former. In some situations we know the standard deviation of the experimental measurements or have an estimate of it. In the analytical example, for instance, one investigates effects on the basis of an existing analytical procedure and one might have determined its precision. If one does not know the standard deviation, one must estimate it. Three different ways to achieve this are described below.

22.6.1 Determination of the standard deviation of the effects by using duplicated experiments

In Chapter 2.1.4.4, we learned that the pooled variance of duplicated experiments is given by

$$s^2 = \sum d_i^2 / 2n \qquad (22.6)$$

where s^2 is the estimated variance of experimental error, d_i is the difference be-

tween the two duplicate experiments y_{i1} and y_{i2} and $n = 2^k$, the number of different experiments. The number of the degrees of freedom is also n.

The effects are computed from eq. (22.4), which, for duplicated experiments becomes:

Effect = ($\sum$ positive level runs – $\sum$ negative level runs) / (N/2)

where N is the total number of experiments (i.e. $N = 2\,n$), or

Effect = mean of positive level runs – mean of negative level runs

The variance of an effect, s^2_{effect}, equals

$s^2_{\text{effect}} = s^2$ for (mean positive levels – mean negative levels)

$$= s^2\left(\frac{1}{n}+\frac{1}{n}\right)=\frac{4s^2}{N} \qquad (22.7)$$

where s^2 is obtained from eq. (22.6).

From eq. (22.7) one can derive s_{effect} and therefore complete the table of effects by adding a standard deviation or else, one can compute the confidence limits at the 1 – α level by writing:

effect $\pm\, t_{\alpha/2;n} \cdot s_{\text{effect}}$

Let us apply this to the analytical example, where all experiments were duplicated. The d-values are obtained from Table 22.8. For instance $y_1 - y_2$ for the first experiment is $d_1 = 0.02$. Because $\sum d_i^2 = 0.0821$, it follows that $s^2 = 0.0821/16$ and $s^2_{\text{effect}} = 0.00513 \times (1/4) = 0.00128$ and $s_{\text{effect}} = 0.036$. The confidence limits are: Effect $\pm$ 2.30 × 0.036 = Effect $\pm$ 0.083 where 2.30 is the t-value for α = 0.05 and 18 degrees of freedom. For all effects with an absolute value larger than 0.083, the confidence interval will not include 0 and all such effects must considered significant. This is the case for A, B, AB and C. The others are not significantly different from 0.

22.6.2 Determination of the standard deviation of the effects by neglecting higher interactions

Although it is relatively easy to understand the physical significance of a two-factor interaction, it is often difficult to understand what a three-factor or still higher interaction means. For this reason the estimates of these interactions are often considered to be due to experimental error and one can derive from them an estimate of the standard deviation s_{effect}. This is applied in Table 22.10 for the pharmaceutical example. One first obtains the sum of squares by summing the squared effects and divides by the number of effects to obtain the variance of the effects.

TABLE 22.10

Estimation of the standard deviation of the effects from the higher interactions for the pharmaceutical example

Factor	Effect	(Effect)^2
ABC	0.14	0.0196
ABD	–0.09	0.0081
ACD	–0.19	0.0361
BCD	–0.39	0.1521
ABCD	0.41	0.1681

Sum of squares = 0.3840
$s^2_{effect} = 0.3840/5 = 0.0768$
$s_{effect} = 0.277$

One can use these standard deviations in combination with the critical value of a t-distribution, in this case with 5 degrees of freedom. At the level of 95% confidence, this then leads to the conclusion that all effects larger in absolute value than $2.57 \times 0.277 = 0.73$ are significant. This would be the case for A, D and C. The conclusion for C is not very clear since the estimate of the effect is 0.79. This explains why we were not able to draw a clear conclusion in Section 22.5.2.

Box et al. [4] state that when one has the choice, it is preferable to take an additional factor into account than to duplicate experiments, both leading to the same number of experiments. This also means that they advocate the procedure described in this section to determine the significance of effects. Of course, when higher interactions are really meaningful, then this procedure will lead to error.

22.6.3 Determination of the standard deviation of the effects by using the centre point

In some cases it may be useful to determine a high, a low and a medium level. This is possible only with quantitative factors and is useful particularly when the 0 level can be considered as a starting or nominal level. For instance, one has a certain process available, but wants to investigate what the effect of changing the level of factors is. The existing values for the factors are called the zero levels and one can define + and – levels at equal distances from this starting point (see Fig. 22.7). This starting point then becomes the centre point of the design. This centre point has several uses (see also Section 22.9.3) and by replicating it, one can obtain an estimate of the experimental uncertainty, which can then be used to obtain s_{effect}. This approach has the disadvantage that s is estimated only in the centre point but is considered to be an estimate for the whole experimental domain.

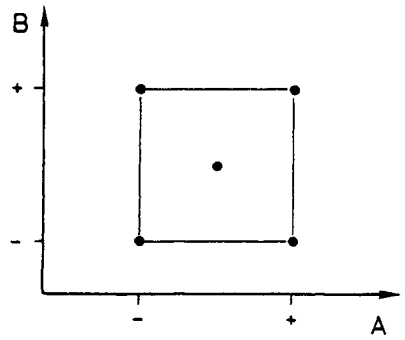


Fig. 22.7. A 2^2 design with centre point.

22.7 Significance of the estimated effects: by ANOVA

Since three factors are studied in the analytical example, acid/base (A), solvent (B) and disintegration time (C), the data can be rearranged as in Table 22.11. This is a three-way ANOVA with somewhat fewer replicates than we were used to in Chapter 6. In the terms of Chapter 6, it is a fixed effect ANOVA and the techniques described in that chapter can be used to determine the significance of each of the effects. This also includes *t*-tests, but attention should be paid to whether one decides to do this comparison-wise or experiment-wise. ANOVA, in fact, is the most evident way of carrying out the statistical analysis. This is done in Table 22.12 for the analytical example. There is one degree of freedom for each of the seven effects, since there are two levels for each of the factors. It is concluded that all main effects are significant and that this is also the case for the interaction AB.

Turning now to the pharmaceutical example, one realizes that this is a four-way ANOVA. There is, however, a problem, namely that there are no replicates and in

TABLE 22.11

A simple three factor duplicated design

		A −		A +	
		B −	B +	B −	B +
Factor C	−	replicate 1	replicate 1	replicate 1	replicate 1
		replicate 2	replicate 2	replicate 2	replicate 2
	+	replicate 1	replicate 1	replicate 1	replicate 1
		replicate 2	replicate 2	replicate 2	replicate 2

TABLE 22.12

ANOVA for the data of Table 22.8 (N/4) Effect^2

Effect	df	SS		MS		F
A	1	21.091		21.091		4136
B	1	1.375		1.375		269.6
A×B	1	2.052		2.052		402.3
C	1	0.050	/df=	0.050	/Res.=	9.8
A×C	1	0.015		0.015		2.94
B×C	1	0.007		0.007		1.37
A×B×C	1	0.026		0.026		5.18
Residual	8	0.041		0.0051		
$F_{(1,8)} = 5.32$						

TABLE 22.13

ANOVA for the data of Table 22.3. The column with header (6) is taken over from Table 22.7.

(N/4) Effect^2

Factor	(6)	SS	df		MS		F
A	−18.1	20.48	1		20.48		67.1
B	3.9	0.95	1		0.95		3.11
C	6.3	2.48	1		2.48		8.13
D	−7.1	3.15	1		3.15	/Res=	10.33
AB	−3.3	0.68	1		0.68		2.23
AC	−2.9	0.53	1		0.53		1.74
BC	0.7	0.031	1		0.031		–
AD	−1.1	0.075	1	/df=	0.075		–
BD	−1.9	0.225	1		0.225		–
CD	−1.5	0.14	1		0.14		–
ABC	1.1	0.076	1		0.076		
ABD	−0.7	0.031	1		0.031		
ACD	−1.5	0.14	1		0.14		
BCD	−3.1	0.60	1		0.60		
ABCD	3.3	0.68	1		0.68		
Triple + quadruple interactions		1.527	5		0.305		

Section 6.6 it was seen that this means that one cannot test the significance of all the interactions. This is very often the case in experimental design. In such a case one does what was already described in Section 22.6.2, namely one supposes that the higher interactions are not significant and one incorporates the sum of squares for these interactions into the residual sum of squares. This is shown in Table 22.13. One sums the relevant SS (0.076 + 0.031 + 0.14 + 0.60 + 0.68) and divides by the 5 degrees of freedom to obtain the MS. There is, of course, a problem with

this approach, i.e. that it occasionally happens that third-order interactions are significant. One will then overestimate residual variance and consequently it is possible that one will not detect some significant factor.

In the pharmaceutical example, the ANOVA leads to the conclusion that there is a significant effect of A, C and D. None of the interactions is significant. It should be noted that the sums of squares can be obtained, either by the use of appropriate ANOVA or experimental design software, or else from Yates' table. Indeed one can use the column containing the estimate of the effects not yet divided by the number of experiments. The sum of squares is obtained by squaring this result and dividing by the number of experiments represented. This was applied in Table 22.13. For instance, $(-18.1)^2/16 = 20.48$. For the analytical example too, one needs to divide by 16 because of the replication. The F-values are the ratios between the relevant MS and the MS comprising all triple and quadruple interactions. Only F-values higher than 1 are given. The critical values are: $F_{0.05;1,5} = 6.61$ and $F_{0.01;1,5} = 16.26$.

ANOVA can be recommended as the most convenient method of analyzing the significance of effects in the analysis of two-level factorial designs. When one does not replicate the experiments, it is good practice to determine also the normal probability plots of Section 22.5.2 (to avoid incorporating a significant triple interaction as explained higher). One should, in any case, perform the visual analysis of Section 22.5.1 and it is usually a good idea to determine also the centre point when this is relevant.

22.8 Least-squares modelling

For a 2^2 design, ANOVA leads to a linear fixed effects model of the type

$$y = \mu + a + b + (ab) + e \tag{22.8}$$

which we applied in Chapter 6.

The effects a, b and ab are associated with the main effects due to x_1, x_2 and the interaction between x_1 and x_2, respectively, so that one can also write the above relationship as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + e \tag{22.9}$$

This is a regression equation. In this equation, b_1 estimates β_1 and β_1 is a measure of the effect of x_1. In fact, it describes the effect on y when x_1 goes from 0 to + 1. *It should be noted that in this definition all effects are half those obtained according to the definition of an effect applied in the preceding sections.* Indeed, in those sections, an effect was defined as the difference of the value of y at the + 1 and – 1 level of a factor.

From the point of view of regression, the designs introduced in this chapter, present the advantage that the variables are not correlated (see Section 10.5). They are *orthogonal*. This means that the matrix $(\mathbf{X}^T\,\mathbf{X})$ is diagonal. This can be verified by substituting + 1 and − 1 for + and − in Table 22.4 and calling A, B and C, respectively x_1, x_2 and x_3. The correlations (and the covariances) between all x-pairs are 0. If we consider here only a simple model $y = b_1\,x_1 + b_2\,x_2 + b_3\,x_3$, the $(\mathbf{X}^T\,\mathbf{X})$ matrix is

$$\mathbf{X}^T\,\mathbf{X} = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 8 \end{bmatrix} = 8 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

or $\mathbf{X}^T\,\mathbf{X} = 8\,\mathbf{I}$

From eq. (10.18) it follows that

$$\mathbf{V}(b) = s_e^2\,(\mathbf{X}^T\,\mathbf{X})^{-1} = 1/8\;s_e^2\,\mathbf{I}$$

so that the covariance terms between the regression coefficients in $\mathbf{V}(b)$ are zero. In other words, the estimates for b_1, b_2, b_3 do not influence each other. In the same way, the confidence limits for the true regression parameters (eq. 10.15) are independent of each other, which means that tests on the significance of a factor can be carried out by testing each b-coefficient separately. The coefficients are considered significant if the confidence interval does not include zero. As we will show in Chapter 24, after introducing some more theory, all estimates for b-coefficients in the equation:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3$$

are uncorrelated, so that the design is indeed orthogonal.

The regression coefficients are usually computed for the scaled variables (i.e. on the scale −1 to +1). It is possible to re-express the coefficients as a function of the original variables. This is done as follows. Let us call the coded value x_i and the original value x_i^*. It can be verified that:

$$x_i = (x_i^* - x_{i,0}^*)\,/(x_{i,+1}^* - x_{i,0}^*) \tag{22.10}$$

where $x_{i,0}^*$ = value for x_i^* for which $x_i = 0$, i.e. the centre of the design, and $x_{i,+1}^*$ value for x_i^* for which $x_i = +1$, i.e. the extremum of the design.

To re-express the regression equation in the original variables, one substitutes eq. (22.10) in eq. (22.9). Suppose that it has been computed that:

$$y = 20 + 10\,x_1 + 30\,x_2 + 5\,x_1\,x_2$$

If the levels + 1 and − 1 for factor 1 are $x_1^* = 100$ and 80 and for factor 2, $x_2^* = 50$ and 40, then $x_{1,0}^* = 90$ and $x_{2,0}^* = 45$. Consequently:

$y = 20 + 10\,(x_1^* - 90)/\,10 + 30\,(x_2^* - 45)/5 + 5[(x_1^* - 90)/10]\,[(x_2^* - 45)/5]$ or

$y = 65 - 3.5\,x_1^* - 3\,x_2^* + 0.1\,x_1^*\,x_2^*$

It is now possible to obtain y for any value of x_1^* and x_2^*. Of course, the computed response is valid only for values of x_1^* and x_2^* between 80 and 100 and 40 and 50, respectively; one should not extrapolate outside the experimental region.

We should note here that the regression notation is the only one that will be used when more-than-two level designs will be described in Chapter 24. This notation might therefore be preferred in the present chapter and also in Chapter 23. However, as in the literature it is more usual to apply the ANOVA-approach of Section 22.7 to two-level designs, we preferred to conform to usage and apply it ourselves. It is important, however, to understand that two-level designs can be interpreted with linear regression models and that one can apply linear regression methodology to interpret the result. It is possible, for instance, to determine confidence intervals around the b-coefficients. The coefficients for which the confidence intervals include zero, indicate which variables are not significant. Using the regression approach is especially useful in some cases, for instance when one wants to predict y for intermediate values of some or all of the factors or when one is not able to apply exactly the levels – 1 and + 1 in all cases, as may happen in an industrial environment. An example of the latter is given by Goupy [5].

22.9 Artefacts

In the preceding sections we have illustrated full two-level factorial designs with two real examples that give good results. Experimental design and the interpretation of the results of factorial experiments are not always so straightforward. In this section, we will discuss some problem situations leading to artefacts and how to detect or avoid them.

22.9.1 Effect of aberrant values

An aberrant value leads to many artificially high interaction effects. To show this let us suppose that as mean value of experiment b in Table 22.8 one obtains 8.14 instead of 4.14. The result is that all calculated effects, where this experiment has a + sign in eq. (22.4) are increased by 1 and all effects where it has a – sign are decreased by 1. The following calculated effects are then obtained:

A: + 1.30, B: + 0.41, AB: – 1.72, C: – 0.89, AC: + 1.06, BC: – 1.04, ABC: + 0.92.

If we use the duplication of the experiments to estimate s_{effect} all effects will probably be found to be significant. When this happens and certainly when it happens with third and higher order interactions, one should suspect that an

aberrant value is present. Examples can be found in Sundberg's review [6], where it is also explained how to detect the aberrant value. For three variables there is only one higher interaction, so that one cannot apply the determination of s_{effect} based on neglecting higher interactions of Section 22.6.2. When there are more variables and one applies this method, aberrant values lead to higher values of the interactions, so that s_{effect} becomes larger and significant main effects may go undetected.

Another possibility, also illustrated by Sundberg, is that a few points were determined outside the normal range of operation, so that very different responses are obtained. These responses are not aberrant in the sense that they are wrong, but they are in the sense that they belong to a domain that should not have been investigated.

22.9.2 Blocking and randomization

Factorial experiments can be described as ANOVA experiments as shown in Section 22.7. Therefore the interpretation of such factorial experiments is affected by the same sources of error as described in Chapter 6.2. In particular, we have seen that blocking may be required and that randomization is usually needed. Blocking is described to some extent in the context of the description of Latin squares in Chapter 24. However, it cannot be described in sufficient detail in a general book on chemometrics. The reader should realize, however, that it is an important topic and consult the more specialized literature on experimental design, such as Refs. [1–9] of Chapter 21.

Let us consider the analytical example and suppose we cannot carry out all 8 duplicated experiments on the same day but carried out 4 duplicated experiments per day. As analytical chemists well know, there is a danger that a between-day effect occurs. Suppose that our measurements on day 1 are systematically higher by an amount d compared to day 2 and suppose that we carry out the experiments in the order given by Table 22.5. Then the effect of A would be overestimated by the amount d:

$$\text{Effect of A} = [y_1 (+d) + y_2 (+d) + y_3 (+d) + y_4 (+d) - y_5 - y_6 - y_7 - y_8]/4.$$

There would be no such error for the other effects as there are two experiments at the + level and two at the – level for all other effects. A better arrangement is shown in Table 22.14. As should become clear after reading Chapter 23, one has confounded the blocking effect with that of the interaction ABC, i.e. the average difference between the experiments in block 1 and 2 is also what one would compute for the interaction ABC.

Time and spatial order often lead to the presence of additional sources of variation. An interesting example, where the time effect does not occur between days or blocks, but within blocks can be found in Sundberg [6]. In that example, the first measurement within a block of four was always too low.

TABLE 22.14

A two-block arrangement for a 2^3 design

Run	A	B	C	
1	–	–	–	block (day) 1
2	+	+	–	
3	+	–	+	
4	–	+	+	
5	+	–	–	block (day) 2
6	–	+	–	
7	–	–	+	
8	+	+	+	

22.9.3 Curvature

In Section 22.6.3 it was explained that including a centre point is sometimes useful. It can be used also to obtain a better idea about the way an effect of a factor evolves over the experimental domain investigated. Suppose one obtains the results of Fig. 22.8. Clearly the behaviour of factor B is non-linear. On going from the – level (mean $y = 11$) to 0 ($y = 20$) there is a much larger increase than in going from 0 to the + level (mean $y = 21$).

If there was no curvature, then the centre point should be equal (within experimental uncertainty) to the average of the four corner points of the design. A 95% confidence interval for this difference is obtained by applying eq. (5.3):

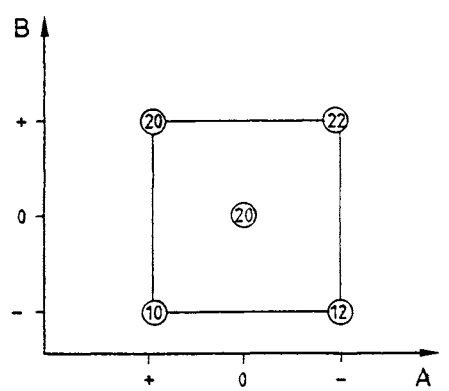


Fig. 22.8. Factor plot for a 2^2 design with centre point.

(mean of responses at +1 and −1 levels − response of centre point) ± 1.96 $\sigma \sqrt{\frac{1}{1}+\frac{1}{4}}$

where σ is the experimental standard deviation common to centre and corner points. If an estimate s was obtained, then 1.96 should be replaced by a t-value for the appropriate number of degrees of freedom. If zero falls in the confidence interval, then one concludes there is no curvature. Of course, replicating the centre point leads to a more precise conclusion.

One can also determine the significance of b_{11} and b_{22} in

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2 \tag{22.11}$$

We do not have enough experimental points to determine separately b_{11} and b_{22} but it has been shown [4] that:

$$b_{11} + b_{22} = \bar{y}_f - \bar{y}_c \tag{22.12}$$

where $\bar{y}_f$ = average of the corner points, and $\bar{y}_c$ = average of the centre points.

If $b_{11} + b_{22}$ is significantly different from zero, this is considered to mean that curvature exists. To determine whether significant curvature exists, one again needs to know the experimental error σ, or to obtain an estimate, s.

References

1. H.J. Malinowski and W.E. Smith, Effect of spheronization process variables on selected tablet properties. J. Pharm. Sci., 63 (1974) 285–288.
2. F. Yates, The Design and Analysis of Factorial Experiments. Imperial Bureau of Soil Science, Harpenden, UK, 1937.
3. G.T. Wernimont, Use of Statistics to Develop and Evaluate Analytical Methods, W. Spendley (Ed.). Association of Official Analytical Chemists, VA, 1985.
4. G.E.P. Box, W.G. Hunter and J.S. Hunter, Statistics for Experimenters, An Introduction to Design, Data Analysis and Model Building. Wiley, New York, 1978.
5. J. Goupy, Unconventional experimental designs. Theory and application. Chemom. Intell. Lab. Syst., 33 (1996) 3–16.
6. R. Sundberg, Interpretation of unreplicated two-level factorial experiments, by examples. Chemom. Intell. Lab. Syst., 24 (1994) 1–17.

Chapter 23

Fractional Factorial Designs

23.1 Need for fractional designs

The number of experiments in a full factorial design increases in an exponential manner with the number of factors, k. For $k = 2$, the number of experiments $n = 4$; for $k = 3$, $n = 8$; for $k = 4$, $n = 16$ and for $k = 7$, $n = 128$. The 128 experiments for 7 factors are used for the estimation of the mean value, 7 main effects and 21 two-factor, 35 three-factor, 35 four-factor, 21 five-factor, 7 six-factor, 1 seven-factor interactions. Three and more factor interactions are usually considered to be unimportant, so that the 128 experiments are used to determine the mean, 7 main effects and 21 binary interactions. Clearly there is a large redundancy and one might expect that it is possible to define smaller experimental designs.

To achieve this one takes one half, one quarter, one eighth, etc. of a full factorial design. The resulting designs are called *fractional designs*. They are symbolized by subtracting from the exponent in the 2^k design a number such that the resulting computation yields the number of experiments. For example, a 2^{4-1} fractional factorial design is a design for 4 factors. Of the 16 experiments required for the full factorial design only half are carried out, i.e. 8 experiments, $2^{4-1} = 8$. A 2^{7-4} fractional factorial is a design for 7 factors, consisting of 8 experiments (instead of the 128 required for a full design).

The designs must be balanced and chosen so that the experiments map the experimental domain as well as possible and orthogonality is preserved. A 2^{3-1} factorial design is shown in Fig. 23.1 and Table 23.1. It consists of half the experiments of the 2^3 design. The data concern an optimization study for the separation of fluoride and phosphate in capillary zone electrophoresis [1]. The authors wanted to know whether the resolution (R_S) between the two anions is affected by the pH (–: 8.0 and +: 9.0), the concentration of background electrolyte, BE (–: 0.008 M and +: 0.010 M) and the concentration of an electroendosmotic flow modifier, CEM (–: 0.0008 M and +: 0.0025 M).

The experiments in this design are located such that they describe a tetrahedron, which is the most efficient way of mapping the experimental domain with only 4 experiments. Each factor is twice at the + level and twice at the – level, so that a balanced design results. The authors concluded that resolution is indeed affected

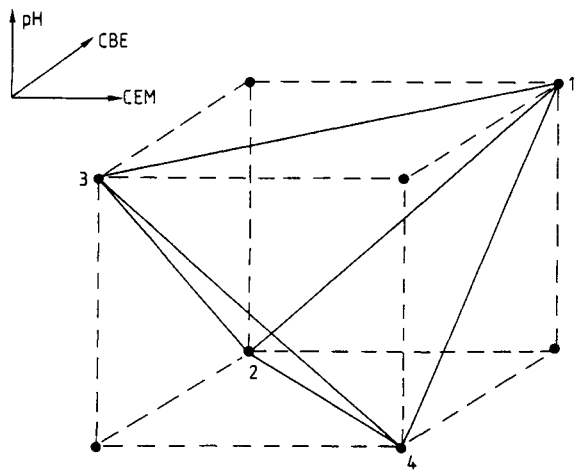


Fig. 23.1. A half-replica of a three factor design (2^{3-1} design).

TABLE 23.1

A 2^{3-1} experiment for determining effects on resolution in capillary zone electrophoresis [1]

Experiment	pH	CBE	CEM	R_S
1	+	+	+	2.75
2	–	+	–	2.53
3	+	–	–	1.95
4	–	–	+	2.36

by one or all of the factors, but that in all cases R_s was larger than the minimum required, so that they were able to stop their investigation at that stage.

To explain how fractional factorial designs are developed we will first consider a 2^{k-1} design. This is called a *half-fraction* or a *half-replica* (of a full factorial) *design*. Similarly a 2^{k-2} design is called a *quarter-fraction* or a *quarter-replica design*. The discussion of 2^{k-1} designs will be followed by 2^{k-p} designs and eventually by so-called saturated fractional factorials. The latter term will be defined in Section 23.7.1.

23.2 Confounding: example of a half-fraction factorial design

Table 23.2 describes a half-fraction design for the spheronization example described in Chapter 22. If we compare Table 22.3 with Table 23.2, we observe that indeed half of the original experiments were deleted and half remain.

TABLE 23.2

Half-replicate of the full factorial design of Table 22.3

		A–		A +	
		B–	B+	B–	B+
C–	D–	6.1 (1)			4.4 ab
	D+		6.3 bd	3.9 ad	
C+	D–		8.3 bc	4.8 ac	
	D+	6.7 cd			3.7 abcd

TABLE 23.3

A 2^4 design and some interactions. The first 8 rows are the experiments selected for a half-fraction design (2^{4-1})

A	B	C	D	BC	AD	ABC	BCD	ABCD	$\bar{y}$	Expt.	Resp.
+	+	+	+	+	+	+	+	+	+	abcd	y_1
+	+	–	–	–	–	–	+	+	+	ab	y_2
+	–	+	–	–	–	–	+	+	+	ac	y_3
+	–	–	+	+	+	+	+	+	+	ad	y_4
–	+	+	–	+	+	–	–	+	+	bc	y_5
–	+	–	+	–	–	+	–	+	+	bd	y_6
–	–	+	+	–	–	+	–	+	+	cd	y_7
–	–	–	–	+	+	–	–	+	+	(1)	y_8
+	+	+	–	+	–	+	–	–	+	abc	y_9
+	+	–	+	–	+	–	–	–	+	abd	y_{10}
+	–	+	+	–	+	–	–	–	+	acd	y_{11}
+	–	–	–	+	–	+	–	–	+	a	y_{12}
–	+	+	+	+	–	–	+	–	+	bcd	y_{13}
–	+	–	–	–	+	+	+	–	+	b	y_{14}
–	–	+	–	–	+	+	+	–	+	c	y_{15}
–	–	–	+	+	–	–	+	–	+	d	y_{16}

The eight experiments of Table 23.2 were of course not selected at random: they are spread out over the experimental domain so that they map it as efficiently as possible. In Section 23.3 we will describe how to perform this selection, but let us first investigate how to interpret the results. To do this we have reproduced a complete 2^4 design in Table 23.3. As already stressed in Chapter 22, experiments in a factorial design, whether a full or a fractional design, should be carried out in random order. The experiments are presented here in an ordered fashion to help understanding. The eight first rows are the experiments selected for the half-replica design of Table 23.2. It should be remembered that in Table 23.3 the experiments are determined completely by the columns A, B, C and D. The columns for the

interactions are obtained by the multiplication rules explained in Chapter 22 and are needed only for the computation of the interaction effects. To avoid making the table too large, only a few of all possible interactions are given (not included are AB, AC, BD, CD, ACD, ABD). We have also added a column for the mean response, $\bar{y}$. This is, of course, obtained by summing all responses and dividing by the number of experiments. To be able to compute it from the adapted eq. (22.4)

$$\bar{y} = [\Sigma(+ \text{ levels}) - \Sigma(- \text{ levels})] / (n/2) \tag{23.1}$$

which is used for the computation of all main and interaction effects, we filled in + signs for the whole column. Because of this formalism the mean response is also called the mean effect.

The interpretation of the half-fraction design is similar to that described in Chapter 22 for full factorials. For instance, the effect of A is computed as follows

$$\text{Effect A} = 1/4(y_1 + y_2 + y_3 + y_4 - y_5 - y_6 - y_7 - y_8) \tag{23.2a}$$

and, to give another example

$$\text{Effect BCD} = 1/4(y_1 + y_2 + y_3 + y_4 - y_5 - y_6 - y_7 - y_8) \tag{23.2b}$$

By comparing the two equations, we observe that we use the same equation for both effects, because they both have the same pattern of + and – signs in the first half of Table 23.3. To understand what exactly is computed with the right part of the eqs. (23.2a) and (23.2b) let us go back to the full factorial design and compute the effects of A and of BCD from that design:

$$\text{Effect A} = 1/8(y_1 + y_2 + y_3 + y_4 + y_9 + y_{10} + y_{11} + y_{12} - y_5 - y_6 - y_7 - y_8 - y_{13} - y_{14} - y_{15} - y_{16})$$

$$\text{Effect BCD} = 1/8(y_1 + y_2 + y_3 + y_4 + y_{13} + y_{14} + y_{15} + y_{16} - y_5 - y_6 - y_7 - y_8 - y_9 - y_{10} - y_{11} - y_{12})$$

and adding them:

$$\text{Effect (A + BCD)} = 1/4(y_1 + y_2 + y_3 + y_4 - y_5 - y_6 - y_7 - y_8) \tag{23.3}$$

What we computed to be the effects of A or of BCD from the half-fractional factorial with eqs. (23.2a and b) is now seen to be what we would have obtained for the sum of the effects A and BCD from the full factorial. We now also see that several other columns show the same pattern of + and –. We observe for instance that AD and BC on the one hand and ABC and D on the other hand have the same pattern. One says that ABC and D, AD and BC, A and BCD are *confounded*. It can be verified also that the mean effect is confounded with the ABCD interaction. One also says that the confounded effects are *aliases* of each other. In general, each effect in a 2^{k-1} design is confounded with one other effect or, to say it in another way, there are pairs of aliases. This should not surprise us: one cannot expect to

determine the mean effect, 4 main effects, 6 two-factor, 4 three-factor and one four-factor interaction with 8 experiments.

The practical meaning of this is that we no longer obtain estimates of single effects, but the sum of two. For instance

$$[(y_1 + y_2 + y_3 + y_4) - (y_5 + y_6 + y_7 + y_8)]/4 = \text{effect (A + BCD)}.$$

One usually assumes that the triple interaction is unimportant compared to the effect of the main factor and if the effect (A + BCD) were to be found important or significant, this would in a first instance be assigned to the effect of A. Of course, there is a danger that the assumption that BCD is much less important than A is wrong, but this is the price one pays for doing less experiments. Another consequence is that pairs of two-factor interactions are confounded and there is no simple way to decide whether a significant effect of (AD + BC) is due to AD, to BC or to both.

The calculations for the spheronization example of Table 23.2, performed here with the Yates algorithm, are given in Table 23.4. To apply this algorithm, it is necessary to view the 2^{4-1} design as derived from the 2^3 one. Why this is done so will become clear from the discussion of the generation of half-fraction designs. One writes down the experiments in standard order for a 2^3 design (see Section 22.3), i.e. without taking into account the D factor. The standard order is then (1), a, b, ab, c, ac, bc, abc. For the experiments carried out at the level D^+, i.e. experiments described by a letter combination including d, the d is then added without changing the order already obtained. One starts with (1). The second rank in the standard order is a. Since there is an experiment ad, one gives this the second rank, etc.

One observes that the main conclusion of the full factorial design (see Chapter 22), the large effect of A (+ BCD, but ascribed to A alone) is found again. The fact that B is unimportant is also observed. One would note that the effects D and C are larger than the effects for interactions and conclude that they may be significant. The normal probability plot is shown in Fig. 23.2. The effect of A is clear and the fact that B is unimportant too. One would hesitate about C and D.

TABLE 23.4

Interpretation of the half-fraction factorial design of Tables 23.2 and 23.3

Run	y	(3)	(4)	(5)	Effect	Factor
(1)	6.1	10.0	20.7	44.2	5.52	$\bar{y}$+ABCD
ad	3.9	10.7	23.5	−10.6	−2.65	A+BCD
bd	6.3	11.5	−4.1	1.2	0.3	B+ACD
ab	4.4	12.0	−6.5	−2.4	−0.6	AB+CD
cd	6.7	−2.2	0.7	2.8	0.7	C+ABD
ac	4.8	−1.9	0.5	−2.4	−0.6	AC+BD
bc	8.3	−1.9	0.3	−0.2	−0.05	BC+AD
abcd	3.7	−4.6	−2.7	−3.0	−0.75	D+ABC

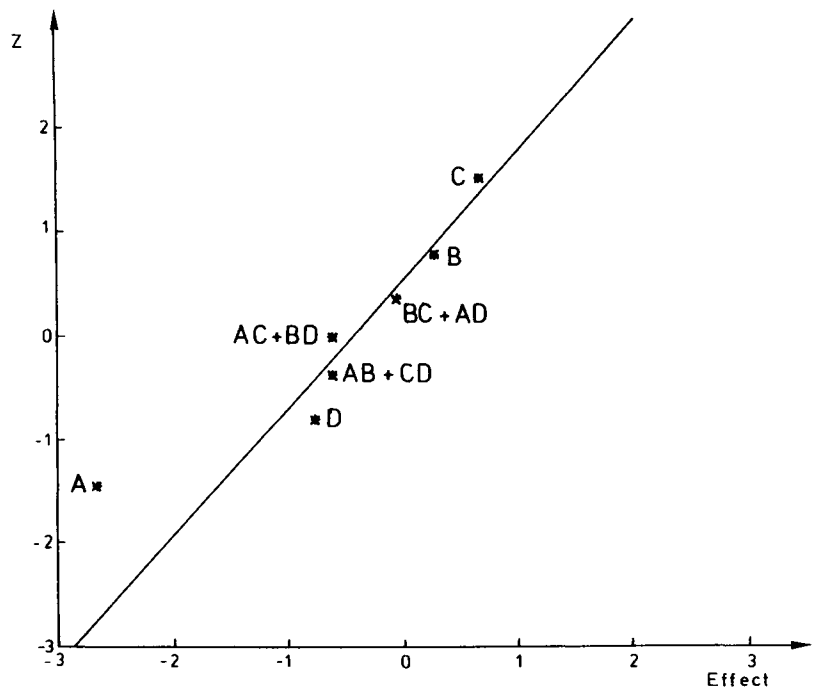


Fig. 23.2. The normal probability plot for the responses in the half-fractional spheronization design (Table 23.2).

23.3 Defining contrasts and generators

In the preceding section we selected 8 out of 16 experiments from a 2^4 design. In this section we explain how we decided which experiments to select. To understand this, we first need to redefine some algebraic rules, which we will first apply to find all aliases and later to generate fractional factorial designs. As before, an interaction, such as AB, is obtained by multiplying A and B (see Table 23.3). More exactly, to obtain the level for AB for a particular experiment, one multiplies the levels for A and for B for that experiment. Now, if one multiplies A by A, i.e. A^2, it is easy to verify that all levels for A^2 are equal to +, the same is true for B, etc.

A	A^2	ABCD
+	+	+
+	+	+
+	+	+
+	+	+
−	+	+
−	+	+
−	+	+
−	+	+

The column A^2 is also equal to ABCD and it corresponds to $\bar{y}$ (see Table 23.3).

We now define a new algebraic rule stating that a squared effect has a value of 1 and at the same time, for ease of notation, we equate the squared effects to I. It follows that:

$A^2 = B^2 = ABCD = I = 1$

We call the *defining contrast* or *defining relation*:

$I = ABCD$

The confounding patterns can be obtained by multiplication of the defining contrast with the effect for which one wants to know what the alias is. For instance the alias of A is obtained by

$AI = A^2BCD = BCD$

which means that A is confounded with BCD.

For AC, it is

$ACI = A^2BC^2D = BD$

so that AC is confounded with BD.

Summarizing, one obtains the aliases:

$ABCD = I, A = BCD, B = ACD, C = ABD, D = ABC, AB = CD, AC = BD, AD = BC.$

It is now possible to generate designs. To do this, it is necessary to note that the first 3 columns of the upper half of Table 23.3 constitute the full factorial design for 3 variables. We could say that to derive the 2^{4-1} from the 2^3 factorial, we have added to the 2^3 factorial a 4th column for factor D and that we have chosen it such that the levels for D are the same as for ABC. This is equivalent to saying that, to make the 2^{4-1} design, we have deliberately sacrificed ABC by confounding it with D.

To generate the design we have first written D = ABC. This is called the *generator*. The generator leads to the defining contrast, since we can derive that $D^2 = ABCD = I$.

This procedure can be applied to construct any 2^{k-p} factorial from a full 2^q factorial ($q = k-p$). To show how this is done let us construct a 2^{5-2} factorial from a 2^3 one, i.e. a quarter-replica of a 2^5 design. We start by writing down the generator.

First, we decide to sacrifice two of the higher interactions to accommodate D and E. For instance, it seems logical to sacrifice the ternary interaction ABC by putting it equal to D, as we already did for generating the 2^{4-1} design. We still need another such decision to include E and we decide that BC = E. This kind of decision should be made on the basis of a knowledge of the system. Usually the experts deciding on the factors to be studied know that certain interactions are less likely to be important than others. These can then be confounded with a main factor.

TABLE 23.5
A 2^{5-2} design

Effects	Experiments							
	1	2	3	4	5	6	7	8
A + BCD + ABCE + DE	+	+	+	+	−	−	−	−
B + ACD + CE + ABDE	+	+	−	−	+	+	−	−
C + ABD + BE + ACDE	+	−	+	−	+	−	+	−
D + ABC + BCDE + AE	+	−	−	+	−	+	+	−
E + ABCDE + BC + AD	+	−	−	+	+	−	−	+
AB + CD + ACE + BDE	+	+	−	−	−	−	+	+
AC + BD + ABE + CDE	+	−	+	−	−	+	−	+
$\bar{y}$ + ABCD + BCE + ADE	+	+	+	+	+	+	+	+

The quarter-replica design 2^{5-2} is now obtained by writing down the 8 experiments of the 2^3 design such that A, B, C are taken over from the 2^3 design, the D levels are written in the way one would write the levels for ABC in Table 23.3 and the E levels in the way one would write BC. This yields Table 23.5. The first 5 rows in that table define the experiments. The two two-factor interactions AB and AC were not used for adding main factors and therefore their levels can be computed in the usual way. It turns out that these interactions are confounded with other interactions. For all these factors we added all aliases and for good measure we also added the levels for computing $\bar{y}$ and its aliases. To obtain the aliases we derive the defining contrasts from the generator ABC = D, BC = E

ABC = D, or $ABCD = D^2 = I$ or

$$I = ABCD$$

and BC = E or $BCE = E^2 = I$ or

$$I = BCE$$

Moreover, we should now compute I^2. This yields an additional defining contrast. Indeed, if I = 1, $I^2 = I$. It follows that $I \times I = I = ABCD \times BCE = AB^2C^2DE = ADE$ so that I = ADE.

In summary, there are three defining contrasts

$$I = ABCD = BCE = ADE$$

which also means that the mean obtained from the 8 experiments is confounded with the effects ABCD, BCE and ADE ($\bar{y} = I$). We can now write all the aliases by multiplying each effect with the 3 defining contrasts. For instance A is confounded with

$A \times ABCD = BCD$

$A \times BCE = ABCE$

$A \times ADE = DE$

It should be noted that other plans would have been obtained if we had chosen to write the E levels as the AB or AC interaction. The confounding pattern is important, since it decides which factors or interactions can be evaluated without interference. If certain specific interactions are of interest to the experimenter, we will try not to confound them with each other.

23.4 Resolution

As explained in the preceding section, the fractional factorial design chosen depends on the generator. Certain factorial designs are better than others. For instance, we would prefer a design that does not confound main factors with two-factor interaction rather than one that does. To describe the quality of a fractional factorial design, we use the concept of *resolution*. The resolution is described by Roman numbers and is defined as follows: a design has resolution R if no p-factor effect is confounded with an effect containing less than $R - p$ factors. A design with R = III then has the property that no $p = 1$ (= main) effect is confounded with interactions containing less than $3 - 1 = 2$ factors. In other words, an R = III design is such that it does not confound main effects with effects that contain less than two factors, i.e. with other main effects.

The resolution depends on the generator and therefore on the defining contrasts. It is equal to the shortest defining contrast. In the 2^{5-2} design of Section 23.3, the shortest defining contrast has length 3 (BCE or ADE), so that the resolution is III. This is appended to the description of the design. The design of the preceding section is therefore a 2^{5-2} (III) design.

The following general rules apply:

– In designs of R = III the main effects are not confounded with other main effects, but they are confounded with two-factor interactions. Table 23.6 shows a 2^{7-4} (III) design.

– In designs of R = IV the main effects are not confounded with each other, nor with two-factor interactions. However two-factor interactions are confounded with other two-factor interactions.

– In designs of R = V the main effects and the two-factor interactions are not confounded with each other.

In all cases main effects and two-factor interactions are confounded with higher order interactions.

Computer programs for experimental design are usually able to propose designs with the highest resolution possible. It is possible to do this manually by careful consideration of the generator, but it is not very easy. One useful procedure is then to apply *folding-over*. As an example, let us consider the construction of a 2^{8-4}(IV) design from a 2^{7-4}(III) design (Table 23.7). We first add a column of + signs for the additional factor (H) and then "fold over", meaning that we add 8 experiments by reversing all signs.

TABLE 23.6

Saturated fractional factorial design for 7 factors: 2^{7-4}(III). Generators: D = ABC, E = AB, F = AC, G = BC

Experiment	Factors						
	A	B	C	D	E	F	G
1	−	−	−	−	+	+	+
2	+	−	−	+	−	−	+
3	−	+	−	+	−	+	−
4	+	+	−	−	+	−	−
5	−	−	+	+	+	−	−
6	+	−	+	−	−	+	−
7	−	+	+	−	−	−	+
8	+	+	+	+	+	+	+

TABLE 23.7

A 2^{8-4} (IV) design obtained from a 2^{7-4} (III) design by folding-over

Experiment	Factors							
	A	B	C	D	E	F	G	H
1	−	−	−	−	+	+	+	+
2	+	−	−	+	−	−	+	+
3	−	+	−	+	−	+	−	+
4	+	+	−	−	+	−	−	+
5	−	−	+	+	+	−	−	+
6	+	−	+	−	−	+	−	+
7	−	+	+	−	−	−	+	+
8	+	+	+	+	+	+	+	+
9	+	+	+	+	−	−	−	−
10	−	+	+	−	+	+	−	−
11	+	−	+	−	+	−	+	−
12	−	−	+	+	−	+	+	−
13	+	+	−	−	−	+	+	−
14	−	+	−	+	+	−	+	−
15	+	−	−	+	+	+	−	−
16	−	−	−	−	−	−	−	−

23.5 Embedded full factorials

In Section 23.2 we concluded that factor B was not important for the spheronization process and we could therefore decide not to consider it. The first 8 experiments constituting the 2^{4-1} design then constitute a full 2^3 design for A, C and D, as can be verified from looking at the + and – signs. One says that the full factorial for A, C and D is *embedded* in the half-fraction factorial design. This also means that we can now interpret the experiment as a full factorial design without any confounding between interactions of the remaining factors. Although this would not be of interest in this specific case, we can verify that it is also possible to eliminate factor A from consideration, which would leave a full factorial design for B,C and D. Similarly, full factorial experiments for A,B and D are obtained by eliminating C, or A,B and C by removing D. As a rule a full factorial can be obtained for every set of R–1 factors. The 2^{4-1} design constituted by the first 8 experiments of Table 23.3 has R = IV and therefore we should indeed find embedded full factorials for each combination of R – 1 = 3 factors.

23.6 Selection of additional experiments

The conclusion of the pharmaceutical example of Section 23.2 was that A and perhaps C and D have an effect. No interactions were found to be significant. However, suppose now that we had found that AC + BD have an important effect. We would then like to know whether the interaction AC is responsible or BD or perhaps even whether AC and BD both have an effect. If this is important enough we would decide to carry out the 8 experiments at the bottom of Table 23.3, thereby completing the full factorial so that we would now be able to decide unambiguously which interactions are important. The second half of Table 23.3 is also a half-replica of the full factorial. In other words, if we have carried out a half-replica factorial design and find that we should really have carried out the full factorial design, we can simply carry out the other half-replica factorial design. It is clear that such a strategy saves work and time: with some luck the first eight experiments will tell us what we want to know and, if not, we can still carry out the rest of the experiment without loss of information (or, at least, with very little loss as will be explained later). In general, the use of fractional factorial experiments as a starting design is to be recommended. Indeed, the first experimental design often serves to point out deficiencies in the setting up of the experiments. They may show, for instance, that the levels were not well chosen. If there is a doubt about whether a factor is indeed of interest, beginners in the practice of experimental design will often have the reflex to start with a full factorial design of what they think to be the more important factors. They will then have many experiments to do in order to

include the additional factor at a later stage. Carrying out a half-fraction design with $n + 1$ factors is more economical than carrying out the full factorial design with n factors.

The strategy outlined in this paragraph leads to a problem, namely that a block effect has to be included in the analysis of the full factorial. Indeed, we carried out the experiments in two blocks (in our example two blocks of eight experiments) and we might expect that this could also have an effect. The block effect is obtained from Table 23.3, by subtracting the results obtained within block 2 from those obtained within block 1.

$$\text{Block effect} = \left(\sum_{1}^{8} y_i - \sum_{9}^{16} y_i \right) / 8$$

It can be verified that the block effect is confounded with that of interaction ABCD. Since it is quite probable that this is negligible, the loss of information referred to earlier in this section is indeed very small.

When the number of variables is large or the cost of an experiment high, this strategy can be taken further. For instance, we could first carry out a quarter-fraction design, then as a second stage add a second quarter-fraction design, etc. Blocking then becomes a more important issue and more attention has to be paid to avoid loss of information. We refer to the literature for more details on how to choose the blocking pattern and select the second quarter-fraction [2]. In fact, it may be necessary, just as for full fractional factorial designs, to carry out the experiments of a fractional factorial design in blocks. Again, we refer to the literature [2] for indications of how to select the blocking pattern.

23.7 Screening designs

23.7.1 Saturated fractional factorial designs

When starting with an investigation there are often so many factors that it is desirable to first carry out a screening to know which are the most important, i.e. those that have a clear effect, and continue further work with those factors. From the way we explain this, it is clear that, at this stage, we are not interested in interactions. From our definition of resolution, it follows that we need at least an R = III design, since in such designs main effects are not confounded. Moreover, we would like to obtain the information we want with the minimum of experiments. This restricts the designs that are applied to a R = III design, since these are most economical. In practice, we often apply 2^{3-1}(III) or 2^{7-4}(III) designs (Table 23.6). These designs allow us to study (up to) 3 or 7 factors with 4 or 8 experiments. They are called *saturated* because they are designs with the smallest fractional

experiment possible. Indeed, the 2^{7-4}(III) design is a 1/16 fractional factorial design. A 1/32 design for 7 factors is not possible since this would require $2^{7-5} = 4$ experiments for 7 factors.

We do not necessarily need to carry out a 2^{3-1} or 2^{7-4} design. Highly fractional, but not necessarily saturated, designs can also be used. For instance, if we were to study 6 factors, we could use the 2^{7-4}(III) design, with a dummy factor (see further), but might prefer a 2^{6-2}(IV) design, because of its higher resolution. On the other hand, the use of *supersaturated* designs has also been described [3]. These designs use fewer experiments than there are main factors. This means, of course, that some main factors are confounded. They are used when many factors (sometimes more than 100) are considered and only a few are likely to be important.

A particular application of screening designs is the measurement of the ruggedness of a process [4] (see also Chapter 13). In this case, one has developed a process, for instance a measurement procedure, and wants to know whether small departures from the process parameters (when the process is a measurement procedure, these are factors that describe the analytical procedure) have an influence on the quality of the process. An example is given in Table 23.8. The procedure concerns a determination of tetracycline by HPLC [5]. The mobile phase contains an aqueous solution containing ammonium salts (0.1 M ammoniumoxalate/0.2 M ammoniumphosphate) and dimethylformamide (270 ml), the pH of which is adjusted to 7.65. The flow rate is 1 ml/min and the integration parameter is 2. The sixth factor, the age of the column, has no nominal value. Table 23.9 shows the design (a 2^{6-2}(IV) design). Table 23.10 gives the effects. They are obtained in the usual way, i.e. by applying eq. (23.1). Clearly, there is one overriding effect: the age of the column. The design allows the evaluation of the main effects (confounded with higher order effects, not mentioned in the table), while the

TABLE 23.8

Factors and their levels for the determination of tetracycline HCl [5]

Factors		Levels	
		−	+
A.	Inorganic substances in mobile phase Ratio $\frac{\text{M(ammoniumoxalate)}}{\text{M(ammoniumphosphate)}}$	$\frac{0.0975\ \text{M}}{0.195\ \text{M}}$	$\frac{0.1025\ \text{M}}{0.205\ \text{M}}$
B.	Dimethylformamide in mobile phase	260 ml	280 ml
C.	pH of mobile phase	7.50	7.80
D.	Flow of mobile phase	0.9 ml/min	1.1 ml/min
E.	Integration parameter (SN-ratio)	1	3
F.	Age of column	new column	2 weeks used

TABLE 23.9

A 2^{6-2}(IV) design. Application to the example of Table 23.8 [5]. The response, y, is the capacity factor of tetracycline. (E = ABC, F = BCD)

Experiment	A	B	C	D	E	F	y
1	–	–	–	–	–	–	1.59
2	+	–	–	–	+	–	1.48
3	–	+	–	–	+	+	1.12
4	+	+	–	–	–	+	1.13
5	–	–	+	–	+	+	1.28
6	+	–	+	–	–	+	1.18
7	–	+	+	–	–	–	1.42
8	+	+	+	–	+	–	1.69
9	–	–	–	+	–	+	1.21
10	+	–	–	+	+	+	1.24
11	–	+	–	+	+	–	1.36
12	+	+	–	+	–	–	1.37
13	–	–	+	+	+	–	1.58
14	+	–	+	+	–	–	1.70
15	–	+	+	+	–	+	1.20
16	+	+	+	+	+	+	1.21

TABLE 23.10

Evaluation of the results of the quarter-fraction factorial design for the example of Tables 23.8 and 23.9

Factors	Capacity factor effect
A	0.030
B	–0.094
C	0.097
D	–0.004
E	0.021
F	–0.328
AB+CE	0.043
AC+BE	0.045
AD+EF	0.014
AE+BC	0.040
AF+DE	–0.042
BD+CF	–0.050
BF+CD	0.034
$E_{critical}$	0.094

two-factor interactions are pairwise confounded. Since in screening designs, we consider only the main effects, we also consider that the interaction effects are negligible and use them to determine s_{effect} in the same way as higher order interactions were used in Section 22.6.2. This allows us to compute $E_{critical}$. Effects larger than $E_{critical}$ are considered significant.

23.7.2 Plackett–Burman designs

For more than 7 factors, we would need to carry out a design based on 16 experiments. Plackett and Burman [6] have proposed experimental designs for $n \times 4$ experiments, i.e. 4, 8, 12, 16, 20, etc., that are suitable for studying up to 3, 7, 11, 15, 19, etc. factors, respectively. In some cases where $n \times 4 = 2^k$, the *Plackett–Burman* design is a specific fraction of a full factorial design and saturated fractional factorial designs can be used just as well. However, this is not the case for multiples of 4 that are not equal to a power of 2. Let us consider the case of 12 experiments for 11 factors.

The Plackett–Burman designs have the particularity that they are cyclical. Consider, for example, the 11-factor, 12-experiment design. It is obtained from a first line given in their paper and which in this instance is

+ + − + + + − − − + −

and describes the first experiment. Experiments 2 to 11 are obtained by writing down all cyclical permutations of this line. The last experiment, 12, always contains only minus signs. The complete design is therefore given in Table 23.11. We can verify that in this way each factor is measured at 6 + and 6 − levels and it is also possible to verify that the main factors are not confounded when the effects are determined in the usual way, i.e.

$$\text{Effect} = \tfrac{1}{6}\left[\sum (y \text{ at } + \text{ levels}) - \sum (y \text{ at } - \text{ levels})\right]$$

When fewer than the maximum possible number of factors are to be studied, dummy factors are added. Suppose that 8 factors must be investigated. This requires an $n = 12$ design and such a design accomodates 11 factors. The solution to this difficulty is to carry out the $n = 12$ design where 3 factors are dummy factors. As Youden and Steiner [7] stated, we should "associate with such factors some meaningless operations such as solemnly picking up the beaker, looking at it intently and setting it down again". In Table 23.11 this means we would derive effects only for the first 8 factors (A–F). However, we can also compute effects for dummy factors *d*1–*d*5, although these effects are meaningless. Nevertheless, the computation is useful. Indeed, these dummy effects can be used in the same way as the interaction effects of Table 23.10 to compute an s_{effect}. This is done for the

ruggedness example of Table 23.8 (see Table 23.12). Here, the author could have chosen to carry out a design with 8 experiments; however, he preferred to carry out 12, so that he would have five dummy factors and be better able to determine the significance of the estimated effects.

TABLE 23.11

Plackett–Burman design for eleven factors. Application to the example of Tables 23.8 and 23.9 [5]. The response, y, is the capacity factor of tetracycline. $d1$ stands for dummy 1.

Experiment	Factor											
	A	$d1$	B	$d2$	C	$d3$	D	$d4$	E	$d5$	F	y
1	+	+	−	+	+	+	−	−	−	+	−	1.84
2	−	+	+	−	+	+	+	−	−	−	+	1.16
3	+	−	+	+	−	+	+	+	−	−	−	1.45
4	−	+	−	+	+	−	+	+	+	−	−	1.64
5	−	−	+	−	+	+	−	+	+	+	−	1.44
6	−	−	−	+	−	+	+	−	+	+	+	1.21
7	+	−	−	−	+	−	+	+	−	+	+	1.15
8	+	+	−	−	−	+	−	+	+	−	+	1.27
9	+	+	+	−	−	−	+	−	+	+	−	1.46
10	−	+	+	+	−	−	−	+	−	+	+	1.13
11	+	−	+	+	+	−	−	−	+	−	+	1.24
12	−	−	−	−	−	−	−	−	−	−	−	1.53

TABLE 23.12

Effects on the capacity factor of tetracycline from the Plackett–Burman design

Factors	Capacity factor effect
A	0.048
B	−0.128
C	0.071
D	−0.067
E	0.000
F	−0.367
Dummy 1	0.080
Dummy 2	0.086
Dummy 3	0.035
Dummy 4	−0.060
Dummy 5	−0.010
$E_{critical}$	0.157

References

1. M. Jimidar, M.S. Khots, T.P. Hamoir and D.L. Massart, Application of a fractional factorial experimental design for the optimization of fluoride and phosphate separation in capillary zone electrophoresis with indirect photometric detection. Quim. Anal., 12 (1993) 63–68.
2. G.E.P. Box, W.G. Hunter and J.S. Hunter, Statistics for Experiments. Wiley, New York, 1978.
3. D.K. Lin, Generating systematic supersaturated designs. Technometrics, 37 (1995) 213–225.
4. M.M.W.B. Hendriks, J.H. De Boer and A.K. Smilde, Robustness of Analytical Chemical Methods and Pharmaceutical Technological Products. Elsevier, Amsterdam, 1996.
5. Y. Vander Heyden, K. Luypaert, C. Hartmann, D.L. Massart, J. Hoogmartens and J. De Beer, Ruggedness tests on the high-performance liquid chromatography assay of the United States Pharmacopeia XXII for tetracycline hydrochloride. A comparison of experimental designs and statistical interpretation. Anal. Chim. Acta, 312 (1995) 245–262.
6. R.L. Plackett and J.P. Burman, The design of optimum multifactorial experiments. Biometrika, 33 (1946) 305–325.
7. W.J. Youden and E.H. Steiner, Statistical Manual of the Association of Official Analytical Chemists. The Association of Official Analytical Chemists, Arlington, VA, 1975.

Chapter 24

Multi-level Designs

24.1 Linear and quadratic response surfaces

Two-level designs allow the estimation of the effect of all the factors and their interactions. Multi-level designs are used in different contexts. For qualitative factors, one often has no other choice but to create as many levels as there are different values for that factor. Designs used in this context are described in Section 24.7.

In this chapter the emphasis will mainly be on the multi-level designs for quantitative factors. The two-level designs of the preceding chapters can describe only lines, planes or hyperplanes. Consider, for example, Fig. 24.1. This describes the situation for a one-factor design. In Fig. 24.1a two levels were measured. This allows the drawing of a straight line through the measurement points. In Fig. 24.1b three levels were included and a second-order model, a parabolic curve, can then be drawn through the three points. Curved models therefore require measurements at three or more levels. This is true for each of the factors that are expected to show a curvilinear relationship. Designs with more than two levels for quantitative factors are described in Sections 24.3 and 24.4.

As explained for non-linear regression (Chapter 11), one has in general a choice between so-called mechanistic and empirical modelling. In the context of experimental design, empirical models are nearly always used and the models are mainly second-order, i.e. quadratic. Higher-order models are rare and are used only when quadratic models are clearly inadequate, for instance when a sigmoid relationship must be described, such as when pH is involved. One might then choose a third-order model, an appropriate transform (such as the logistic transform [1]) or a mechanistic (physical) model [2]. In the latter case the non-linear regression techniques of Chapter 11 are required (see further Section 24.6). However, as already stated, in an experimental design context this is unusual and most of this chapter will therefore be devoted to using quadratic models.

Such a quadratic model includes a constant term, first and second-order terms and the interaction between factors (usually limited to the two-factor interactions). For two variables, the model is:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \qquad (24.1)$$

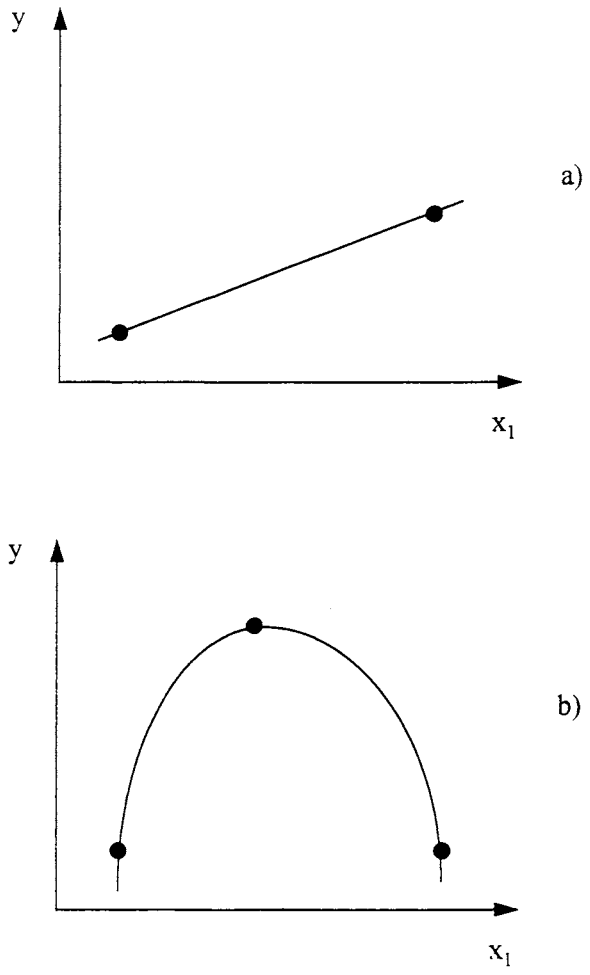


Fig. 24.1. (a) Univariate linear model derived from measurement at two levels; (b) univariate quadratic model derived from measurement at three levels.

The design leads to the estimation of the β-coefficients, by *b*-values, resulting in the following equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2 \qquad (24.2)$$

For three variables the quadratic equation is

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 \qquad (24.3)$$

How well the *b*-coefficients are estimated depends on the experimental design. Quality criteria for these designs are described in Section 24.2. Readers for whom

this chapter is a first introduction to multi-level designs may prefer to skip this section and read Section 24.3 first in which the more usual designs are described.

When the *b*-coefficients have been obtained by multiple regression, one can use them to predict the response $\hat{y}$ as a function of the x-factors. This leads to the construction of so-called *response surfaces*. Some typical quadratic response surfaces for two factors are shown in Fig. 24.2. As said above, models of order higher than two are rarely used, because in many cases the true response surface can be approximated by a second-order model. When this is not the case (see also Chapter 21.6), this is often due to an inadequate choice of response. An example of the use of response surfaces is given in Section 24.5.

Several books or review papers on response surface methodology have been published. Apart from the books already cited in Chapter 21, one can refer for

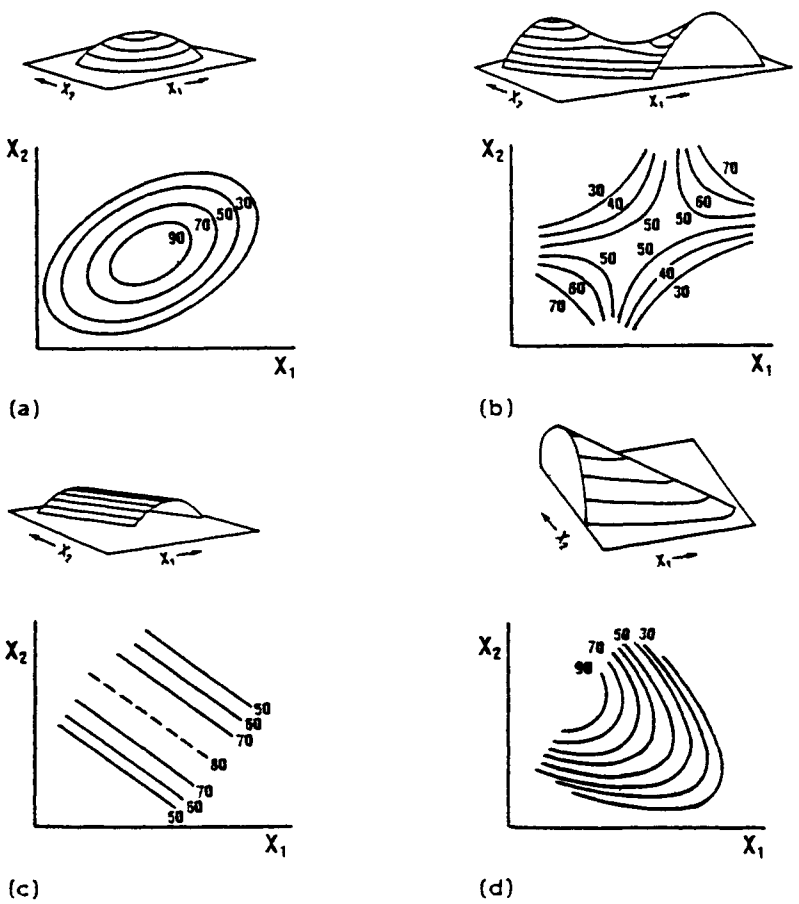


Fig. 24.2. Some quadratic bivariate response surfaces and the resulting contour plots (adapted from Ref. [3]): (a) mound, (b) saddle surface, (c) stationary ridge, and (d) rising ridge.

instance to books and articles by Atkinson [3,4], Draper and Box [5], Phan-Tan-Luu [6], Frantz et al. [7], Morgan et al. [8] and Fernandes de Aguiar et al. [9].

24.2 Quality criteria

24.2.1 D-, M- and A-optimality

In Chapter 10.2 we derived that the *b*-coefficients in equations such as (24.2) and (24.3) are given by

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y} \tag{24.4}$$

where **b** is the column vector of coefficients, **y** is the column vector of n measurements, and **X** is the *independent variable matrix*, sometimes also called *model matrix*. For instance, for eq. (24.2):

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_{11} \\ b_{22} \\ b_{12} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}^2 & x_{12}^2 & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}^2 & x_{22}^2 & x_{21}x_{22} \\ & & & \cdot & & \\ & & & \cdot & & \\ & & & \cdot & & \\ 1 & x_{n1} & x_{n2} & x_{n1}^2 & x_{n2}^2 & x_{n1}x_{n2} \end{bmatrix} \tag{24.5}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

X is derived from the model and from the *design matrix* **D**. The latter contains the factor combinations at which one should carry out the experiments. Here the design matrix would be

$$\mathbf{D} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{n1} & x_{n2} \end{bmatrix}$$

In Section 10.5 it was pointed out that the $(\mathbf{X}^T\mathbf{X})^{-1}$ matrix determines the variance of the regression parameters, i.e. the quality of the model that will be obtained. $\mathbf{X}$ depends only on the model and on the values chosen for $x_1, \ldots x_m$, i.e. on the design. This means that for a given model the distribution of the experiments, i.e. the values of the x-factors over the experimental domain, determines the quality of a model.

It is very important to note that $(\mathbf{X}^T\mathbf{X})^{-1}$ does not include information about the response. In other words, all the information required to evaluate the effect of the design on the quality of the estimated model is present before any experiment has been carried out. This leads to the important conclusion that, for a given experimental error, the quality of the estimation of $\mathbf{b}$ and of the prediction of $\mathbf{y}$ is determined exclusively by the experimental design.

The ideal situation is when the design matrix is *orthogonal*. This means that there is no correlation between the factors and that the estimation of $\mathbf{b}$ is best. This is obtained when the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ is diagonal. Often this is not possible and one will then try to approach this situation as closely as possible. More generally, it can be shown that the estimation of $\mathbf{b}$ is best, when the determinant of $(\mathbf{X}^T\mathbf{X})^{-1}$ is minimal or in other words the volume of the joint confidence interval for the b's is minimal. Since:

$$\det((\mathbf{X}^T\mathbf{X})^{-1}) = 1/\det(\mathbf{X}^T\mathbf{X})$$

the determinant of $(\mathbf{X}^T\mathbf{X})$ should be maximal. As $\det(\mathbf{X}^T\mathbf{X})$ generally increases with the number of experimental points, it follows that the estimation will be better when there are more points. For designs with the same number of points, one can compare the determinants of $(\mathbf{X}^T\mathbf{X})$ to decide which one will give the best estimation of $\mathbf{b}$. A design is called *D-optimal* compared to other designs with the same number of experimental points, when its $\det(\mathbf{X}^T\mathbf{X})$ is largest. The classical two-level factorial designs (both the full design of Chapter 22 and the fractional factorial ones of Chapter 23) are both D-optimal for first-order models and, in fact, orthogonal. One should note again that D-optimality has to do with the quality of estimation of a certain set of b's. This set depends on the model and therefore one computes D-optimality for a given model.

In Section 24.4 we will see that in certain cases the usual symmetric designs cannot be used. It is then necessary to apply explicitly the D-optimality principle

to select the best design. A general property of D-optimal designs [3] is that they concentrate experimental effort on only a few sets of conditions. For instance, for a single factor x and a model $\hat{y} = b_0 + b_1 x + b_{11} x^2$ it is shown that 3 levels of x are optimal i.e. the two extreme levels of x (1 and + 1) and a value at the centre of this region (0). If 3 experiments are carried out, they should be performed at these levels; if a fourth experiment is performed, D will be optimal when it is carried out by replicating one of these three levels and not, as one might perhaps expect, by introducing a fourth level so that measurements are carried out at -1, $-1/3$, $+1/3$ and $+1$.

Several other criteria exist. For instance the M-criterion is given by det $(\mathbf{M}) = (\mathbf{X}^T\mathbf{X})/n$. It compares designs with different numbers of experiments. The A-criterion is based on the trace of $(\mathbf{X}^T\mathbf{X})^{-1}$. In eq. (10.18) the variance–covariance matrix of the b coefficients is introduced. The variances of the coefficients are the diagonal elements of $\mathbf{V}(b)$ and, therefore, these diagonal elements should be small. The values of the elements are determined by $(\mathbf{X}^T\mathbf{X})^{-1}$ (see eq. (10.18)). The variance s_e^2 can be considered a constant depending on the quality of the measurement, and not on the design. The influence of the design on the variance of the estimates is therefore determined by the trace or sum of the diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$, which should be small: in that case the average variance of the b's is small. This leads to the formulation of the A-optimality criterion (with the "A" of "average"). A design is called *A-optimal* compared with other designs with the same number of experimental points, when its $\mathrm{tr}(\mathbf{X}^T\mathbf{X})^{-1}$ is smallest.

Let us consider as an example the following situation [9]. Six experiments described by two variables could be carried out:

$$\mathbf{D} = \begin{bmatrix} 2.1 & 3.0 \\ 3.5 & 1.5 \\ 4.9 & 4.0 \\ 5.1 & 2.6 \\ 5.7 & 1.0 \\ 7.0 & 2.4 \end{bmatrix} \begin{matrix} (1) \\ (2) \\ (3) \\ (4) \\ (5) \\ (6) \end{matrix}$$

From these six candidate points, we would like to use only four. The question then is which set of four should we select: will e.g. the set {(1), (2), (3), (4)} perform better than any other set of four experiments. This question can be answered only with respect to a model. Let us suppose this model is:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2^2$$

This, of course, is an unusual model, which is applied here only for illustration. It should be noted that we do not give a vector of responses $\mathbf{y}$: the idea is to choose the best set of 4 experiments i.e. the best design, and this must be done before experiments are carried out.

The model matrix can now be determined for each of the combinations of four experiments. For example for the set {(1), (2), (3), (5)} it is given by:

$$\mathbf{X}_{1235} = \begin{bmatrix} 1 & 2.1 & 9.00 \\ 1 & 3.5 & 2.25 \\ 1 & 4.9 & 16.00 \\ 1 & 5.7 & 1.00 \end{bmatrix}$$

The D-criterion is given by $\det(\mathbf{X}^T\mathbf{X}) = 4.285 \times 10^3$.

It should be noted that D can take on large values. To avoid numerical problems, one often first scales the variables.

The A-criterion, for the same example would be

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 2.99 & -0.57 & -0.06 \\ -0.57 & 0.13 & 0.00 \\ -0.06 & 0.00 & 0.01 \end{bmatrix}$$

$$\mathrm{tr}(\mathbf{X}^T\mathbf{X})^{-1} = 2.99 + 0.13 + 0.01 = 3.13$$

In the present case, it is found that the best combination according to the D-optimality criterion is {(1), (2), (3), (6)}. This is also the best for the A-criterion.

24.2.2 Rotatability, uniformity and variance-related criteria

A design is called *rotatable*, when the variance of the prediction does not depend on the direction in which one looks starting from the centre point, but only on the distance from the centre point. The two-level designs, applied in Chapters 22 and 23 are both orthogonal and rotatable. A design is rotatable only when the experiments are roughly situated on a (hyper)sphere. However, not all spherical designs are rotatable (e.g. the Doehlert design, Section 24.3.4). By adequate selection of the number of centre points, it is possible to arrange that the precision of the response of a predicted design is similar over the whole domain. Such a design is said to have *uniform precision.*

The *variance function* is a measure of the uncertainty in the predicted response. From eq. (10.19), one can derive that:

$$\mathrm{Var}(\hat{y}_i) = \mathbf{x}_i^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_i s_e^2 = d(x_i)\, s_e^2$$

where $\hat{y}_i$ is the predicted response in point i, $\mathbf{x}_i$ is the vector describing a certain experiment, s_e is the experimental standard deviation and $d(x_i)$ is the variance function at point i. Of course, one wants $\mathrm{Var}(\hat{y}_i)$ to be as small as possible, but, in

general, one considers $d(x_i) = 1$ acceptable, because this means that $\mathrm{Var}(\hat{y}_i) = s_e^2$, i.e. that the prediction uncertainty is equal to the experimental uncertainty [6]. From the variance function, one derives the G-optimality criterion. Designs are considered *G-optimal* when, for the same number of experiments n, the maximal value of the variance function in the experimental region is minimal. To compare designs with different n, one can compute the G-efficiency as G-eff = $p/(d(x)\cdot n)$ where $d(x)$ is the maximal value of $d(x_i)$ and p is the number of coefficients in the model.

In the preceding section, we saw that it is possible to evaluate the quality of designs, for which the model is known. Moreover, as will be stressed further in Section 24.6, these models have to be linear in the regression sense (see Section 11.1) to be able to apply such criteria in a simple way (quadratic and in general polynomial models are linear in the regression sense). Several experimental design specialists such as Scheffé, Plackett and Doehlert [10] have stated that in cases, where the criteria of Section 24.2.1 cannot be (easily) applied, the experimenter ought to look for designs with an equally spaced distribution of points. Designs which show this property are said to show *uniformity* of space filling.

24.3 Classical symmetrical designs

In this section we will introduce the most often used designs. These designs are all highly symmetrical. Most of them score very well on the criteria described earlier. The experimental domain they describe can be (hyper)spherical or (hyper)cubic. Most of the designs described later are spherical. This is the case for the central composite designs (except the face-centred central composite design), the Box–Behnken and the Doehlert design. Cubic designs are the 3-level factorial design and the face-centred central composite design. It should be noted that the terms cubic and spherical are used even for a two-variate situation. One should consider what experimental domain exactly one wants to describe and take good care not to extrapolate outside the region described when making predictions. The prediction error then becomes much larger and the model may not be correct outside the experimental region.

It should be noted that the experimental design describes the experiments that have to be performed to obtain the model relating y to x. It can be useful to do additional experiments for various reasons. For instance, one can replicate the centre point in the design to have an idea of the experimental error. Replication of experimental points allows validation of the model (see Section 10.3.1) and the determination of additional points, different from the experimental design points allows validation of the prediction performance (see Section 10.3.4).

TABLE 24.1

A two-factor three-level (3^2) design

Experiment	x_1	x_2
1	−1	−1
2	−1	0
3	−1	+1
4	0	−1
5	0	0
6	0	+1
7	+1	−1
8	+1	0
9	+1	+1

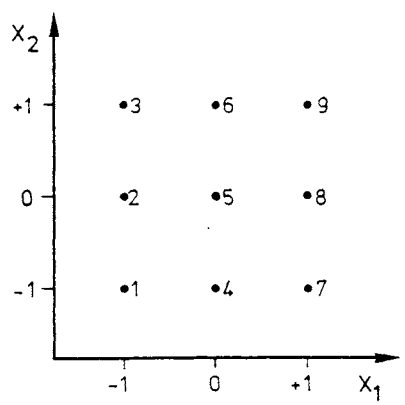


Fig. 24.3. A 3^2 factorial design.

24.3.1 Three-level factorial designs

The full three-level factorial design, 3^k, can be used to obtain quadratic models, but except for very small k, it requires rather many experiments (9 for $k = 2$, 27 for $k = 3$, 81 for $k = 4$). The $k = 2$ design is given in Table 24.1 and Fig. 24.3. When one has the resources to do more than 9 experiments and $k = 2$, then it has been shown that good D-optimality is obtained for $n = 13$. The 4 additional experiments are not situated at additional levels (see Section 24.2.1) but are used to replicate the four corner points of the design.

The three-level factorial design is the only often applied multi-level design that is completely orthogonal. It is, however, not rotatable. For the two-factor design of Table 24.1 the **X**-matrix for the model of eq. (24.2) is

$$\mathbf{X} = \begin{bmatrix} +1 & -1 & -1 & +1 & +1 & +1 \\ +1 & -1 & 0 & +1 & 0 & 0 \\ +1 & -1 & +1 & +1 & +1 & -1 \\ +1 & 0 & -1 & 0 & +1 & 0 \\ +1 & 0 & 0 & 0 & 0 & 0 \\ +1 & 0 & +1 & 0 & +1 & 0 \\ +1 & +1 & -1 & +1 & +1 & -1 \\ +1 & +1 & 0 & +1 & 0 & 0 \\ +1 & +1 & +1 & +1 & +1 & +1 \end{bmatrix}$$

and

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 9 & 0 & 0 & 6 & 6 & 0 \\ 0 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 \\ 6 & 0 & 0 & 6 & 4 & 0 \\ 6 & 0 & 0 & 4 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

We notice that most covariance coefficients are 0, except for some coefficients related to b_0 and the term for (b_{11}, b_{22}), i.e. the term that describes the covariance between the quadratic terms. The parameter b_0 does not describe a factor and by taking the average of the columns into consideration, the covariance between the quadratic terms vanishes: by subtracting 6/9 (the average for the fourth and fifth columns in **X**) from all values in those columns, the covariance for (b_{11}, b_{22}) becomes zero.

Fractional factorial designs have also been proposed. These so-called *orthogonal arrays* are described with the same notation as in Chapter 23 for two-level fractional factorials. For instance, a 3^{4-2} fractional design is the $(1/3^2)$ part of a 3^4 design. It therefore consists of 9 experiments (Table 24.2). These designs have not, however, been applied very often.

TABLE 24.2
A 3^{4-2} design

Experiment	x_1	x_2	x_3	x_4
1	−1	−1	−1	−1
2	−1	0	0	0
3	−1	+1	+1	+1
4	0	−1	0	+1
5	0	0	+1	−1
6	0	+1	−1	0
7	+1	−1	+1	0
8	+1	0	−1	+1
9	+1	+1	0	−1

24.3.2 Central composite designs

To solve the problem of economy so-called *central composite designs* have been proposed. Examples are given in Tables 24.3 and 24.4 and in Figs. 24.4 and 24.5. Central composite designs always consist of the following three parts (Fig. 24.4):

– A two-level factorial design. In Table 24.3, the first four experiments constitute a full 2^2 design and in Table 24.4 the first eight a full 2^3 design.

– A *star design*. To add more levels so as to be able to describe curvature one adds points, which are described as a star. Points 5–8 in Table 24.3 and points 9–14

TABLE 24.3

A two-factor central composite design

Experiment	x_1	x_2
1	−1	−1
2	+1	−1
3	−1	+1
4	+1	+1
5	$-\sqrt{2}$	0
6	$+\sqrt{2}$	0
7	0	$-\sqrt{2}$
8	0	$+\sqrt{2}$
9 etc.	0	0

TABLE 24.4

A three-factor central composite design

Experiment	x_1	x_2	x_3
1	−1	−1	−1
2	+1	−1	−1
3	−1	+1	−1
4	+1	+1	−1
5	−1	−1	+1
6	+1	−1	+1
7	−1	+1	+1
8	+1	+1	+1
9	−1.682	0	0
10	+1.682	0	0
11	0	−1.682	0
12	0	+1.682	0
13	0	0	−1.682
14	0	0	+1.682
15 etc.	0	0	0

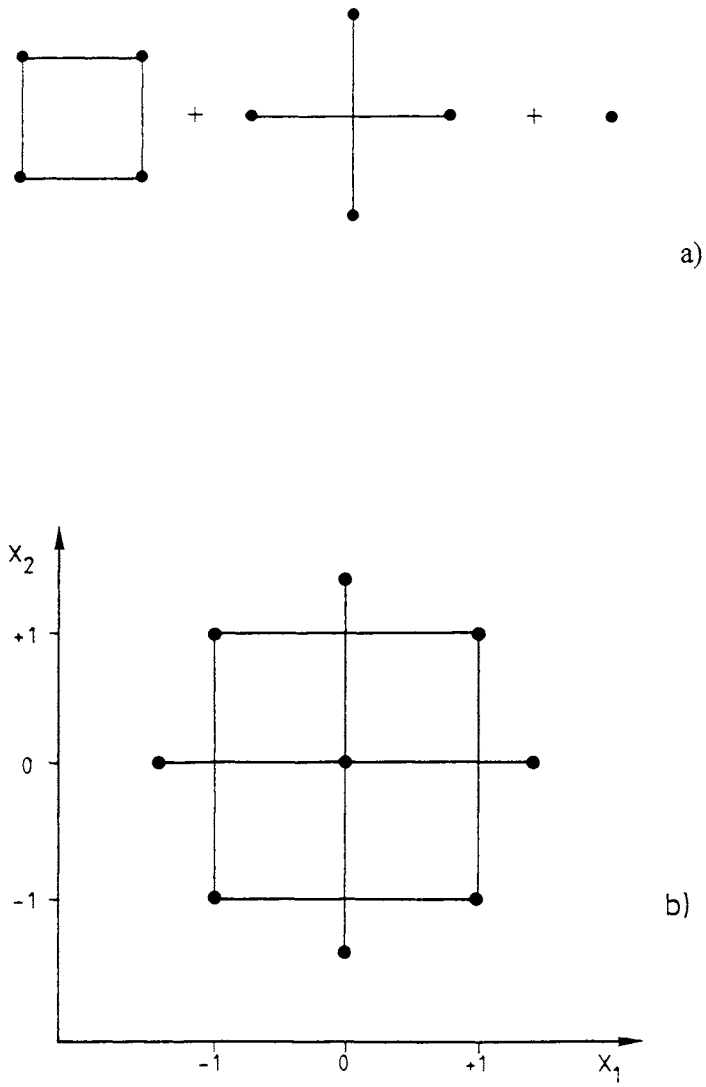


Fig. 24.4. Two-factor central composite design. (a) The design consists of a 2^2 factorial design plus a star design and the centre point. (b) The composite design.

in Table 24.4 are the star points. They are situated in general at a distance α (here 1.404 and 1.682, respectively) from the centre of the design. How to decide on the value of α is described further.

– The centre point. This is often replicated. For this reason it is designed as 9, etc. and 15, etc. in Tables 24.3 and 24.4.

In general, there are therefore

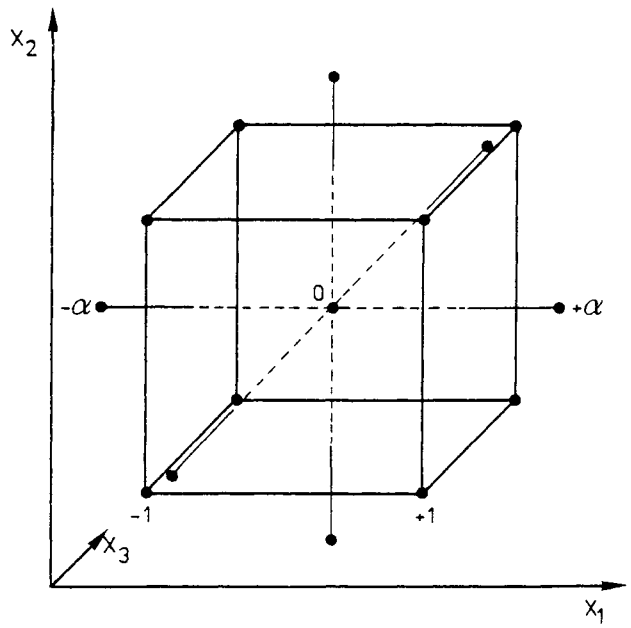


Fig. 24.5. Central composite design for 3 factors. The levels are indicated only for x_1.

$n_c = 2^k$ cube points (for a full factorial) with levels of – 1 and + 1

$n_s = 2^k$ star or axial points with levels of $-\alpha$ and $+\alpha$

n_0 1.68 centre points with all levels equal to 0.

Each factor is encountered at five levels ($-\alpha, -1, 0, +1, +\alpha$) or at three (for $\alpha = 1$, which is however unusual). The number of experiments is much less than for a 3-level factorial design. For four factors, one needs 16 ($= n_c$) + 8 ($= n_s$) + at least 1 ($= n_0$) = at least 25 experiments (compared with $3^4 = 81$). In Tables 24.3 and 24.4 the most usual value of $\alpha (= \sqrt{2} = 1.414$ for $k = 2$ and $\sqrt[4]{8} = 1.68$ for $k = 3$) is given, but other values are possible as explained below.

Three types of central composite design are sometimes considered. Those of Tables 24.3 and 24.4 are then called *central composite circumscribed* (CCC) because $|\alpha|$ is larger than 1. One thinks of the – 1 and + 1 levels as the boundaries of the experimental design set by the user and the axial points come outside this region. When these boundaries should not be exceeded because it is not possible for experimental reasons, one employs the *central composite inscribed* (CCI). One sets $+\alpha$ and $-\alpha$ equal to the boundaries, so that the design is completely included within the experimental boundaries. If one were to set $+\alpha = +1$ for x_1, then (for $k = 2$), one could set the levels of the 2^2 design at $1/\sqrt{2} = 0.712$ and -0.712 instead of + 1 and – 1. A third type is the *central composite face-centred* (CCF), for which

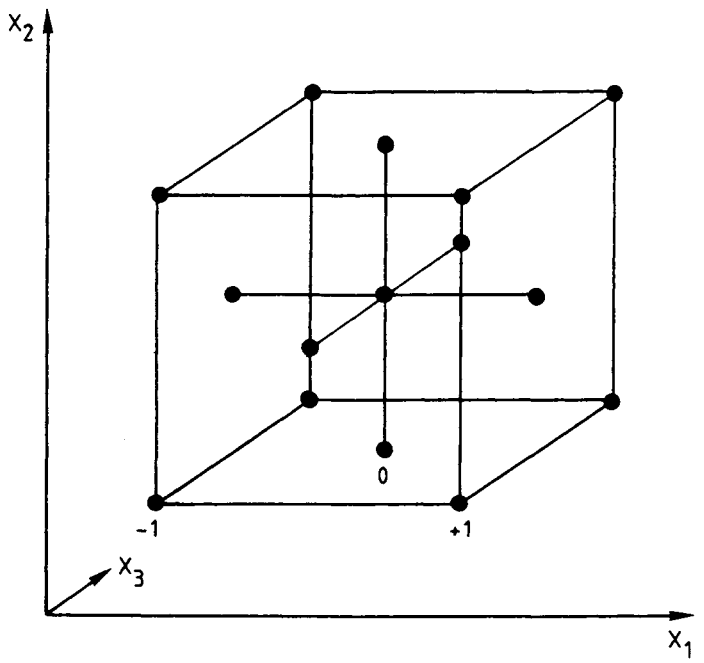


Fig. 24.6. Face centred central composite design. The levels are indicated only for x_1.

$|\alpha| = 1$. The result is a face-centred cube (see Fig. 24.6). It is a cubic design, in contrast with the two others which are spherical. It is, however, used much less than the spherical CCC designs and in what follows, we will discuss the latter designs.

In Fig. 24.7 all points, both the cube and the star points, are situated on a circle so that the design is rotatable, with $\alpha = r$ and $d = +1$ or -1. Since $\sin\gamma = d/r$, it follows that $\alpha = \pm(1/\sin\gamma)$. Since $\sin\gamma = \sin 45° = 1/\sqrt{2}$, $\alpha = \sqrt{2} = 1.414$. More generally, it can be shown that rotatability is achieved when $\alpha = (n_c)^{1/4}$. Table 24.5 gives the values of α required for the lower values of k, both for full factorial and fractional factorial cubic parts of the design.

The centre points are often replicated. There are good practical reasons for this (see also Chapter 22). The replicated centre points give an immediate idea of experimental precision. When one blocks the designs, comparing the results of centre points in each block gives an indication of whether block effects occur. Of special interest is the value of n_0 required to achieve (near) orthogonality. In this case one achieves both (near) orthogonality and rotatability. These values of n_0 are given in the third column of Table 24.5. As an example, for $k = 3$, $\alpha = 1.68$ and $n_0 = 9$ yield both an orthogonal and rotatable design. If orthogonality is considered less important than uniform (prediction) precision, then smaller n_0 numbers are usually recommended. These numbers are given in the fourth column of Table 24.5.

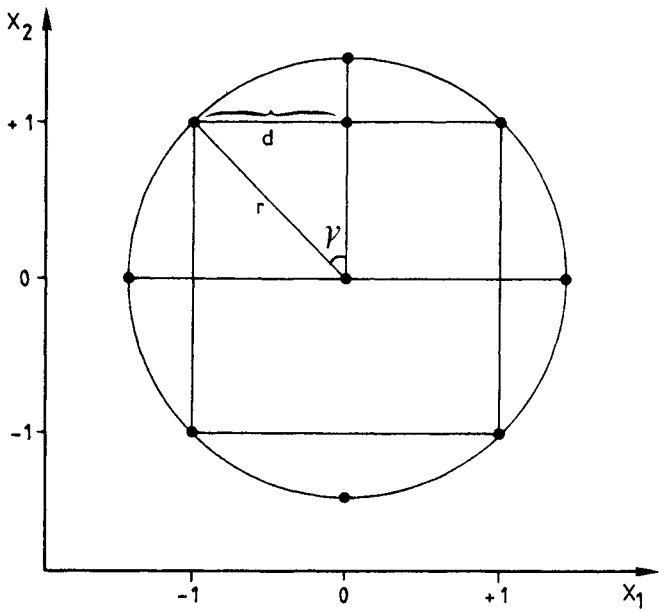


Fig. 24.7. Central composite (circumscribed) design. All points, except the centre point, are situated on a circle of radius r. The meaning of d and γ is explained in the text.

TABLE 24.5

Starpoint distances α and number of centre points n_0 for central (circumscribed) composite designs (adapted from Ref. [8])

k	α for rotatable central composite	n_0 for combined orthogonality and rotatability	n_0 for uniform precision
2	1.40	8	5
3	1.68	9	6
4	2.00	12	7
5	2.38	17	10
5*	2.00	10	6
6	2.83	24	15
6*	2.38	15	9

* The factorial design is fractional.

The cubic part of the design usually is a full factorial design. It is however also possible to use fractional factorial designs for this purpose. A biotechnical application is given in [11]. This study is concerned with among others the yield of *S. cerevisiae* as a function of the content of glucose, NH_4^+, K^+, $H_2PO_4^-$ and Mg^{2+}, in the culture. The design is described in Table 24.6. The first 16 points constitute a 2^{5-1} design.

TABLE 24.6

Central composite design with half-fractional cubic part for a biotechnological example [11]

Run	[Glucose]	$[NH_4^+]$	$[K^+]$	$[H_2PO_4^-]$	$[Mg^{2+}]$
1	−1	−1	−1	−1	−1
2	1	−1	−1	−1	1
3	−1	1	−1	−1	1
4	1	1	−1	−1	−1
5	−1	−1	1	−1	1
6	1	−1	1	−1	−1
7	−1	1	1	−1	−1
8	1	1	1	−1	1
9	−1	−1	−1	1	1
10	1	−1	−1	1	−1
11	−1	1	−1	1	−1
12	1	1	−1	1	1
13	−1	−1	1	1	−1
14	1	−1	1	1	1
15	−1	1	1	1	1
16	1	1	1	1	−1
17	−2	0	0	0	0
18	2	0	0	0	0
19	0	−2	0	0	0
20	0	2	0	0	0
21	0	0	−2	0	0
22	0	0	2	0	0
23	0	0	0	−2	0
24	0	0	0	2	0
25	0	0	0	0	−2
26	0	0	0	0	2
27	0	0	0	0	0
28	0	0	0	0	0
29	0	0	0	0	0
30	0	0	0	0	0
31	0	0	0	0	0
32	0	0	0	0	0

24.3.3 Box–Behnken designs

The *Box–Behnken* design for $k = 3$ is shown in Fig. 24.8 and in Table 24.7. It is a spherical, rotatable design. Viewed on a cube (Fig. 24.8a), it consists of the centre point and the middle points of the edges. It can also be viewed (Fig. 24.8b) as consisting of three interlocking 2^2 factorial designs and a centre point. It is economical, since it requires 13 experiments for $k = 3$. It should be stressed that,

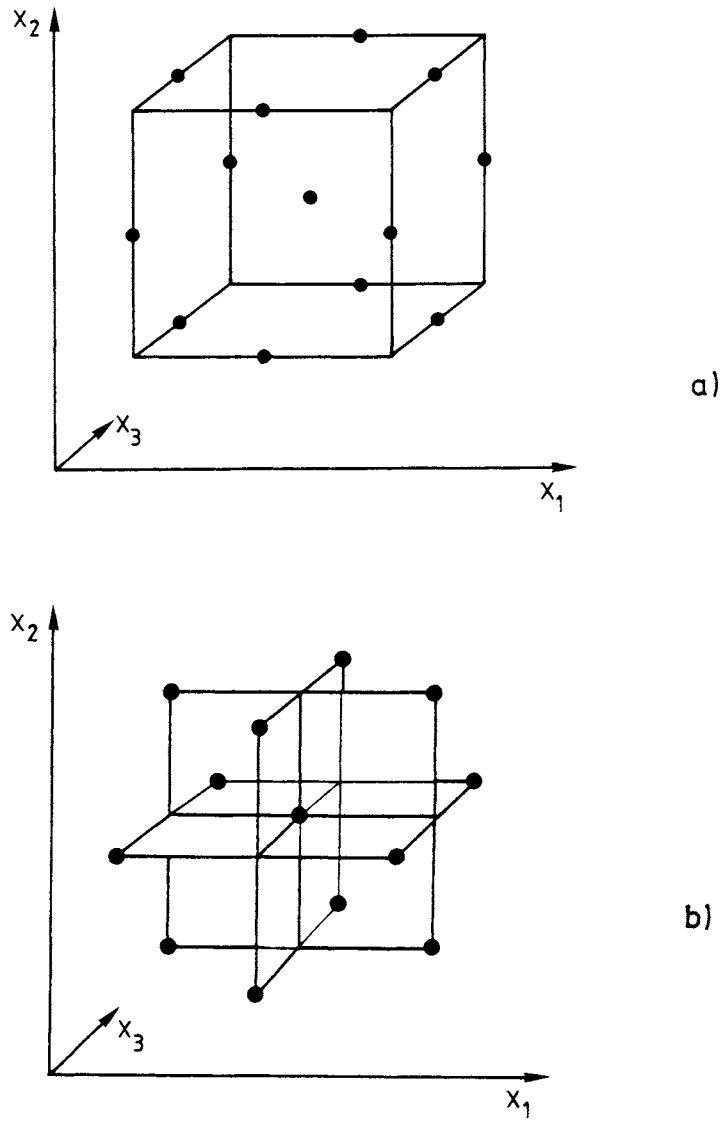


Fig. 24.8. Box–Behnken design. (a) The design, as derived from a cube. (b) Representation as interlocking 2^2 factorial experiments.

although the design can be derived from a cube, it is spherical, so that part of the cubic domain is not covered by the resulting model. The prediction in this part is then an extrapolation, which should be avoided.

TABLE 24.7

The Box–Behnken design for $k = 3$

Expt.	x_1	x_2	x_3
1	+1	+1	0
2	+1	−1	0
3	−1	+1	0
4	−1	−1	0
5	+1	0	+1
6	+1	0	−1
7	−1	0	+1
8	−1	0	−1
9	0	+1	+1
10	0	+1	−1
11	0	−1	+1
12	0	−1	−1
13 etc.	0	0	0

24.3.4 Doehlert uniform shell design

A less well known, but very useful type of design is the uniform shell design introduced by Doehlert [10]. It is sometimes called the Doehlert uniform network or, simply, *Doehlert design*. It describes a spherical experimental domain, but with less points than the central composite design and it stresses uniformity in space filling (see Section 24.2.2). For two factors, the central composite design can be viewed as consisting of one central point and eight points situated at equal intervals on a circle. The Doehlert design for two factors consists of one central point and six points forming a hexagon, i.e. also situated on a circle (see Fig. 24.9a). In three dimensions it consists of a centred dodecahedron (Fig. 24.9b). It can be verified that the distances between all neighbouring experimental points in a Doehlert design are the same.

The respective design matrices are given in Tables 24.8 and 24.9. They are generated as follows. One starts with a simplex in the space considered. For a two-factor space, this means that one starts with an equilateral triangle. The points forming the simplex are labelled S in Fig. 24.9a. Its coordinates are (0,0) (1,0) and (0.5, 0.866) (expts. 1, 3 and 7 in Table 24.8). The other points can be obtained easily using a simple rule. One must subtract each point from each other. Subtraction of point 1 from 3 yields for instance (0.5 – 1, 0.866 – 0) = (– 0.5, 0.866) (point 4) and 3 from 1 (1 – 0.5, 0 – 0.866) = (0.5, – 0.866) (point 5).

To construct the Doehlert design for k factors, one needs the simplex for the same number of factors. The coordinates for $k = 3$ to 10 are given in Chapter 26, where the simplex is discussed. Let us apply this for $k = 3$. From Section 26.2.2 it

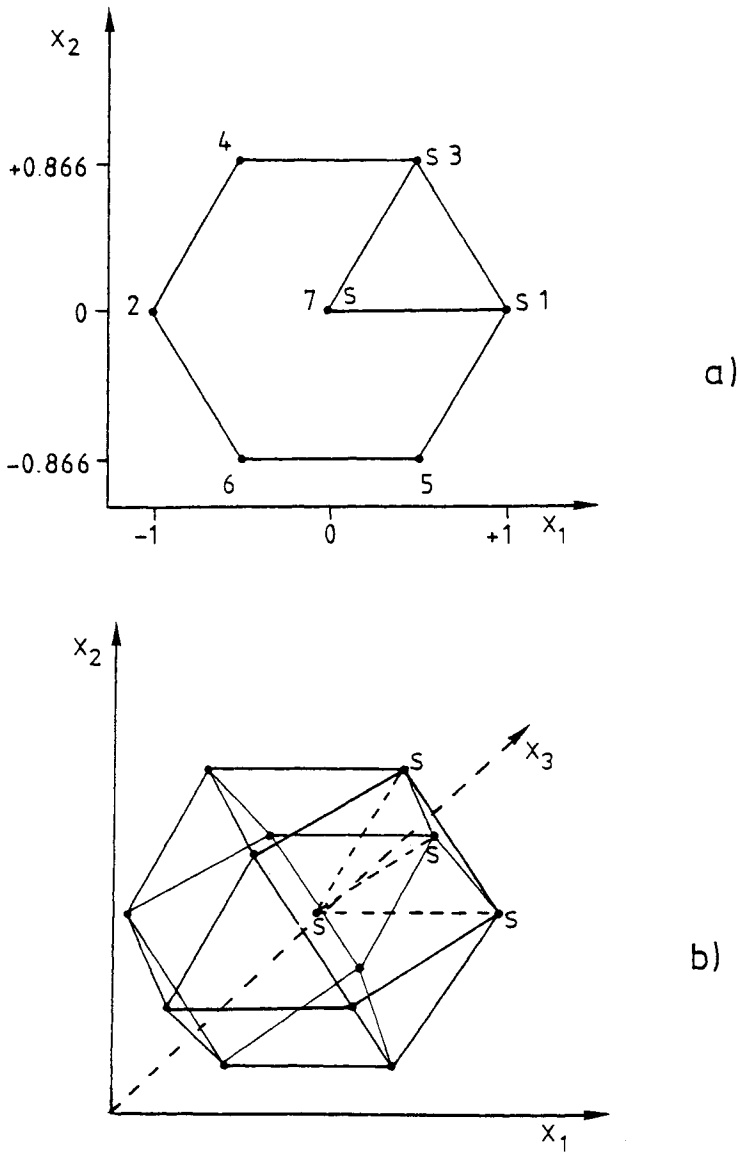


Fig. 24.9. The Doehlert design. The points indicated with S constitute the generating simplex for (a) two variables, (b) three variables.

is derived that the coordinates of the vertices of this simplex are (0, 0, 0); (1, 0, 0); (0.5, 0.866, 0); (0.5, 0.289, 0.817). In Table 24.9, they constitute the first four experiments. All other experiments can be derived by pairwise subtractions of the coordinates of one point from the other, for instance: 0.5 – 0.5 = 0; 0.866 – 0.289 = 0.577; 0 – 0.817 = – 0.817 (point 10).

TABLE 24.8

Doehlert design for 2 factors

Expt.	x_1	x_2
1	1	0
2	–1	0
3	0.5	0.866
4	–0.5	0.866
5	0.5	–0.866
6	–0.5	–0.866
7	0	0

TABLE 24.9

Doehlert design for 3 factors

Expt.	x_1	x_2	x_3
1	0	0	0
2	1	0	0
3	0.5	0.866	0
4	0.5	0.289	0.817
5	–1	0	0
6	–0.5	–0.866	0
7	–0.5	–0.289	–0.817
8	0.5	–0.866	0
9	0.5	–0.289	–0.817
10	0	0.577	–0.817
11	–0.5	0.866	0
12	–0.5	0.289	0.817
13	0	–0.577	0.817

The Doehlert design is not rotatable. This is not surprising, since x_1 is measured at a different number of levels compared to x_2. For the two-factor design there are 5 levels for x_1 (– 1, – 0.5, 0, 0.5, 1) and 3 for x_2 (– 0.866, 0, 0.866). There are, however, several advantages. A first advantage is that the design is efficient, where efficiency is defined as the number of b-coefficients estimated divided by the number of experiments. Table 24.10 describes the efficiency of the central composite design compared to the Doehlert design. For all k, the Doehlert design is more efficient. The Doehlert design is also more efficient in mapping space: adjoining hexagons can fill a space completely and efficiently, since the hexagons fill space without overlap.

TABLE 24.10

Comparison of efficiency of central composite and Doehlert design

k	Number of b coefficients (p)	Number of expts. with central composite* (fc)	Number of expts. with Doehlert (fd)	p/fc	p/fd
2	6	9	7	0.67	0.86
3	10	15	13	0.67	0.77
4	15	25	21	0.60	0.71
5	21	43	31	0.49	0.68
8	45	273	73	0.16	0.62

* One centre point only.

Apart from its uniformity characteristics, the most important advantage is, however, its potential for sequentiality. For instance, one can easily re-use experiments when the boundaries were not well chosen. Suppose one has carried out a Doehlert design and, from the responses one has observed, that it would be useful to investigate outside the original area in the direction of the arrow in Fig. 24.10. One would then carry out experiments 8, 9 and 10 and obtain a new hexagon in that direction. Similarly, it is possible to look at two variables first and add the third afterwards, if this is felt necessary. Indeed, in Table 24.9 one observes that there are 7 experiments with $x_3 = 0$. Their x_1 and x_2 levels coincide with the levels for the $k = 2$ design. One could therefore carry out a design for x_1 and x_2 with 7 experiments at a constant $x_3 = 0$. If a sufficiently good response is obtained in this way, one can stop here. Otherwise, one might consider these 7 experiments to be part of a $k = 3$ design and add points 4, 7, 9, 10, 12 and 13. These points convert the circle described by the $k = 2$ hexagon into the sphere described by the $k = 3$ cubooctahedron.

In a number of cases it is difficult to make decisions at the start of the investigation about certain aspects. For instance, it may be difficult to decide on the variables to include and on the experimental boundaries (should one limit the pH range to 3–7 or rather to 5–7?). As multi-level designs are rather costly in the number of experiments to be carried out, this decision has important consequences. One could then decide on sequential strategies. For instance, one could decide to investigate first the pH domain 5 to 7 and, depending on the results obtained, decide later whether or not to include also the domain between 3 and 5. Or, one could decide to optimize a chromatographic experiment with as variables solvent strength and pH and decide later to add ionic strength or temperature. In all those cases, one would not like to have to start a completely new set of experiments in the second or later stages of the investigation but prefer to be able to re-use as many experiments as possible from the first or prior stages.

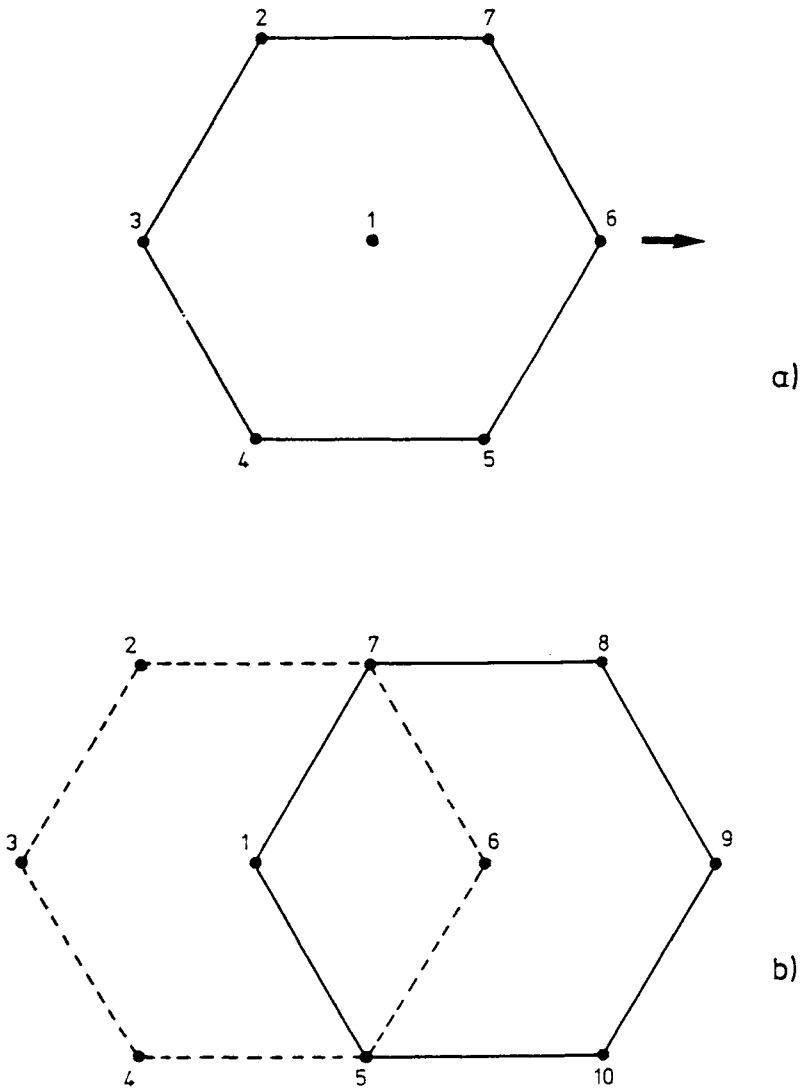


Fig. 24.10. Sequentiality in space. (a) The first Doehlert design was carried out and the results indicated that there may be an optimum outside the original experimental domain in the direction of the arrow. (b) Experiments 8, 9, 10 are added to obtain a new design in that direction.

24.4 Non-symmetrical designs

24.4.1 D-optimal designs

It may happen that it is not possible to carry out one of the classical symmetric designs of the preceding section. Bourguignon et al. [12], for instance, have

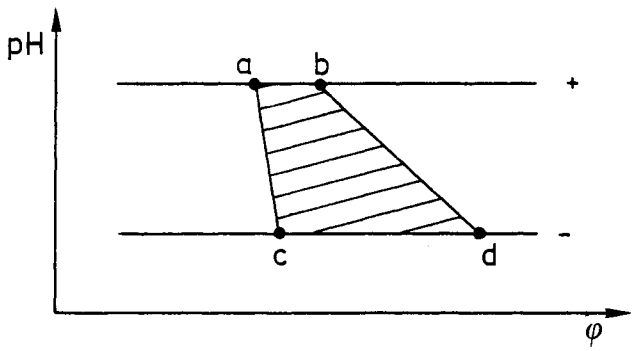


Fig. 24.11. Non-symmetrical experimental domain. Preliminary HPLC experiments at the pH – and + level indicate that the percentage of methanol (φ) at which adequate retention is obtained is included between respectively a and b and c and d. Lines between a and c and b and d delimit the feasible experimental area for subsequent separation optimization.

described irregular experimental domains in chromatography, when optimizing at the same time pH and percentage of methanol in the mobile phase. The possible experimental domain is delimited first by a retention boundary map (see Fig. 24.11), i.e. one determines with a few experiments the area in which it is possible to have suitable retention. The resulting area can be very irregular in form.

A common reason for irregularly shaped experimental regions is that one of the combinations of extreme levels of the variables is practically not possible. For instance, if a reaction is being studied, the combination of high concentration of the reactant and high temperature may make the reaction explosive. As an example suppose that the experimental domain to be investigated is rectangular with boundaries of x_1 and x_2 equal to + 1 and – 1. Then a logical choice would be the 3^2 design. However, suppose also that the experiment with $x_1 = 1$ and $x_2 = 1$ is not possible and that this leads to a constraint $x_1 + x_2 \leq 1$ (see Fig. 24.12a) [13]. One might try to include the 3^2 design within the practically possible experimental domain, but this would exclude from consideration an important part of the domain (Fig. 24.12b). To avoid this, one selects a number of candidate points, for instance using a grid over the whole experimental domain (Fig. 24.12c). From this candidate set, one then has to select a number of experiments. If the decision was to select 8 experiments, one would then compute D (or some other quality criterion from Section 24.2) and select the set of $n = 8$ experiments with smallest D. In Fig. 24.12d the resulting selection is shown. Of course, computing D for all possible combinations of 8 out of 27 experiments (more than 2 million) would require much computation time and so-called *exchange algorithms* have been proposed to shorten this time [14, 15]. Genetic algorithms (Chapter 27) can also be applied [16]. As explained further, one should determine whether the number of experiments selected allows to determine the model sufficiently well.

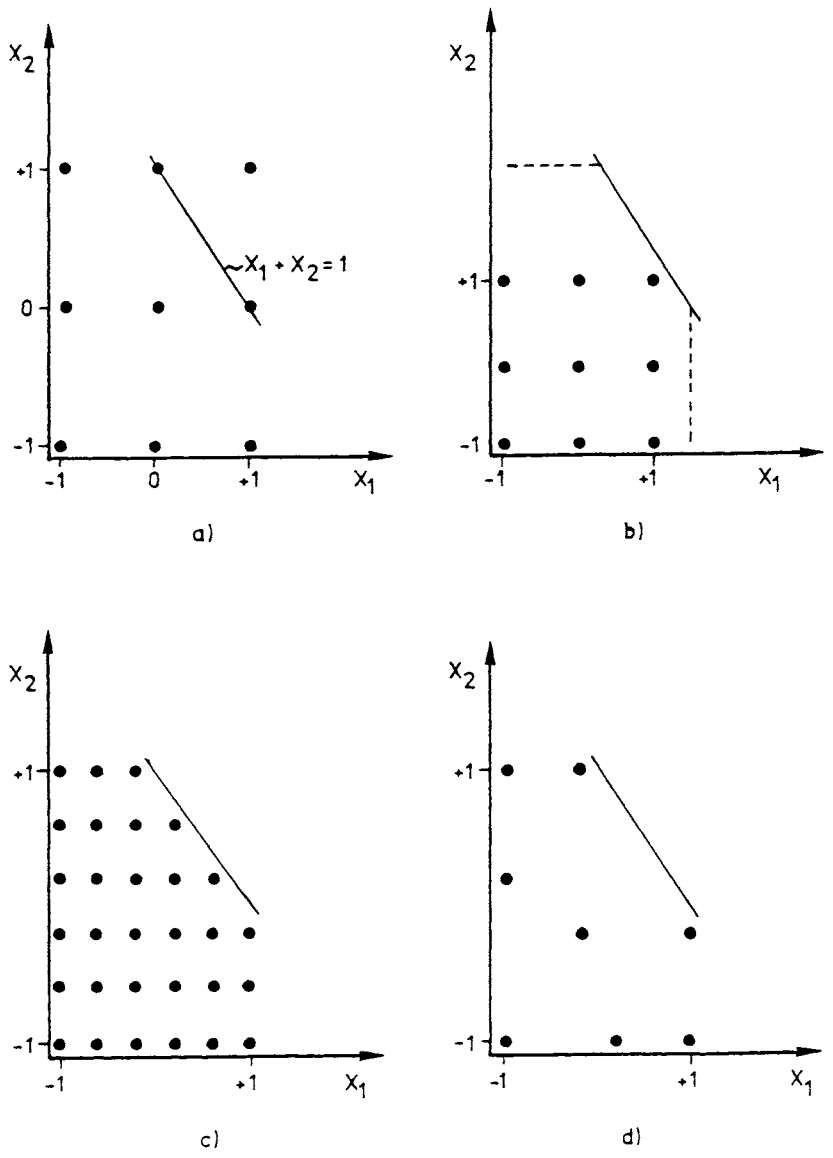


Fig. 24.12. D-optimal design selection (adapted from Ref. [13]). (a) A 3^3 design in a rectangular domain; (b) a 3^3 design in an asymmetrical domain; (c) 27 candidate points; (d) the 8 experiments selected from the 27 candidate points.

It is interesting to consider the solution of Fig. 24.12d in somewhat more detail. In the preceding chapters and sections, we have seen that confidence intervals for the true b-coefficients are smallest when the experiments are situated as far from each other as possible. This conclusion was already reached in Section 8.2.4 on

univariate regression. This principle is systematically applied in experimental design. Indeed, in Chapter 22, except for a possible centre point, the experiments are situated on the corners of the experimental domain (a square, cube or hypercube) and in Section 24.3, the experiments are situated most often on a circle, a sphere or a hypersphere and again these experiments are situated on the boundaries of the experimental domain. The D-optimal design selected here also consists of points that characterize the boundary (and a central point).

Other non-standard situations are those where qualitative variables have to be tested at many different levels. Broudiscou et al. [16] describe such a situation in which 6 factors are investigated at 3 to 7 levels for each factor, in total 30 levels, with 28 experiments.

One of the interesting features of D-optimal designs is the flexibility they give. Not only do they allow to work in experimental domains that are not cubical or spherical, but also one can impose that certain experiments must be included (for instance, because in a preparation phase certain experiments were already carried out) and then compute which additional points are needed to complete a design.

The D-optimality strategy is also used for the selection of experiments when the levels are set by nature. Let us explain this with an example from organic chemistry. The example was described by Phan-Tan-Luu et al. [6]. They describe the influence of the substituent on the reaction rate in the Menschutkin reaction. The factors are the inductive effect, σ_I, and the resonance effect, σ_R, of the substituents. Fourteen substituents are chosen. They are given in Table 24.11. The experimenter would prefer not

TABLE 24.11

Factor levels for the study of the Menschutkin reaction [7]

No.	Substituent	σ_I	σ_R
1	tBu	–0.07	–0.18
2	Et	–0.05	–0.23
3	Me	–0.04	–0.25
4	H	0	0
5	Vinyl	0.05	–0.21
6	NMe_2	0.06	–1.75
7	Phenyl	0.10	–0.30
8	NHCOMe	0.26	–0.86
9	OMe	0.27	–1.02
10	COMe	0.28	0.16
11	CO_2Me	0.30	0.14
12	Br	0.44	–0.30
13	Cl	0.46	–0.36
14	CN	0.56	0.13

TABLE 24.12

Best (D-optimal) experiments for a first order model with interactions for the Menschutkin reaction [7]

n	$\det(\mathbf{X}^T\mathbf{X})$	Combination
4	0.061	1 – 6 – 14 – 4
5	0.121	1 – 6 – 14 – 4 – 13
6	0.389	1 – 6 – 14 – 4 – 13 – 2
7	0.339	1 – 6 – 14 – 4 – 13 – 2 – 9
8	0.487	1 – 6 – 14 – 4 – 13 – 2 – 9 – 12
9	0.662	1 – 6 – 14 – 4 – 13 – 2 – 9 – 12 – 3
10	0.850	1 – 6 – 14 – 4 – 13 – 2 – 9 – 12 – 3 – 10

to have to carry out the reaction with all 14 substituents for reasons of economy: an experimental design with the minimum of experiments involved is desired. However, it is not possible to freely choose levels of σ_I and σ_R, so as to conform with the designs given above. The levels are not determined by the experimenter, but by nature. All the experimenter can do is to make a selection from the 14 possible experiments.

In Table 24.12 the selection of $n = 4$, 5, etc. experiments based on the D-optimality criterion is given for the first-order model with interaction. For instance, the D-optimal selection for $n = 4$ is obtained by selecting substances 1, 6, 14 and 4. In Fig. 24.13 one observes that one selects experiments that characterize the boundary of the experimental region. Before starting this selection it should be verified whether a solution with sufficient quality can be obtained. One way to do this is to compute the inflation factors. In Section 10.5 it was explained that the inflation factors (VIF) allow to decide whether b-coefficients can be estimated sufficiently well. For the first-order model with interaction and for all 14 substances VIF (x_1) = 1.35, VIF (x_2) = 2.36, VIF (x_1x_2) = 2.50, which is below the rejection limit of VIF = 5. For a second-order model VIF (x_1) = 20.0, VIF (x_2) = 64.6, VIF (x_1x_2) = 14.7, VIF (x_1^2) = 13.3, VIF (x_2^2) = 36.0; so that one concludes that the design including all 14 compounds does not allow to obtain a sufficiently good quadratic model.

24.4.2 Uniform mapping algorithms

As was already explained in Section 24.2.2, in cases where no model is known or where it is known but D-optimality and related criteria are not easily computed (e.g. non-linear models), one prefers uniform spacing algorithms. In Section 24.3.4 the Doehlert uniform shell design was proposed for filling in a uniform way a spherical domain. In a non-symmetrical region a uniform mapping algorithm such

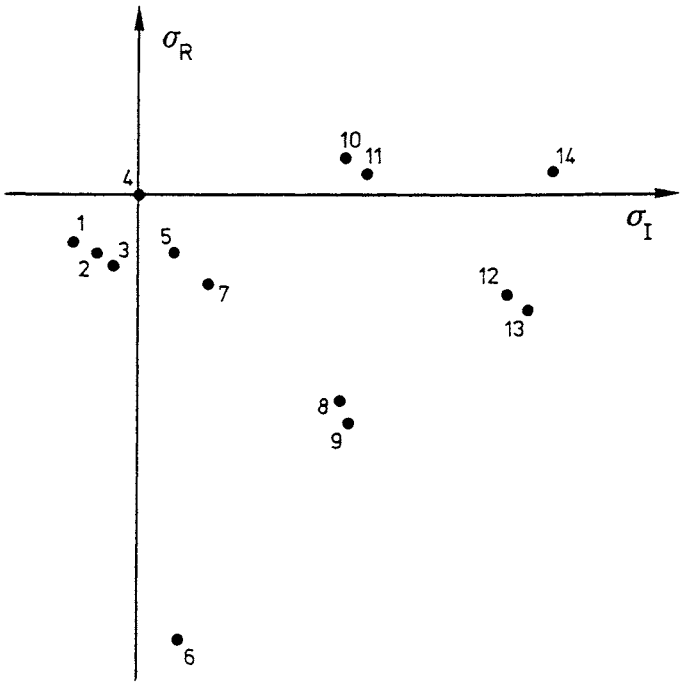


Fig. 24.13. The 14 experiments of Table 24.11 in the σ_I–σ_R plane.

as that of Kennard and Stone [17] can be applied. Their reasoning can be understood as follows. Different experiments chosen should contain different information. They should therefore not occupy similar locations in the multidimensional space described by the factors. On the contrary, one would like to cover the space as uniformly as possible, making sure at the same time that the experiments are as far from each other as possible. The Kennard and Stone algorithm consists of maximizing the minimal distance between each selected point and all the others. The distance is the Euclidean distance (see Chapters 9 and 30) and is given by:

$$d_{ij} = \sqrt{\sum_{l=1}^{k} (x_{il} - x_{jl})^2} \tag{24.6}$$

where l (ranging from 1 to k) identifies the variables and i and j identify the two points. An example of how to compute the distance between two points in multidimensional space is given in Chapter 30.

One can initiate the algorithm in two ways. In practical instances, one might already have performed some experiments and could decide to include these as starting points. When this is not the case, one would determine the distance between all pairs of points and select the largest one

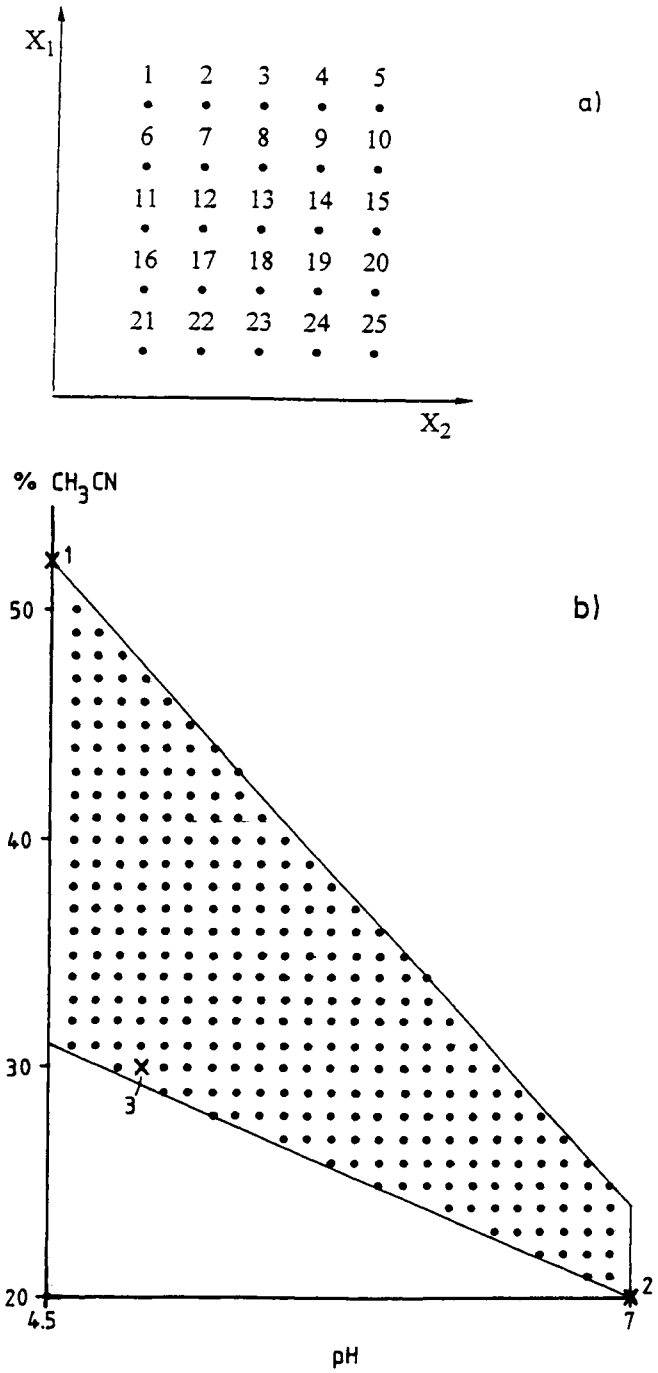


Fig. 24.14. Kennard and Stone design selection. (a) From a grid of possible experiments in a rectangular domain; (b) from a grid of possible experiments in an irregular domain.

$d_{\text{selected}} = \max\ (d_{ij})$

This leads to the selection of the first two points. In the example of Fig. 24.14a, this would result in the selection of either 1 and 25 or 21 and 5. Let us opt for the former combination. One now enters consecutively additional points by computing for each point i_0 not yet selected the distance to each selected point i and select the i_0 for which

$$d_{\text{selected}} = \max_{i_0} \left(\min_{i} (d_{i,i_0})\right) \tag{24.7}$$

In words, one measures the distance for a not included i_0 to each already selected i and determines to which i it is closest ($\min_i (d_{i,i_0})$). This distance is obtained for all i_0 and the point is included for which it is highest: one maximizes the distance to the closest point already included. In Fig. 24.14a this would lead to the inclusion of point 21 or 5. If 5 were chosen and the whole operation repeated, one would then select 21. A possible series of selected points, would then be

1, 25, 5, 21, 13, 3, 11, 15, 23, 19, etc.

This makes sense since the first 4 together form the 2^2 factorial best suited to describe this experimental domain, the first 5 form a centred 2^2 factorial, the first 9 the 3^2 design (which for $k = 2$ is also the face-centred central composite), which would be the most sensible way of choosing an experimental design for a second order model.

In Fig. 24.14b, we see how to apply the algorithm in an irregular domain. The first two points to be selected would be points 1 and 2, while the third point to be added would be point 3. It should be added that several variants of the algorithm can be distinguished. For instance, one could decide that one needs the centre point and select this as point 1.

24.5 Response surface methodology

Let us consider an example from Morgan et al. [8]. They want to optimize burner height (x_H) and lamp current (x_L) with as criterion the signal-to-noise ratio (y) of an atomic absorption spectrophotometer. They decided to apply a central composite design. Instead of applying the rotatable design of Table 24.3, they preferred an orthogonal design with $\alpha = 1.267$ and 5 centre points. The design is shown in Table 24.13. The information matrix is then given by:

$$\mathbf{X} = \begin{array}{c} \begin{matrix} x_0 & x_H & x_L & x_H^2 & x_L^2 & x_H x_L \end{matrix} \\ \begin{bmatrix} +1 & -1 & -1 & +1 & +1 & +1 \\ +1 & +1 & -1 & +1 & +1 & -1 \\ +1 & -1 & +1 & +1 & +1 & -1 \\ +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & -1.267 & 0 & +1.605 & 0 & 0 \\ +1 & +1.267 & 0 & +1.605 & 0 & 0 \\ +1 & 0 & -1.267 & 0 & +1.605 & 0 \\ +1 & 0 & +1.267 & 0 & +1.605 & 0 \\ +1 & 0 & 0 & 0 & 0 & 0 \\ +1 & 0 & 0 & 0 & 0 & 0 \\ +1 & 0 & 0 & 0 & 0 & 0 \\ +1 & 0 & 0 & 0 & 0 & 0 \\ +1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

and

$$\mathbf{X}^T\mathbf{X} = \begin{array}{c} \begin{matrix} b_0 & b_H & b_L & b_{HH} & b_{LL} & b_{HL} \end{matrix} \\ \begin{bmatrix} 13 & 0 & 0 & 7.21 & 7.21 & 0 \\ 0 & 7.21 & 0 & 0 & 0 & 0 \\ 0 & 0 & 7.21 & 0 & 0 & 0 \\ 7.21 & 0 & 0 & 9.154 & 4 & 0 \\ 7.21 & 0 & 0 & 4 & 9.154 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix} \end{array}$$

(a few errors in the article were corrected). This yields:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.196 & 0.000 & 0.000 & -0.108 & -0.108 & 0.000 \\ 0.000 & 0.139 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.139 & 0.000 & 0.000 & 0.000 \\ -0.108 & 0.000 & 0.000 & 0.194 & 0.000 & 0.000 \\ -0.108 & 0.000 & 0.000 & 0.000 & 0.194 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.250 \end{bmatrix} \tag{24.8}$$

and, eventually, this leads to the model:

$$\hat{y} = 87.39 + 8.18\, x_H + 9.05\, x_L - 3.97\, x_H^2 - 18.92\, x_L^2 - 5.00\, x_H x_L \tag{24.9}$$

Much information about the effects can be obtained by visual observation of the results as given in Fig. 24.15a. Comparison of the results for the two-level factorial inscribed in the design shows that, by applying the methods of Chapter 22, one

TABLE 24.13

Central composite design for the optimization of the signal-to-noise ratio (from Ref. [8])

Expt.	Height (x_H)	Lamp current (x_L)	Signal-to-noise y
1	−1	−1	41
2	+1	−1	71
3	−1	+1	64
4	+1	+1	74
5	−1.267	0	76
6	+1.267	0	91
7	0	−1.267	44
8	0	+1.267	75
9	0	0	80
10	0	0	83
11	0	0	97
12	0	0	75
13	0	0	100

expects that both x_H and x_L may have an effect (although this is not evident, because the precision on the measurement as derived from the replicates of the centre point is rather low). The effect of x_L at the high level of x_H seems smaller than that at the low level, so that an interaction is possible. All values at the centre point are higher than those at the corner points, so that curvature in both directions may occur. In summary, it seems possible that all coefficients in eq. (24.9) are significant.

These coefficients can be tested by applying eq. (10.15) or (10.16). We will apply the latter:

$$t = b_i / s_{b_i}$$

To compute s_{b_i}, we apply eq. (10.18)

$$\mathbf{V}_{b_i} = (s_{b_i})^2 = (s_e)^2 (\mathbf{X}^T\mathbf{X})^{-1}$$

In this equation s_e is the experimental standard deviation. This can be obtained from the ANOVA of the multiple regression model described in Section 10.3.1.1: the mean squares, MS, are estimates of variance. If there is no lack-of-fit, then the residual mean square provides a value for s_e^2. The ANOVA results are given in Table 24.14. One finds that the lack-of-fit term is not significant, so that one concludes that $s_e^2 = 86.80$. Incidentally, one can note that the pure error term is obtained from the 5 replicates of the centre point. The pure error mean square could have been used as value for s_e^2. However, this is obtained with only 4 degrees of freedom, while the residual mean squares has 7 degrees of freedom, so that the latter is considered to be the better estimate of experimental variance.

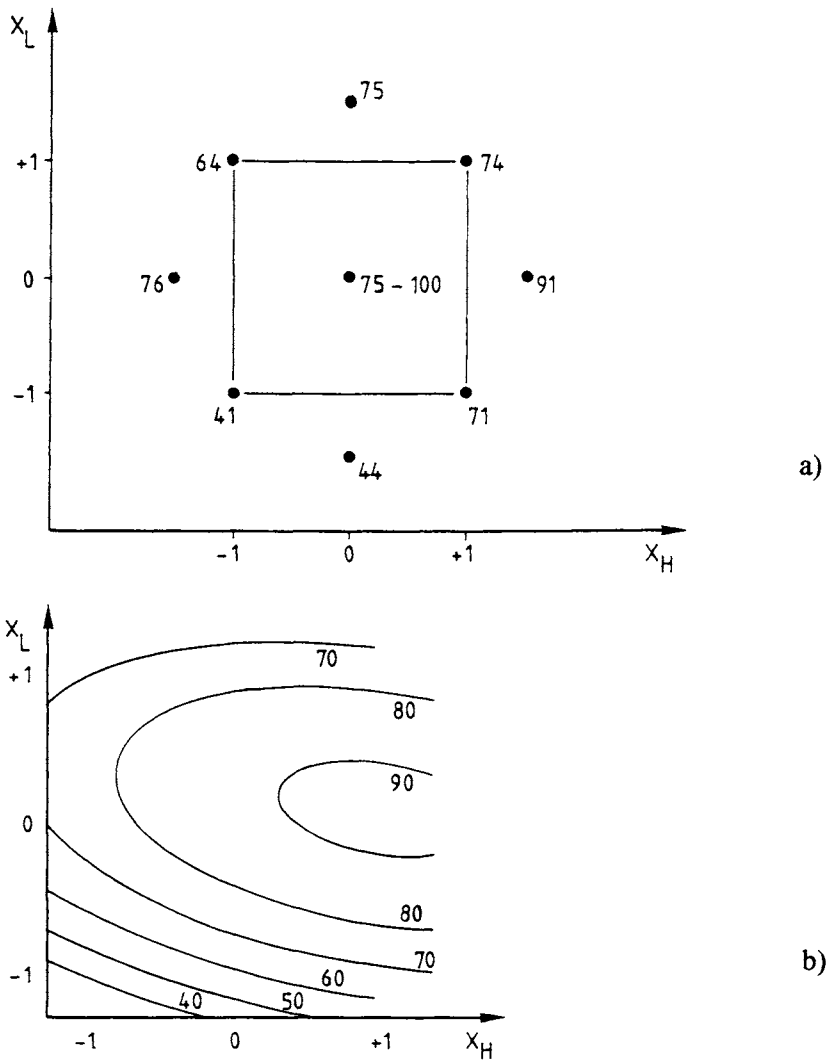


Fig. 24.15. (a) Response for the central composite design of Table 24.13 in the x_H–x_L plane; (b) contour plot for eq. (24.9).

One can now apply eq. (10.18), using matrix (24.8). This yields for s_{b_0}:

$$s_{b_0} = \sqrt{(0.196)(86.8)} = 4.12$$

Similarly, $s_{b_H} = s_{b_L} = 3.47$, $s_{b_{HH}} = s_{b_{LL}} = 4.10$, $s_{b_{HL}} = 4.65$.

The calculated t-value for b_0 is: 87.39/4.12 = 21.2, which is much larger than the tabulated value $t_{0.025,n-p}$ ($n = 13$, $p = 6$) = 2.365; parameter b_0 is therefore significant.

TABLE 24.14

Analysis of variance for the central composite design of Table 24.13

Source of variation	SS	df	MS	F
Due to regression	3101	5	620	7.15
Residual	608	7	86.8	
Lack of fit	129	3	42.9	0.36
Pure error	479	4	119	

It can be verified that this is also the case for all parameters, except b_{HL} and b_{HH}, while the significance of b_H is on the limit. It is instructive to compute $\hat{y}$, first without the x_H^2 term and then without the $x_H x_L$ term. The results are, respectively,

$$\hat{y} = 85.18 + 9.05\, x_L + 8.18\, x_H - 18.92\, x_L^2 - 5.00\, x_H\, x_L$$

and

$$\hat{y} = 85.18 + 9.05\, x_L + 8.18\, x_H - 18.92\, x_L^2$$

Comparison with eq. (24.9) shows that omitting x_H^2 has an effect only on the b_0 term. This could be expected from the $(\mathbf{X}^T\mathbf{X})^{-1}$ matrix, which shows that b_0 and b_{HH} are correlated to some extent. Omitting $x_H x_L$ has no further effect, which is due to the fact that in the $(\mathbf{X}^T\mathbf{X})^{-1}$ matrix all cross terms involving b_{HL} are zero.

Although one can eliminate non-significant terms from the model, this is often not done, because one reasons that the conclusion, that there is no significance, is based on relatively few degrees of freedom and that the b-value obtained is still the best estimate available. For similar reasons, one usually does not apply the step-wise techniques described in Chapter 10.

Additional statistical analysis can be carried out. For instance, one can validate the fit of the model. This is in practice usually done by analyzing the residuals (Section 10.3.1.3), or by testing the significance of the lack-of-fit term in the ANOVA of Table 24.14. The validation of the prediction accuracy requires that additional experiments are carried out, which are then predicted with the model as such. In experimental optimization, the model is often considered only a means to an optimization end and one will usually proceed by selecting optimal conditions from the response surface and restrict prediction validation to these optimal conditions, i.e. carry out the experiment at those conditions and observe whether it fulfils expectations.

The selection of the optimal conditions is often, but not necessarily, done with the aid of visual representation of the response surface or contour plot, describing y as a function of pairs of variables. In the present case, Fig. 25.15b leads us to conclude that the optimal values of x_H and x_L are around + 1 and + 0.2, respectively.

24.6 Non-linear models

In reversed phase chromatography the capacity factor of a substance with mobile phases with different pH follows a sigmoid relationship. An acidic compound for instance, is not ionized at low pH and therefore strongly retained; at high pH it is ionized and badly retained. Considering only pH as a factor leads to the relationship of Fig. 24.16. The location of this sigmoid relationship is described by the pK of the substance. Quadratic models might be adequate over small pH ranges but the relationship can never be modelled with a quadratic curve over its whole range. In such cases, one can apply non-linear regression. Marques and Schoenmakers [2] developed a model for ln k'.

$$\ln k' = \ln(k'_{oa}\, e^{-S\varphi} + k'_{ob}\, e^{-S\varphi}\, 10^{pK_a} \cdot 10^{pH}) - \ln\,(1 + 10^{-pK_a}\, 10^{pH}) \qquad (24.10)$$

where k' = capacity factor, the ln of which is to be modelled as a function of two variables, the pH and φ, the volume fraction of organic modifier; k'_{oa} and k'_{ob} are the capacity factors of fully protonated and dissociated species; S is the so-called solvent strength factor; pK_a has its usual meaning, i.e. the negative logarithm of the acidity constant of the acid being studied. S, k'_{oa}, k'_{ob} and pK_a are generally not known in the mixed organic–aqueous media being studied and must be derived: they are the coefficients to be estimated in the non-linear model.

The designs to be applied are not evident. In principle, when the model is known, as is the case here, one can apply D-optimal designs. However, it turns out that the D-optimality principle here applies to the Jacobian matrix (Section 11.2.3). The practical consequence is that one needs to know the values of the model parameters (here S, etc.) to compute optimal designs [3]. This leads to a circular reasoning: one

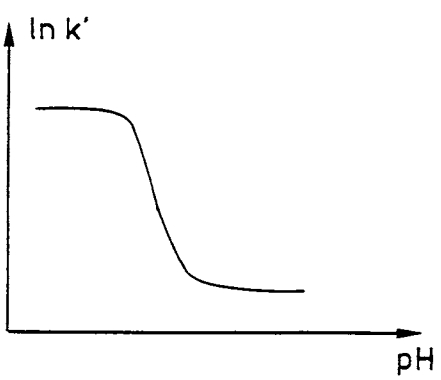


Fig. 24.16. Sigmoid (non-linear) relationship between ln k' and pH.

needs the design to determine the coefficients and the coefficients to determine the optimal design. For that reason, D-optimality is not often applied in this situation, although iterative approaches are possible.

24.7 Latin square designs

A multi-level design of a very different nature is the *Latin square* design. It is used to study a single factor at k levels under circumstances that require blocking (see Chapter 6.9 and 22.9.2). It is convenient to introduce the Latin square method using an example originating from agricultural experimentation. Suppose that five varieties of some economically valuable plant are to be compared in terms of their yields. Planting the five varieties in five plots next to each other may lead to error because the location of the plot may influence the result. To avoid this effect, the field is divided into 25 plots arranged in five rows and five columns. The varieties are planted so that they appear once in each row and once in each column. If the varieties are called a, b, c, d and e, this could lead to the following design

d c e b a
c a d e b
a e b c d
e b a d c
b d c a e

The row and the column direction are two factors, that are blocked, while the effect of the variety is the factor that one wants to study. The same design can be applied without a geometrical context. Suppose we want to study an extraction yield with different solvents but are concerned that environmental factors related to the days and the instrumentation may confuse the effect. Table 24.15 would then be a possible design. One would measure on day 1 the extraction yield of solvent d with instrument I, of solvent c with instrument II, etc.

TABLE 24.15

A latin square arrangement for measuring an extraction yield with solvents a to e

Day	Instrumentation				
	I	II	III	IV	V
1	d	c	e	b	a
2	c	a	d	e	b
3	a	e	b	c	d
4	e	b	a	d	c
5	b	d	c	a	e

The analysis of the results is carried out by ANOVA. An example for a 3×3 design is given below. The design is the following

	I	II	III
1	a	b	c
2	b	c	a
3	c	a	b

The results of the experiments are given by:

	I	II	III
1	16	7	6
2	6	9	17
3	7	14	8

The grand average, $\bar{y}$, is 10. The average values for the rows, $\bar{y}_i$, the columns, $\bar{y}_j$, and the levels, $\bar{y}_t$, are the following.

1: 9.67	I: 9.67	a: 15.67
2: 10.67	II: 10	b: 7
3: 9.67	III: 10.33	c: 7.33

The estimates of the effects of individual columns and rows and of the treatments are then obtained as $\bar{y}_i - \bar{y}$ for the rows, $\bar{y}_j - \bar{y}$ for the columns and $\bar{y}_t - \bar{y}$ for the factors, i.e.

1: –0.33	I: –0.33	a: 5.67
2: +0.67	II: 0	b: –3
3: –0.33	III: 0.33	c: –2.67

These estimates are used to predict the individual results as:

$$\hat{y}_{ijt} = \text{grand average} + \text{effect row}_i + \text{effect column}_j + \text{effect level}_t$$

$$= \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + (\bar{y}_t - \bar{y})$$

and the residuals as

$$e_{ijt} = y_{ijt} - \bar{y} - (\bar{y}_i - \bar{y}) - (\bar{y}_j - \bar{y}) - (\bar{y}_t - \bar{y}) \tag{24.11}$$

One can now compute the ANOVA. The total number of degrees of freedom is $9 - 1 = 8$, the number of degrees of freedom for the columns 2, for the rows 2 and for the effects 2, so that 2 degrees of freedom are left for the residual. The sums of squares are the following

Rows: $(10 - 9.67)^2 + (10.67 - 10)^2 + (9.67 - 10)^2 = 0.6667$
Columns: $(10 - 9.67)^2 + (10 - 10)^2 + (10.33 - 10)^2 = 0.2178$
Factor: $(15.67 - 10)^2 + (7 - 10)^2 + (7.33 - 10)^2 = 48.27$

The residuals are obtained from eq. (24.11) and the residual sum of squares is given by:

$$\sum_i \sum_j (e_{ijt})^2$$

For instance:

$e_{111} = 16 - 10 - (-0.33) - (-0.33) - 5.67 = 1$

$e_{212} = 6 - 10 - 0.67 - (-0.33) - (-3) = -1.33$

The residual sum of squares is 8.64 and the mean square 4.32. The mean square for factors is 24.13 and $F = 5.59$. The tabulated F-value ($\alpha = 5\%$) for 2 degrees of freedom for the effects and for 2 degrees of freedom for the residual is 19.0, so that there is no significance. To obtain significance in small Latin square designs replication is usually necessary.

To avoid confusion, it should be stressed that these designs have in common with the other designs described in this chapter only the fact that they are multi-level. They are used to determine whether a certain treatment has an effect or not. This effect can be fixed or random (see Chapter 6). The Latin square design is the simplest of a family to which belong the *Graeco–Latin square* (3 blocking variables), the *hyper-Graeco–Latin square* (more than 3 blocking variables) and the *balanced incomplete block designs* (designs with unequal number of rows and columns).

References

1. P.F. de Aguiar, B. Bourguignon, M.S. Khots, W. Penninckx and D.L. Massart, The use of logistic transformation in HPLC optimization. Quim. Anal., 12 (1993) 177–182.
2. R.M.L. Marques and P.J. Schoenmakers, Modelling retention in reversed-phase liquid chromatography as a function of pH and solvent composition. J. Chromatogr., 592 (1992) 157–182.
3. A.C. Atkinson, Beyond response surfaces: recent developments in optimum experimental design. Chemom. Intell. Lab. Syst., 28 (1995) 35–47.
4. A.C. Atkinson and A.N. Donev, Optimum Experimental Designs. Clarendon Press, Oxford, 1994.
5. G.E.P. Box and N.R. Draper, Empirical Model-Building and Response Surfaces. Wiley, New York, 1987.
6. R. Phan-Tan-Luu, COMETT Course/ Eguilles, 1988.
7. R.M. Frantz, J.E. Browne and A.R. Lewis, in: Pharmaceutical Dosage Forms — Disperse Form. H.A. Lieberman, M.M. Rieger and G.S. Banker (Eds.). Marcel Dekker, New York, 1988, p. 485.
8. E. Morgan, K.W. Burton and P. Church, Chemom. Intell. Lab. Syst., 5 (1989) 283–302.
9. P.F. de Aguiar, B. Bourguignon, M.S. Khots, D.L. Massart and R. Phan-Tan-Luu, D-optimal designs. Chemom. Intell. Lab. Syst., 30 (1995) 199–210.

10. D.H. Doehlert, Uniform shell designs. Appl. Statist., 19 (1970) 231–239.
11. R. Marchetti and M.E. Guerzoni, Study on the effects of selected mineral nutrients and of their interactions on the yeast fermentation performance. Use of an experimental design. Cerevisia Biotechnol., 16(1) (1991) 24–33.
12. B. Bourguignon, P.F. de Aguiar, M.S. Khots and D.L. Massart, Optimization in irregularly shaped regions: pH and solvent strength in reversed-phase HPLC. Anal. Chem., 66 (1994) 893–904.
13. R.S Discover Statistical Appendices, BBN software. Cambridge, MA, 1988.
14. M.E. Johnson and C.J. Nachtsheim, Some guidelines for constructing exact D-optimal designs on convex design spaces. Technometrics, 25 (3) (1983) 271–277.
15. V.V. Federov, Theory of Optimal Experiments. Academic Press, New York, 1972.
16. A. Broudiscou, R. Leardi and R. Phan-Tan-Luu, Genetic algorithms as a tool for selection of D-optimal design. Chemom. Intell. Lab. Syst., 35 (1996) 87–104.
17. R.W. Kennard and L.A. Stone, Computer-aided design of experiments. Technometrics, 11 (1969) 137–148.

Chapter 25

Mixture Designs

25.1 The sum constraint

In many cases we need to optimize the composition of mixtures. Typical situations are those where different excipients are mixed to obtain optimal characteristics of a tablet, such as hardness or dissolution time, or the optimization of HPLC separations by finding the best solvent composition of the mobile phase. To introduce the problem we will consider a simple example due to Phan-Tan-Luu and colleagues [1]. French wines are often made by mixing ("assemblage") of wines from different grape varieties ("cépages"). The Côteaux d'Aix wines are assembled from Cabernet (C), Syrah (S) and Grenache (G). Suppose now that one has those three wines available and one is set the task to find the optimal composition from the sensory point of view. An experimental design approach would consist of making a restricted set of mixtures that map the experimental domain well. At first sight one could consider the %C, %S and %G as the three factors and apply, for instance, a 2^3 design with as levels 0 and 100%. This would require (see also Fig. 25.1) the combinations of Table 25.1.

Clearly, several of these compositions are impossible. The first one, for instance, would consist of 0% in total and the last one of 300%. Of course, only

TABLE 25.1

The impossible factorial design for optimizing a wine with three grape varieties (C = Cabernet, S = Syrah, G = Grenache)

C (%)	G (%)	S (%)
0	0	0
0	0	100
0	100	0
0	100	100
100	0	0
100	0	100
100	100	0
100	100	100

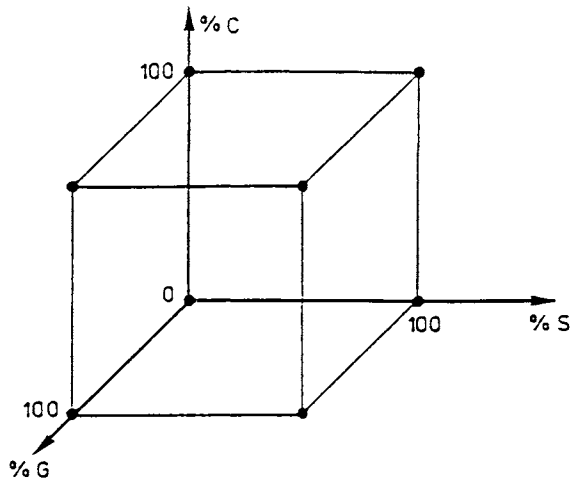


Fig. 25.1. The wine assemblage problem treated as an (impossible) 2^3 design.

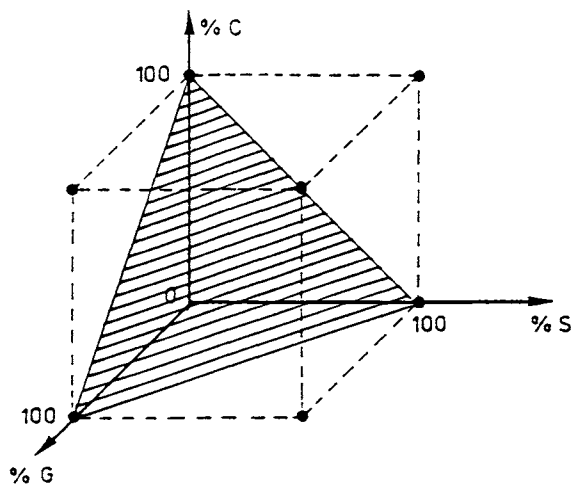


Fig. 25.2. The triangle is the domain in which the mixture experiments of the wine assemblage experiments are possible.

those combinations are possible where the sum of the components is exactly equal to 100%. This constraint is typical for mixture problems. The feasible domain is shown in Fig. 25.2. The constraint that the sum should be 100% has other consequences. Consider for instance an equation such as in Chapter 24:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3$$

This equation is not well suited for mixtures. It is easy to verify that, for instance, there is no need for a b_0 term. To do this let us consider the simpler first-order model $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$. Since $x_1 + x_2 + x_3 = 1$, one can re-write this as

$$y = b_0 (x_1 + x_2 + x_3) + b_1x_1 + b_2x_2 + b_3x_3$$

$$= (b_0 + b_1) x_1 + (b_0 + b_2) x_2 + (b_0 + b_3) x_3$$

$$= b'_1x_1 + b'_2x_2 + b'_3x_3$$

Other consequences for the response surface equation will be described in Section 25.3.

25.2 The ternary diagram

The domain in Fig. 25.2 where experiments can be carried out is an equilateral triangle. This is preferably represented as a ternary diagram. An example of a *trilinear* or *ternary diagram* is shown in Fig. 25.3. Each corner of the diagram consists of 100% of one of the three components or, expressed in fractions, $x_i = 1$ ($i = 1$, 2 or 3). The sides represent binary mixtures. The x_1-coordinate takes on values from $x_1 = 0$ at the bottom to $x_1 = 1$ at the apex, the x_2-coordinate runs from the right side to the bottom left corner and the x_3-coordinate runs from 0 at the apex

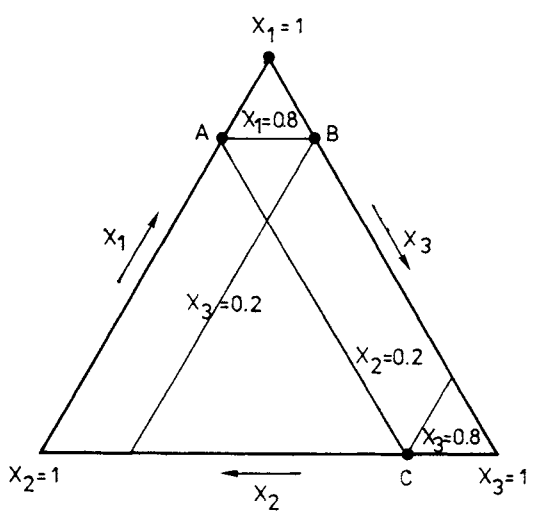


Fig. 25.3. A ternary diagram with binary mixtures A ($x_1 = 0.8$, $x_2 = 0.2$), B ($x_1 = 0.8$, $x_3 = 0.2$) and C ($x_2 = 0.2$, $x_3 = 0.8$).

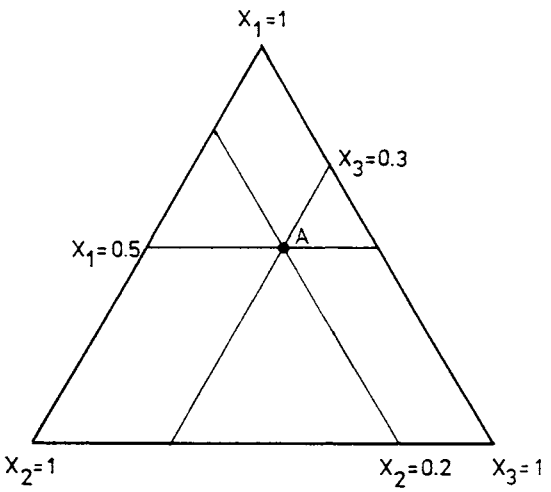


Fig. 25.4. A ternary diagram with ternary mixture A ($x_1 = 0.5, x_2 = 0.2, x_3 = 0.3$).

to 1 in the bottom right corner. Point A represents then a mixture of $x_1 = 0.8$, and since it is situated on the axis between x_1 and x_2, $x_2 = 0.2$ and $x_3 = 0$. In point B, the amount of $x_3 = 0.2$ and $x_2 = 0$.

It is somewhat more difficult to derive the compositions for ternary mixtures. One notices that points A and B both have $x_1 = 0.8$. By connecting them one obtains a line parallel to the bottom side of the triangle. All mixtures on that line have $x_1 = 0.8$ and in the same way other lines parallel to the bottom line can be drawn for other values of x_1. In the same way lines parallel to the right side of the triangle give mixtures of the same composition in x_2 (in the figure $x_2 = 0.2$). It can be verified that point A is $x_1 = 0.8$ and $x_2 = 0.2$, since it falls on the intersection of the lines describing such compositions. Compositions of equal x_3 (in the figure $x_3 = 0.2$) are given by lines parallel to the left side of the triangle.

It is now possible to depict ternary mixtures. Figure 25.4 shows the lines for $x_1 = 0.5, x_2 = 0.2, x_3 = 0.3$. The intersection point A of the three lines is the mixture with that composition.

It is useful to note that the line connecting a particular point on a side of the triangle to the apex opposite to it describes constant proportions of two components. Figure 25.5 is a ternary diagram for mixtures of the microcrystalline cellulose avicel, α-lactose monohydrate and water [2]. The response is a quality parameter for pellets made with the mixtures. One observes that there is a better chance of obtaining good quality pellets around a line with a ratio water:Avicel = 43:57. When there is more than 50% lactose present the quality degrades.

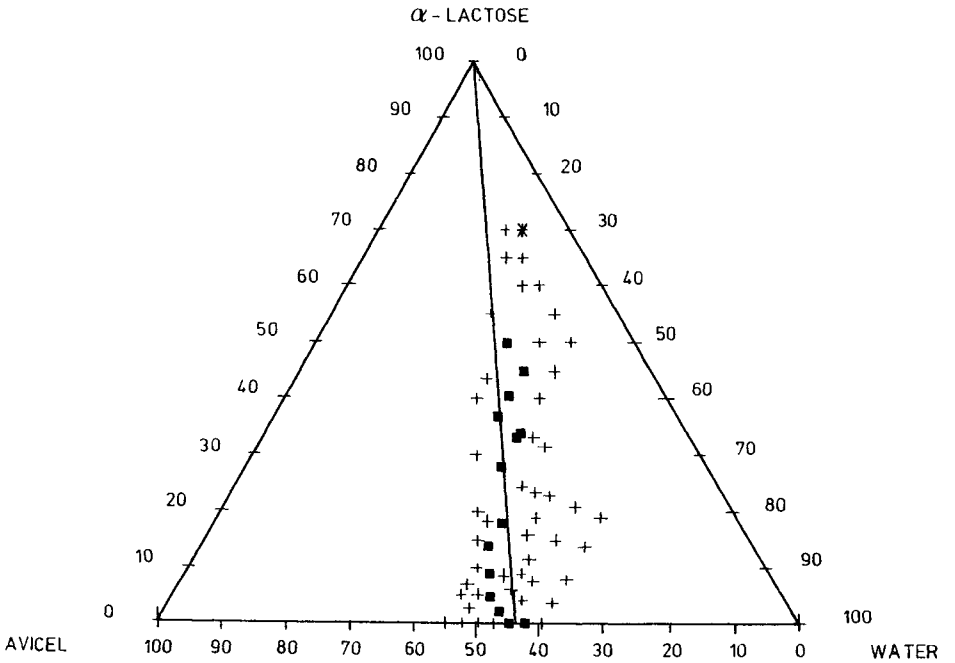


Fig. 25.5. A ternary diagram with the quality of pellets as a function of the components of a tablet (■ good quality, + not acceptable, * non pelletizable mixtures). Adapted from Ref. [2].

25.3 Introduction to the Simplex design

By far the most often used designs are Simplex designs. For a three-component mixture, a Simplex is a triangle, for a four-component mixture, it is a tetrahedron, etc. A definition of the general term simplex is given in Section 26.2.2 and a more precise terminology of Simplex mixture designs will be introduced in later sections of this Chapter. In this section an intuitive introduction is given. It should be noted here immediately that mixture design is a very traditional field, with customs of its own. This section intends to give a flavour of it. A more detailed account is given in Section 25.4.

One of the most often used experimental designs for three components is given in Table 25.2. For reasons that we will explain in the next sections, this is called a (3,3) design. The first 7 points are used to compute the coefficients in eq. (25.1):

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3 \qquad (25.1)$$

This is a typical mixture design equation. It is called the reduced cubic model and is explained in Section 25.4.2.

TABLE 25.2

Three-component simplex design for a wine assemblage optimization

Experiment	x_1	x_2	x_3	Response
1	1	0	0	y_1
2	0	1	0	y_2
3	0	0	1	y_3
4	0.5	0.5	0	y_{12}
5	0.5	0	0.5	y_{13}
6	0	0.5	0.5	y_{23}
7	0.33	0.33	0.33	y_{123}
8	0.67	0.165	0.165	y_8
9	0.165	0.67	0.165	y_9
10	0.165	0.165	0.67	y_{10}

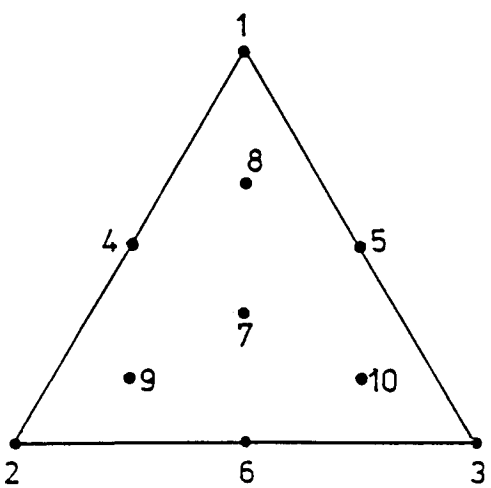


Fig. 25.6. Simplex centroid design. The first 7 points are used to estimate the model, the 3 following to validate it.

The indices of the response refer to the composition of the mixtures. For instance, response y_{123} means that components x_1, x_2, x_3 are present in equal proportions. The distribution of the experimental points in the ternary diagram is shown in Fig. 25.6. Eq. (25.1) contains 7 coefficients so that one requires a minimum of 7 experiments to estimate them and one uses experiments 1 to 7 from Table 25.2 to do so. Table 25.3 explains how this is done for the (3,3) design, which we introduce in this section, and some related designs, that will be described in later sections.

TABLE 25.3

Simplex lattice designs: (3,1), (3,2) and (3,3) reduced designs

The (3,1) design

Experimental points: 1–3 of Table 25.2

Canonical equation: $y = b_1x_1 + b_2x_2 + b_3x_3$

Coefficients: $b_1 = y_1$, $b_2 = y_2$, $b_3 = y_3$

The (3,2) design

Experimental points: 1–6 of Table 25.2

Canonical equation: $y = b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3$

Coefficients: $b_1 = y_1$, $b_2 = y_2$, $b_3 = y_3$

$b_{12} = 4y_{12} - 2(y_1 + y_2)$

$b_{13} = 4y_{13} - 2(y_1 + y_3)$

$b_{23} = 4y_{23} - 2(y_2 + y_3)$

The (3,3) design, reduced cubic model

Experimental points: 1–7 of Table 25.2

Canonical equation: $y = b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3$

$b_1 = y_1$ $\quad b_{12} = 4y_{12} - 2(y_1 + y_2)$

$b_2 = y_2$ $\quad b_{13} = 4y_{13} - 2(y_1 + y_3)$

$b_3 = y_3$ $\quad b_{23} = 4y_{23} - 2(y_2 + y_3)$

$b_{123} = 27y_{123} - 12(y_{12} + y_{13} + y_{23}) + 3(y_1 + y_2 + y_3)$

It is easy to understand how one arrives at the values of Table 25.3. For instance, for $x_1 = 1$, $y = y_1$ and since $x_2 = 0$ and $x_3 = 0$, all the terms in eq. (25.1) except the first become 0, so that:

$$y_1 = b_1 \times 1 = b_1$$

For $y = y_{12}$, $x_1 = 0.5$, $x_2 = 0.5$, $x_3 = 0$, eq. (25.1) yields:

$$y_{12} = b_1 \times 0.5 + b_2 \times 0.5 + b_{12} \times 0.5 \times 0.5$$

and since $b_1 = y_1$ and $b_2 = y_2$:

$$b_{12} = 4y_{12} - 2(y_1 + y_2)$$

In a similar way, one obtains the b_{123} value in Table 25.3.

Equation (25.1) can be used to predict the response over the whole experimental domain, i.e. the whole ternary diagram. To verify that the prediction is good, it is usual in traditional mixture design to carry out some additional experiments. In this case, one would carry out the experiments 8 to 10 and compare the experimentally obtained y_8 to y_{10} with the responses predicted from eq. (25.1).

25.4 Simplex lattice and centroid designs

25.4.1 The (3,2) Simplex lattice design

Simplex mixture designs were first introduced by Scheffé [3]. In many cases the simplex mixture designs proved remarkably successful and they are used very often. The simplex described in the foregoing section is called a *Simplex centroid* design. It will be described in more detail in Section 25.4.3. The original simplex designs are called *Simplex lattice* designs and these will now be discussed further. The *lattice* is the equivalent of the factorial design for process variables, in the sense that experimental points are also taken at the border of the experimental domain and, for more than 2 levels, are evenly spaced along the coordinates representing the factors.

An example is shown in Fig. 25.7b. This is called a (3,2) lattice which is a special case of the general (k, m) lattice. This terminology will now be explained. In the Simplex lattice design each of the k components can take on $m + 1$ equally spaced levels from 0 to 1. For a first-order equation ($m = 1$) the levels are 0 and 1. For a second-order model ($m = 2$) these levels are

$$x_i = 0, \tfrac{1}{2}, 1$$

and more generally:

$$x_i = 0, \frac{1}{m}, \frac{2}{m}, \ldots, 1 \tag{25.2}$$

The first-order model requires that one should form all combinations of 1 and 0. They are (1,0,0), (0,1,0) and (0,0,1), i.e. the three corner points in the triangle, consisting of the three pure components (Fig. 25.7a). This is a rather simple and unchallenging case. Let us therefore consider the second order model. This can be treated with the use of the design of Fig. 25.7b. This is a (3,2) lattice with $k = 3$ components, $3(= m + 1)$ equally spaced experiments along each side of the simplex and in total 6 experiments. They are obtained by performing all the combinations of 0, 0.5 and 1, so that the sum is always 1. The experiments are therefore (1,0,0), (0,1,0), (0,0,1), (0.5, 0.5, 0), (0.5, 0, 0.5) and (0, 0.5, 0.5), i.e. experiments 1 to 6 of Table 25.2. The 6 experiments are sufficient to describe an $m = 2$ polynomial:

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 \tag{25.3}$$

This is the *canonical form* of the general quadratic equation for three components given by eq. (25.1), which we rewrite here as:

$$y = b'_0 + b'_1x_1 + b'_2x_2 + b'_3x_3 + b'_{12}x_1x_2 + b'_{13}x_1x_3 + b'_{23}x_2x_3 + b'_{11}x_1^2 + b'_{22}x_2^2 + b'_{33}x_3^2 \tag{25.4}$$

Consider the term x_1^2. This can be rewritten as:

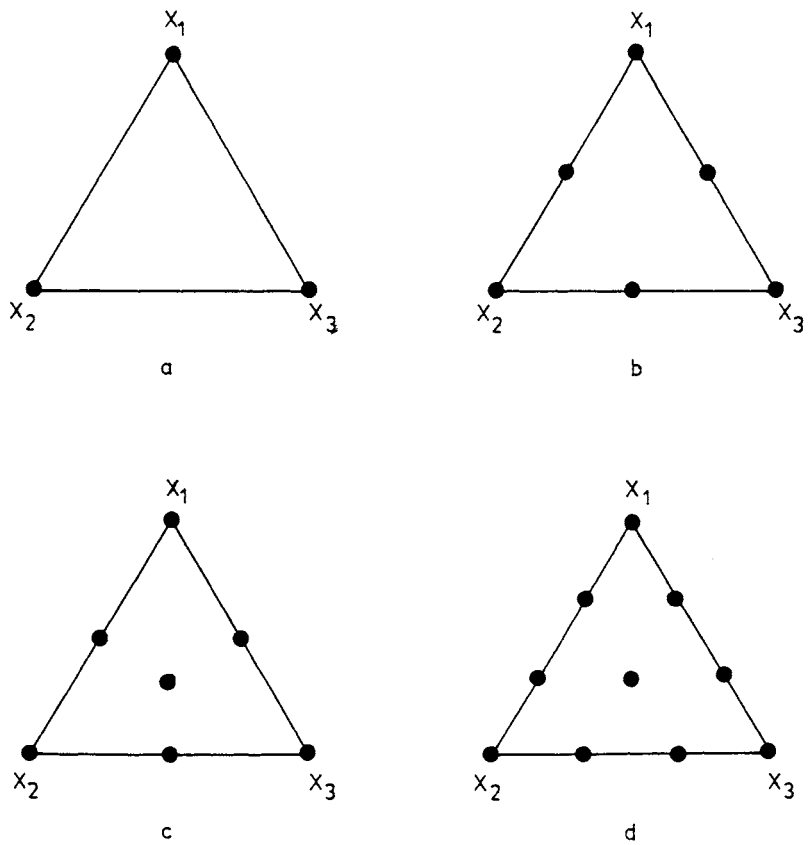


Fig. 25.7. Simplex lattice designs for 3 components: (a) first-order model; (b) second-order model; (c) reduced cubic model; (d) complete cubic model.

$x_1^2 = x_1(1 - x_2 - x_3) = x_1 - x_1x_2 - x_1x_3$, so that

$b'_{11}x_1^2 = b'_{11}x_1 - b'_{11}x_1x_2 - b'_{11}x_1x_3$

This means that one does not need an x_1^2 term, but can include it in the x_1, x_1x_2 and x_1x_3 terms. This is true for all squared terms and also for the b_0 term. The net result is that one reduces the 10-term polynomial of eq. (25.4) to its 6-term canonical form of eq. (25.3). In mixture designs one often determines as many coefficients as there are experiments so that regression is not needed. This has to do with the time at which the theory of mixture design originated. At that time no computers where available, so that much emphasis was put on the simplicity of the calculations. Computations of the b coefficients are made in a way similar to that explained in Section 25.3 (see also Table 25.3).

The reader will have noticed that we do not follow the same path as we did with process variables. With process variables, one often first tries to understand which variables have an effect. In mixture design, this is very difficult. It is easy to understand that one cannot change a mixture variable without changing at the same time at least one other mixture variable. It is also apparent in the equations given above, since e.g. the interaction coefficient b'_{12} is confounded with the coefficients of the quadratic terms b'_{11} and b'_{22}.

25.4.2 (k, m) Simplex lattice designs

Let us now look at Simplex lattice designs in general and first at the (3,3) design. The polynomial equation is:

$$\begin{aligned} y = {} & b'_0 + b'_1x_1 + b'_2x_2 + b'_3x_3 + b'_{11}x_1^2 + b'_{22}x_2^2 + b'_{33}x_3^2 + b'_{12}x_1x_2 \\ & + b'_{13}x_1x_3 + b'_{23}x_2x_3 + b'_{112}x_1^2x_2 + b'_{122}x_1x_2^2 + b'_{113}x_1^2x_3 \\ & + b'_{133}x_1x_3^2 + b'_{223}x_2^2x_3 + b'_{233}x_2x_3^2 + b'_{123}x_1x_2x_3 \\ & + b'_{111}x_1^3 + b'_{222}x_2^3 + b'_{333}x_3^3 \end{aligned} \tag{25.5}$$

From this equation, one can derive two canonical models. The first is called the *complete cubic (canonical) model*:

$$\begin{aligned} y = {} & b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + g_{12}x_1x_2(x_1 - x_2) + \\ & g_{13}x_1x_3(x_1 - x_3) + g_{23}x_2x_3(x_2 - x_3) + b_{123}x_1x_2x_3 \end{aligned} \tag{25.6}$$

The experimental design is described in Table 25.4 and shown in Fig. 25.7d. It is obtained by making all possible combinations between 0, 1/3, 2/3 and 1, so that the sum is always 1. The coefficients are given in Table 25.4.

The second (canonical) model is called the *reduced* or *special cubic model*. It is a simplification of the model of eq. (25.6) and is given by

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3 \tag{25.7}$$

The model is in fact that of eq. (25.1). The experimental design is that of Table 25.2 and Fig. 25.7c and the coefficients are computed as in Table 25.3.

Let us now — without further explanations — apply the principles we have learned for the three-component design to mixtures with four components. The simplex is now a tetrahedron and the (4,1), (4,2) and (4,3) complete and reduced models are given in Tables 25.5 to 25.8 and Figs. 25.8a to d. The (4,1) simplex consists of all combinations with levels 0 and 1, the (4,2) simplex of all possible combinations including 0, 1/2 or 1, the complete (4,3) of all combinations of 0, 1/3,

TABLE 25.4

Simplex lattice (3,3) design: complete cubic design

Expt.	x_1	x_2	x_3	Response
1	1	0	0	y_1
2	0	1	0	y_2
3	0	0	1	y_3
4	0.33	0.67	0	y_{122}
5	0.67	0.33	0	y_{112}
6	0.33	0	0.67	y_{133}
7	0.67	0	0.33	y_{113}
8	0	0.33	0.67	y_{233}
9	0	0.67	0.33	y_{223}
10	0.33	0.33	0.33	y_{123}

Canonical equation: $y = \sum b_i x_i + \sum b_{ij} x_{ij} + \sum g_{ij} x_i x_j (x_i - x_j) + \sum b_{ijk} x_i x_j x_k \qquad 1 \le i < j < k \le 3$

Coefficients: $b_i = y_i$

$b_{ij} = (9/4)\,[y_{iij} + y_{ijj} - (y_i + y_j)]$

$g_{ij} = (9/4)\,[3y_{iij} - 3y_{ijj} - (y_i + y_j)]$

$b_{123} = 27y_{123} - (27/4)(y_{112} + y_{122} + y_{113} + y_{133} + y_{223} + y_{233}) + (9/2)(y_1 + y_2 + y_3)$

TABLE 25.5

The (4,1) Simplex lattice design

Expt.	x_1	x_2	x_3	x_4	Response
1	1	0	0	0	y_1
2	0	1	0	0	y_2
3	0	0	1	0	y_3
4	0	0	0	1	y_4

Canonical equation: $y = b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$

Coefficients: $b_i = y_i$

2/3 and 1 and the reduced (4,3) of all combinations of 0 and 1, all combinations of 0 and 1/2 and all combinations of 0 and 1/3.

For the general case of k components and degree m the number of experiments for a full design is

$$\binom{m+k-1}{m} = \frac{(m+k-1)!}{(k-1)!m!} \tag{25.8}$$

This yields the numbers of experiments of Table 25.9. Of course, the numbers in the lower right corner are impractical, so that such designs are not, or very rarely, applied in practice.

TABLE 25.6

The (4,2) Simplex lattice design

Expt.	x_1	x_2	x_3	x_4	Response
1	1	0	0	0	y_1
2	0	1	0	0	y_2
3	0	0	1	0	y_3
4	0	0	0	1	y_4
5	0.5	0.5	0	0	y_{12}
6	0.5	0	0.5	0	y_{13}
7	0.5	0	0	0.5	y_{14}
8	0	0.5	0.5	0	y_{23}
9	0	0.5	0	0.5	y_{24}
10	0	0	0.5	0.5	y_{34}

Canonical equation: $y = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{14}x_1x_4 + b_{23}x_2x_3 + b_{24}x_2x_4 + b_{34}x_3x_4$

Coefficients: $b_i = y_i$

$b_{ij} = 4y_{ij} - 2(y_i + y_j)$

TABLE 25.7

The (4,3) Simplex lattice design: complete cubic design

Expt.	x_1	x_2	x_3	x_4	Response
1	1	0	0	0	y_1
2	0	1	0	0	y_2
3	0	0	1	0	y_3
4	0	0	0	1	y_4
5	0.33	0.33	0.33	0	y_{123}
6	0.33	0	0.33	0.33	y_{134}
7	0.33	0.33	0	0.33	y_{124}
8	0	0.33	0.33	0.33	y_{234}
9	0.33	0.67	0	0	y_{122}
10	0.67	0.33	0	0	y_{112}
11	0.33	0	0.67	0	y_{133}
12	0.67	0	0.33	0	y_{113}
13	0.33	0	0	0.67	y_{144}
14	0.67	0	0	0.33	y_{114}
15	0	0.33	0.67	0	y_{233}
16	0	0.67	0.33	0	y_{223}
17	0	0.33	0	0.67	y_{244}
18	0	0.67	0	0.33	y_{224}
19	0	0	0.33	0.67	y_{344}
20	0	0	0.67	0.33	y_{334}

Canonical equation: $y = \sum b_i x_i + \sum b_{ij} x_{ij} + \sum g_{ij} x_i x_j (x_i - x_j) + \sum b_{ijk} x_i x_j x_k \qquad 1 \le i < j < k \le 4$

Coefficients: $b_i = y_i$

$b_{ij} = 9/4[y_{iij} + y_{ijj} - (y_i + y_j)]$

$g_{ij} = 9/4[3y_{iij} - 3y_{ijj} - (y_i + y_j)]$

$b_{ijk} = 27y_{ijk} - (27/4)(y_{iij} + y_{ijj} + y_{iik} + y_{ikk} + y_{jjk} + y_{jkk}) + (9/2)(y_i + y_j + y_k)$

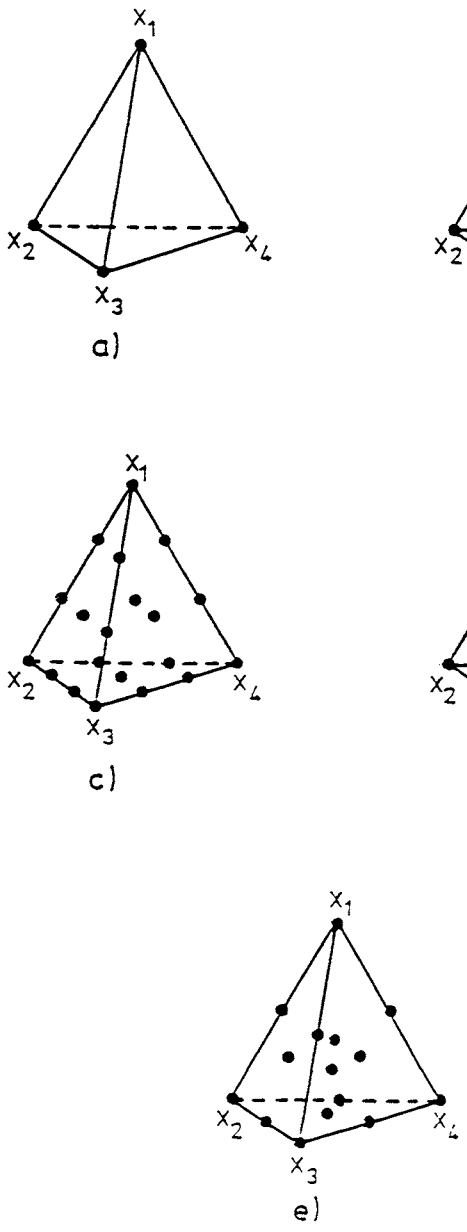


Fig. 25.8. Simplex lattice and centroid designs for $k = 4$; (a) $m = 1$; (b) $m = 2$; (c) $m = 3$; (d) $m = 3$, reduced; (e) centroid design.

TABLE 25.8

The (4,3) Simplex lattice design: reduced cubic design

Expt.	x_1	x_2	x_3	x_4	Response
1	1	0	0	0	y_1
2	0	1	0	0	y_2
3	0	0	1	0	y_3
4	0	0	0	1	y_4
5	0.5	0.5	0	0	y_{12}
6	0.5	0	0.5	0	y_{13}
7	0.5	0	0	0.5	y_{14}
8	0	0.5	0.5	0	y_{23}
9	0	0.5	0	0.5	y_{24}
10	0	0	0.5	0.5	y_{34}
11	0.33	0.33	0.33	0	y_{123}
12	0.33	0	0.33	0.33	y_{134}
13	0.33	0.33	0	0.33	y_{124}
14	0	0.33	0.33	0.33	y_{234}

Canonical equation: $y = \sum b_i x_i + \sum b_{ij} x_i x_j + \sum b_{ijk} x_i x_j x_k \qquad 1 \le i < j < k \le 4$

Coefficients: $b_i = y_i$

$b_{ij} = 4y_{ij} - 2(y_i + y_j)$

$b_{ijk} = 27(y_{ij} + y_{ik} + y_{jk}) + 3(y_i + y_j + y_k)$

TABLE 25.9

Number of experiments in the (k,m) Simplex-lattice design

Degree	Number of components (k)						
m	2	3	4	5	6	7	8
1	2	3	4	5	6	7	8
2	3	6	10	15	21	28	36
3	4	10	20	35	56	84	120
4	5	15	35	70	126	210	330

25.4.3 Simplex centroid design

When applying the Simplex lattice designs, one uses for each combination a maximum of m different compounds. Thus when one carries out the (4,2) simplex lattice designs of Table 25.6 only binary mixtures are considered, which seems somewhat strange for a quaternary problem. Also in some of these designs (see for instance experiments 9 to 20 in Table 25.7) the components appear in unequal proportions, which does not help interpretation of the results. Scheffé [4] developed

the so-called *Simplex centroid* designs to avoid this. He developed a polynomial equation consisting of product terms, such as (for $k = 3$):

$$y = \sum_{i=1}^{k} b_i x_i + \sum_{i}^{k-1} \sum_{j>i}^{k} b_{ij} x_i x_j + b_{123} x_1 x_2 x_3 \tag{25.9}$$

In general, the design points are the k permutations of (1, 0, ..., 0), the $\binom{k}{2}$ permutations of (0.5, 0.5, 0, .., 0), the $\binom{k}{3}$ permutations of (1/3, 1/3, 1/3, 0, .., 0) and the centre point ($1/k$, $1/k$, ..., $1/k$). The total number of design points is then $2^k - 1$.

Let us apply this for the $k = 3$ and $k = 4$ mixtures. For the $k = 3$ design, one first makes the 3 permutations of (1,0,0) yielding experiments 1–3 of Table 25.2, the 3 permutations of (0.5, 0.5, 0) yielding experiments 4 to 6 of the same table and eventually point (1/3, 1/3, 1/3), i.e. point 7 of Table 25.2. The Simplex centroid design is also equal to the reduced cubic (3,3) lattice design. The polynomial is given in eq. (25.1), the computation of the coefficients in Table 25.3 and Fig. 25.7c shows the design.

For the quaternary design, we can refer for a large part to Table 25.8. This gives the 4 permutations of (1, 0, 0, 0) (experiments 1–4), the 6 permutations of (0.5, 0.5, 0, 0) (experiments 5–10), the 4 permutations of (1/3, 1/3, 1/3, 0) (experiments 11–14). One needs to add to this a fourth type of point, namely (0.25, 0.25, 0.25, 0.25).

The polynomial is:

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{14}x_1x_4 + b_{23}x_2x_3 + b_{24}x_2x_4 + b_{34}x_3x_4 + b_{123}x_1x_2x_3 + b_{124}x_1x_2x_4 + b_{134}x_1x_3x_4 + b_{234}x_2x_3x_4 + b_{1234}x_1x_2x_3x_4$$

The coefficients are computed as described in Table 25.8, to which one must add

$$b_{1234} = 256\, y_{1234} - 108(y_{123} + y_{124} + y_{234}) + 32(y_{12} + y_{13} + y_{14} + y_{23} + y_{24} + y_{34}) - 4(y_1 + y_2 + y_3 + y_4)$$

The design is shown in Fig. 25.8e. It is interesting to note that this design contains all the centred simplexes from $k = 1$ to 4. Indeed, points 1 to 4 plus the extra point are the centred tetrahedron (simplex of degree 4). Each plane forming the tetrahedron is a triangle (simplex of degree 3), centred with the (1/3, 1/3, 1/3, 0) points. The vertices are lines (simplex of degree 2), centred with the (0.5, 0.5, 0, 0) points and the corner points can be considered as centred simplexes of degree 1.

25.4.4 Validation of the models

With 6 experiments one can determine the 6 coefficients of eq. (25.3), but, of course, it is not possible to determine how good the resulting equation is at predicting the response for all possible mixtures. It is then useful to determine the response for some additional mixtures.

A strategy which is often adopted in this field is to carry out the experiments for a simple model and some additional points that are needed for a higher order or more complete model. The coefficients of the simple model are determined and one predicts the response for the additional points. If the result is acceptable, so is the model and one stops. If this is not the case one incorporates the additional points into the model and determines some new additional points, that may lead again to a higher order model. For instance, let us consider the situation that one has decided initially on a (3,2) Simplex lattice design.

One reasons that, if the model of eq. (25.3) is not good enough, a model of higher degree ($m = 3$) must be determined and that this will require carrying out additional experiments. One can then carry out some of the experiments needed for the (3, 3) lattice design that are not part of the (3, 2) design and predict the values of these additional points to see how good the prediction power of the $m = 2$ model is. If it proves not good enough, at least some of the additional experiments needed to go to a $m = 3$ model have already been carried out.

In practice, this means that one would carry out the first 6 points of Table 25.2 and use these to obtain the quadratic model of eq. (25.3). At the same time, one would determine as additional points, point 7 which would allow to compute the reduced cubic model if the quadratic one is not good enough, or, else, all or some of points 4 to 10 of Table 25.4, which are missing to make it a complete cubic design.

For $k = 4$, one could start with the (4,1) model and determine 4 points (expt. 1 to 4 of Table 25.8), and determine additionally experiments 5 to 10. If the validation shows that the $m = 1$ model is not good enough, one incorporates these experiments in the model $m = 2$ and determines additionally experiments 11 to 14. If the model is still not good enough one can then determine the $m = 3$ model and, eventually, one can go to the $m = 4$ model. If the model is still not good enough, one should investigate the possibility that discontinuities occur (see also next section).

25.4.5 Designs based on inner points

In some cases, it is necessary that all components should be present in each experiment. Cornell [5,6], for instance describes an example where the composition of a bleach for the removal of ink is optimized. The bleaching agents are bromine, hypochloric powder and dilute HCl and the bleach functions only when all three components are present. The designs in the preceding sections all require a majority of experiments performed with mixtures that do not contain all components. For example the complete cubic design consists of 3 experiments with one-component "mixtures", 6 with binary mixtures and only one with a mixture containing all components. Clearly the designs discussed until now are not adapted

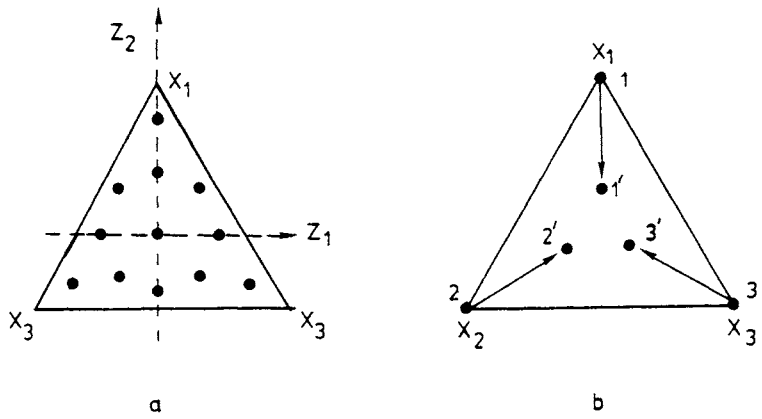


Fig. 25.9. Designs with inner points only according to (a) Draper [7,8] (the centre point is duplicated, so that 13 experiments are required) and (b) Lambrakis [9].

for situations where one needs all components present in all the experiments. In some cases, mixtures without one or more components are physically possible and functional, but their performance is influenced by the missing component to such a degree, that discontinuities in the neighbourhood of the boundaries of the simplex occur. In this case, too, the designs of the preceding sections are not acceptable.

There are two general ways out of this problem. One is to introduce lower bounds below which the concentration of the components is not allowed to descend. This leads to the use of pseudocomponents as explained in Section 25.5.

The other possibility is to apply designs with only interior points, i.e. where no points are situated on the boundary of the designs. Interior points always consist of mixtures with all components. Draper and Lawrence [7,8] situate all points in an interior triangle (Fig. 25.9a). It is noteworthy that they transform the 3-component problem to a 2-factor one by introducing the axes z shown in the figure. They then obtain the b-coefficients in the usual quadratic polynomial

$$y = b_0 + b_1 z_1 + b_2 z_2 + b_{11} z_1^2 + b_{22} z_2^2 + b_{12} z_1 z_2$$

by least-squares fitting.

Lambrakis [9] starts from the usual Scheffé (k,m) lattice designs and projects the experimental points into the interior of the simplex. An example is given for the (3,1) lattice in Fig. 25.9b. The design points 1, 2 and 3 of the lattice become 1′, 2′ and 3′ in the Lambrakis design. This transformation can be carried out for some, but not for all the lattice designs.

It should be noted here that discontinuities can exist not only in the neighbourhood of the boundaries. For instance, ternary diagrams are often used to study phase transitions and it is clear that, when more than one type of phase can exist within the experimental domain, strong discontinuities can be observed at the

borderline between two different phases. This is one of the reasons why a model can fail and, if one finds that a good model cannot be obtained, one should always wonder whether discontinuities can occur [10].

25.4.6 Regression modelling of mixture designs

Of course, the models described so far can be derived equally well with regression and using regression is necessary when more experiments were made than there are coefficients in the model. Let us return to the wine example of Section 25.1. Sergent et al. [11] describe the optimization of an assemblage where the design of Table 25.2 was carried out. All 10 experiments are considered design points and their results (Table 25.10) are used to derive the model coefficients. The model postulated is that of eq. (25.1). Regression leads to the equation:

$$y = 2.81\,x_1 + 3.87\,x_2 + 5.87\,x_3 + 13.37\,x_1x_2 + 49.37\,x_1x_3 + 35.46\,x_2x_3 - 106.41\,x_1x_2x_3$$

This leads to the isoresponse curves (contour plots) of Fig. 25.10.

It is sometimes necessary to apply the optimality criteria of Section 24.2, such as D-optimality. As in Chapter 24, one first defines the experiments that are possible, for instance all experiments on a certain grid covering the experimental domain, the model and the number of experiments one accepts to carry out. This approach is more often used, when there are constraints, such as those described in Sections 25.5 and 25.6. When one first defines the number of experiments one is willing to carry out and then needs to distribute them over the experimental region in the best possible way, one can also use these criteria. In mixture design the regression approach is applied less often, mainly for historical reasons, but we feel that more use of regression should be made. In the same way, the application of criteria such as D-optimality is to be recommended.

TABLE 25.10

Optimization of a wine "assemblage"

	x_1	x_2	x_3	y
1	1	0	0	3
2	0	1	0	4
3	0	0	1	5.5
4	0.5	0.5	0	7
5	0.5	0	0.5	16.5
6	0	0.5	0.5	13.5
7	0.33	0.33	0.33	11
8	0.66	0.17	0.17	9
9	0.17	0.66	0.17	8.5
10	0.17	0.17	0.66	14

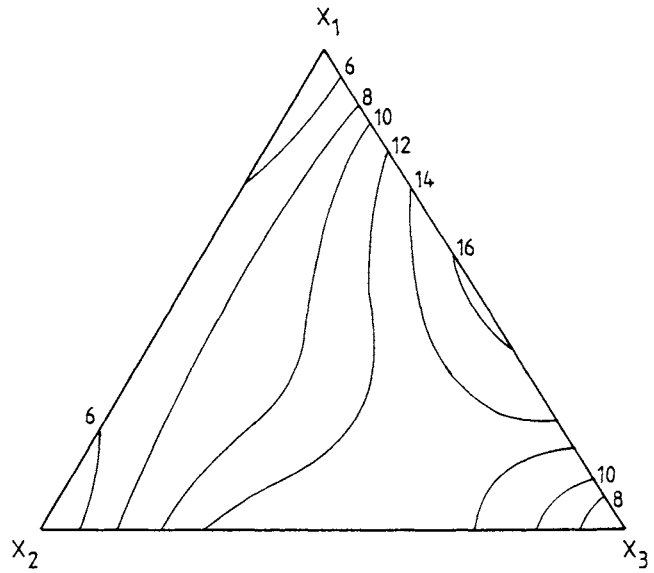


Fig. 25.10. Response surface for the wine assemblage problem described in Table 25.10 (adapted from Ref. [11]).

25.5 Upper or lower bounds

As explained in Section 25.4.5 it often happens that, for practical reasons, one has to limit the experimental domain. In the example of Fig. 25.5 it would not be possible to use the simplex design as such because it is not possible to make pellets with pure lactose ($x_1 = 1, x_2 = 0, x_3 = 0$) or water ($x_1 = 0, x_2 = 0, x_3 = 1$). In some cases, one can define upper and/or lower bounds i.e. the fraction of x_1 must be, at least, a certain percentage or it should not exceed a certain percentage. In this section we will consider first the situation that some or all factors have a lower bound or, else, some or all have an upper bound. The situation where some compounds have an upper bound and, at the same time, others have a lower bound is described in Section 25.6.

Let us consider a numerical example. A mixture of three compounds is to be studied and we require that $x_1 \geq 0.1, x_2 \geq 0.2, x_3 \geq 0.3$. This is shown in Fig. 25.11. The three bounds delimit a new simplex. This is again an equilateral triangle and the apexes can be considered to be new "pure" components and have compositions $x'_1 = 1, x'_2 = 0, x'_3 = 0$; $x'_1 = 0, x'_2 = 1, x'_3 = 0$ and $x'_1 = 0, x'_2 = 0, x'_3 = 1$. Of course, these new components are not really pure and for this reason they are called *pseudo-components*. They have the following compositions:

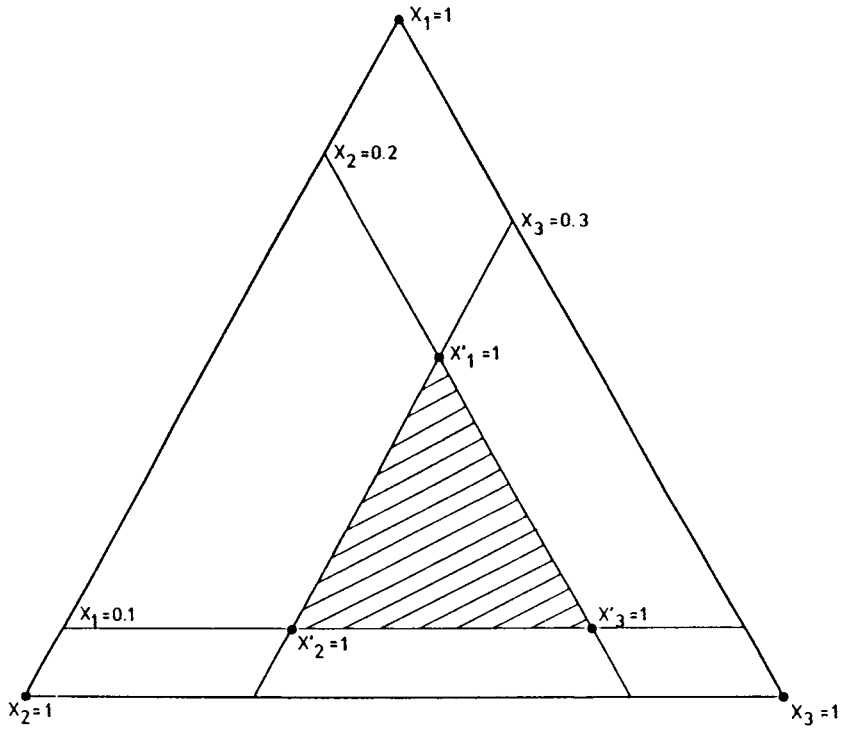


Fig. 25.11. The shaded area is the experimental area defined by the lower bounds $x_1 \geq 0.1$, $x_2 \geq 0.2$, $x_3 \geq 0.3$.

Pseudo-component 1: $x_1 = 0.5, x_2 = 0.2, x_3 = 0.3$
Pseudo-component 2: $x_1 = 0.1, x_2 = 0.6, x_3 = 0.3$
Pseudo-component 3: $x_1 = 0.1, x_2 = 0.2, x_3 = 0.7$

One can relate x'_1, x'_2, x'_3 to x_1, x_2, x_3 by the following equation

$$x'_i = \frac{x_i - c_i}{1 - \sum_{i=1}^{k} c_i} \tag{25.10}$$

where $c_i (c_i \geq 0$ and $\sum c_i < 1)$ is the lower bound for component i. For example:

$$x'_1 = \frac{x_1 - 0.1}{1 - (0.1 + 0.2 + 0.3)}$$

We can verify that at the apex of the new simplex ($x'_1 = 1$):

$x_1 = 1[1 - (0.1 + 0.2 + 0.3)] + 0.1 = 0.5$

and at position $x'_2 = 1$, where $x'_1 = 0$:

$x_1 = 0[1 - (0.1 + 0.2 + 0.3)] + 0.1 = 0.1$

The pseudo-components can form a new simplex design of the appropriate type and order. In Table 25.11 a (3,2) Simplex lattice design is shown for the boundaries given before. The responses y_1 to y_6 obtained with this design can then be used to determine the model of eq. (25.3) with x' replacing x. The b-coefficients are computed as described in Table 25.3. Once this model has been obtained, we can use the relationship of eq. (25.10) to predict y as a function of x_1, x_2 and x_3 in the original experimental domain.

When upper bounds are given, i.e. if some or all of the $x_1 < c$, where c is the upper bound, the simplex is not retained. Figure 25.12 gives examples for the

TABLE 25.11

(3,2) Simplex lattice design for pseudo-components with $x_1 \geq 0.1$, $x_2 \geq 0.2$, $x_3 \geq 0.3$

Expt.	Pseudo-component levels			Original component levels			Response
	x'_1	x'_2	x'_3	x_1	x_2	x_3	
1	1	0	0	0.5	0.2	0.3	y_1
2	0	1	0	0.1	0.6	0.3	y_2
3	0	0	1	0.1	0.2	0.7	y_3
4	0.5	0.5	0	0.3	0.4	0.3	y_4
5	0.5	0	0.5	0.3	0.2	0.5	y_5
6	0	0.5	0.5	0.1	0.4	0.5	y_6

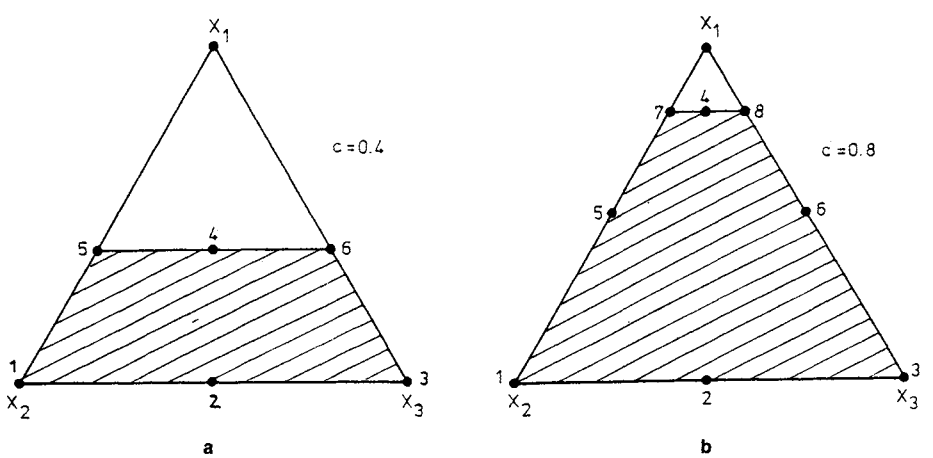


Fig. 25.12. Two situations with upper bounds: (a) $x_1 \leq 0.4$; (b) $x_1 \leq 0.8$.

situations where $x_1 \leq 0.4$ and $x_1 \leq 0.8$. We observe that the shaded area that describes the feasible experimental domain is now no longer a simplex. In choosing the experiments to be carried out, it should be remembered that one needs at least the same number of experiments as the number of coefficients to be estimated in the modelling equation and that these experiments should map the experimental domain efficiently.

One will usually look at the extreme vertices of the experimental domain and experiments, where the unrestricted factors take on a fixed proportion. For instance, in the examples of Fig. 25.12 one would look at situations where $x_2{:}x_3 = 1{:}1$. For $c = 0.4$, this would lead us to retain the vertices described as experiment 1, 3, 5 and 6 in Fig. 25.12a and add experiments 2 and 4, when the extreme levels (0 and 0.4) of x_1 are present and at the same time $x_2{:}x_3 = 1{:}1$. The resulting design is given in Table 25.12.

For $c = 0.8$ the same strategy would then lead to the selection of experiments at points 1, 2, 3, 4, 7 and 8 of Fig. 25.12b. However 4, 7 and 8 are rather close together so that one might prefer to replace 7 and 8 by experiments 5 and 6 of the unrestricted simplex lattice design. The resulting design is given in Table 25.13.

TABLE 25.12

Design for $x_1 \leq 0.4$

Expt.	x_1	x_2	x_3
1	0	1	0
2	0	0.5	0.5
3	0	0	1
4	0.4	0.3	0.3
5	0.4	0.6	0
6	0.4	0	0.6

TABLE 25.13

Design for $x_1 \leq 0.8$

Expt.	x_1	x_2	x_3
1	0	1	0
2	0	0.5	0.5
3	0	0	1
4	0.8	0.1	0.1
5	0.5	0.5	0
6	0.5	0	0.5

25.6 Upper and lower bounds

Suppose now that some or all of the components are subject to both upper and lower boundaries. This leads to the situation described in Fig. 25.13. To the lower bounds of Fig. 25.11 higher bounds were now added so that

$0.1 \leq x_1 \leq 0.4$

$0.2 \leq x_2 \leq 0.5$

$0.3 \leq x_3 \leq 0.6$

The shaded area is the area in which experiments are possible. The models to be fitted over this area still are equations of the type described in Sections 25.4.1–4.3. and usually the (3,2) or (3,3) reduced cubic model. They require that at least 6 or 7 experiments be carried out to determine the *b*-coefficients.

The method that is used most often is the *extreme vertices* method of McLean and Anderson [12]. It requires the determination of the experiments at the 6

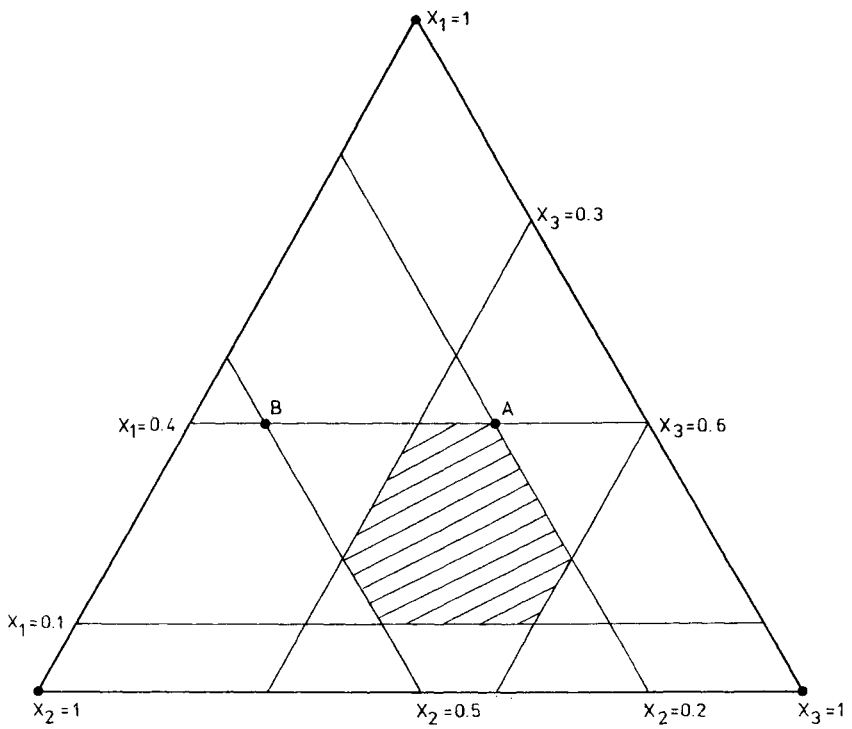


Fig. 25.13. McLean and Anderson's extreme vertices algorithm: first stage. Point A is a possible point, point B is not.

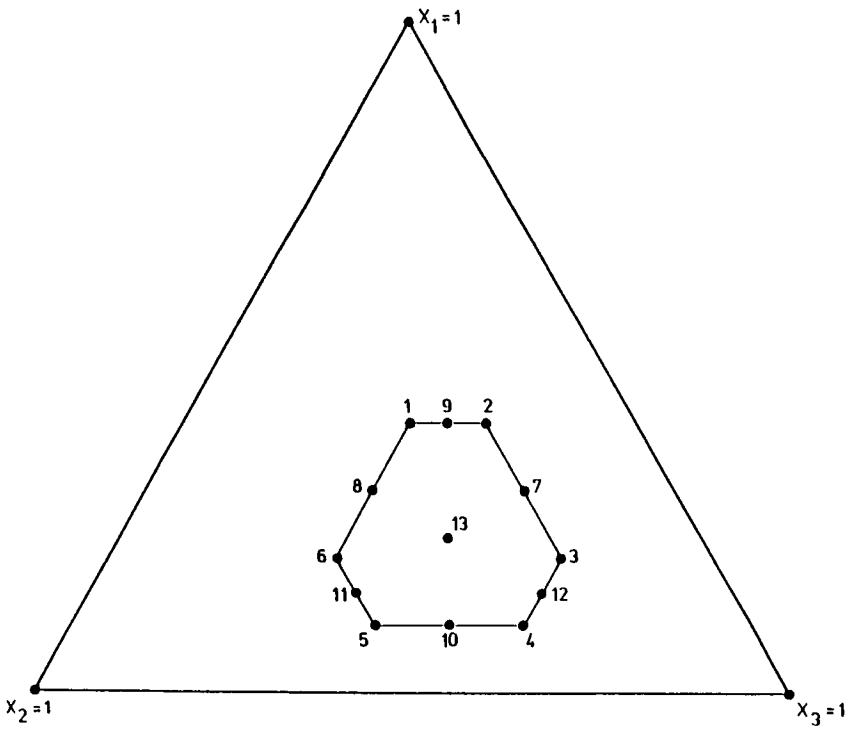


Fig. 25.14. McLean and Anderson's extreme vertices algorithm: candidate experimental points for $0.1 \leq x_1 \leq 0.4$, $0.2 \leq x_2 \leq 0.5$, $0.3 \leq x_3 \leq 0.6$.

vertices (points 1 to 6 in Fig. 25.14). When more experiments are wanted to obtain higher confidence in the model, one adds the face centroids (here 6 experiments, 7 to 12 in Fig. 25.14) and the overall centroid (13 in Fig. 25.14). McLean and Anderson proposed the following two-step algorithm to obtain the coordinates of these experiments:

1. Form all combinations of lower levels and upper levels for the different components in pairs and fill in how much of the third component would then be required to make the mixture. This means that one determines all the points where the lines in the ternary diagram cross.
2. Some of these mixtures (for instance A in Fig. 25.13) are possible, while others are not (such as B). These are detected in the algorithm because the lower or upper bounds for the third component are violated.

To understand this, let us carry out the calculation for the example of Figs. 25.13 and 25.14. The algorithm is summarized in Table 25.14. The four first lines in the table are obtained by first writing down all combinations of $x_1 = 0.1, x_1 = 0.4, x_2 = 0.2, x_2 = 0.5$. One then fills in x_3. The values filled out in this way are underlined in Table 25.14. For instance for the combination $x_1 = 0.4, x_2 = 0.5$ one needs $x_3 = 0.1$.

This experiment is however rejected since $x_3 = 0.1$ falls outside the permitted range $0.3 \leq x_3 < 0.6$. In fact, this is point B of Fig. 25.13. The combination $x_1 = 0.4$, $x_2 = 0.2$ requires $x_3 = 0.4$. This is acceptable and is taken up in the design as point A of Fig. 25.13 and point 2 of Fig. 25.14. The design can now be completed to yield the first six lines of Table 25.15.

To obtain the face centroids one takes all combinations of two vertices with the same value for one of the x's and obtains the others by averaging. For instance vertices 2 and 3 have $x_2 = 0.2$ in common. One obtains x_1 by averaging 0.4 and 0.2 (0.3) and x_3 by averaging 0.4 and 0.6 (0.5). The design is then completed by

TABLE 25.14

First stage of McLean and Anderson's extreme vertices method for $0.1 \leq x_1 \leq 0.4$, $0.2 \leq x_2 \leq 0.5$, $0.3 \leq x_3 \leq 0.6$

x_1	x_2	x_3	Expt. of Fig. 25.14
0.1	0.2	0.7	
0.1	0.5	0.4	(5)
0.4	0.2	0.4	(2)
0.4	0.5	0.1	
0.1	0.6	0.3	
0.1	0.3	0.6	(4)
0.4	0.3	0.3	(1)
0.4	0.0	0.6	
0.5	0.2	0.3	
0.2	0.2	0.6	(3)
0.2	0.5	0.3	(6)
−0.1	0.5	0.6	

TABLE 25.15

Second stage of McLean and Anderson's extreme vertices method

Expt.	Type	x_1	x_2	x_3	
1	Vertex	0.4	0.3	0.3	
2	Vertex	0.4	0.2	0.4	
3	Vertex	0.2	0.2	0.6	
4	Vertex	0.1	0.3	0.6	
5	Vertex	0.1	0.5	0.4	
6	Vertex	0.2	0.5	0.3	
7	Face centroid	0.3	0.2	0.5	2 + 3
8	Face centroid	0.3	0.4	0.3	1 + 6
9	Face centroid	0.4	0.25	0.35	1 + 2
10	Face centroid	0.1	0.4	0.5	5 + 4
11	Face centroid	0.15	0.5	0.35	6 + 5
12	Face centroid	0.15	0.25	0.6	3 + 4
13	Centroid	0.233	0.333	0.433	

averaging the x-values for the vertices to obtain the overall centroid. For instance the x_1 value for this point is $(0.4 + 0.4 + 0.2 + 0.1 + 0.1 + 0.2)/6 = 0.233$.

The number of experiments obtained in this way is usually higher than that required to fit the model. In this case there are 13 experiments and one needs 6 or 7 to determine the coefficients. If the number of experiments is considered to be too large, one can decide to drop some of the experiments. Algorithms to decide which ones to drop have been described by Snee [13]. One can also apply D-optimality concepts.

The visualisation of the results of complex mixture designs is not evident. An example of how to handle this can be found in Hare [14], who studied an instant soup thickener with four components, one of them (x_1) with an upper bound, the others with both upper and lower bounds. This yields the feasible region of Fig. 25.15. The response was a measure of lumping tendency. This was measured at the vertices, the overall centroid, the centroids of the edges and of some faces, 19 experiments in all. The model applied was

$$y = \sum_{i=1}^{4} b_i x'_i + \sum_{i<j}^{4} b_{ij} x'_i x'_j$$

where x'_i and x'_j are pseudo-components (see Section 25.5) of i and j. The b-coefficients were obtained by regression and this allowed to represent the results by

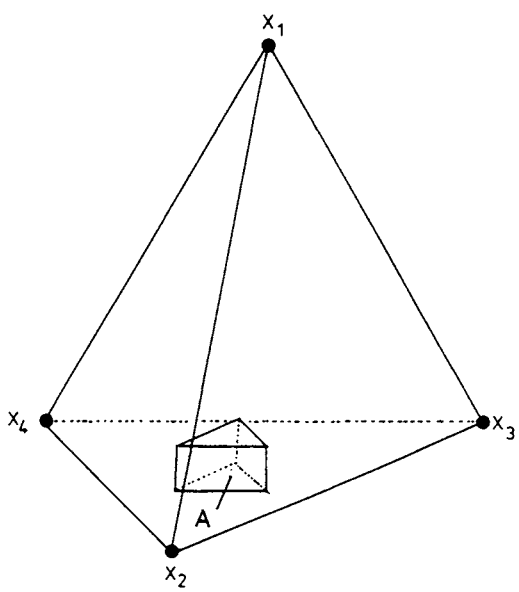


Fig. 25.15. Feasible region for the optimization of lumping tendency for an instant soup thickener (adapted from [14]). Triangle A is situated in the plane $x_2x_3x_4$ ($x_1 = 0$).

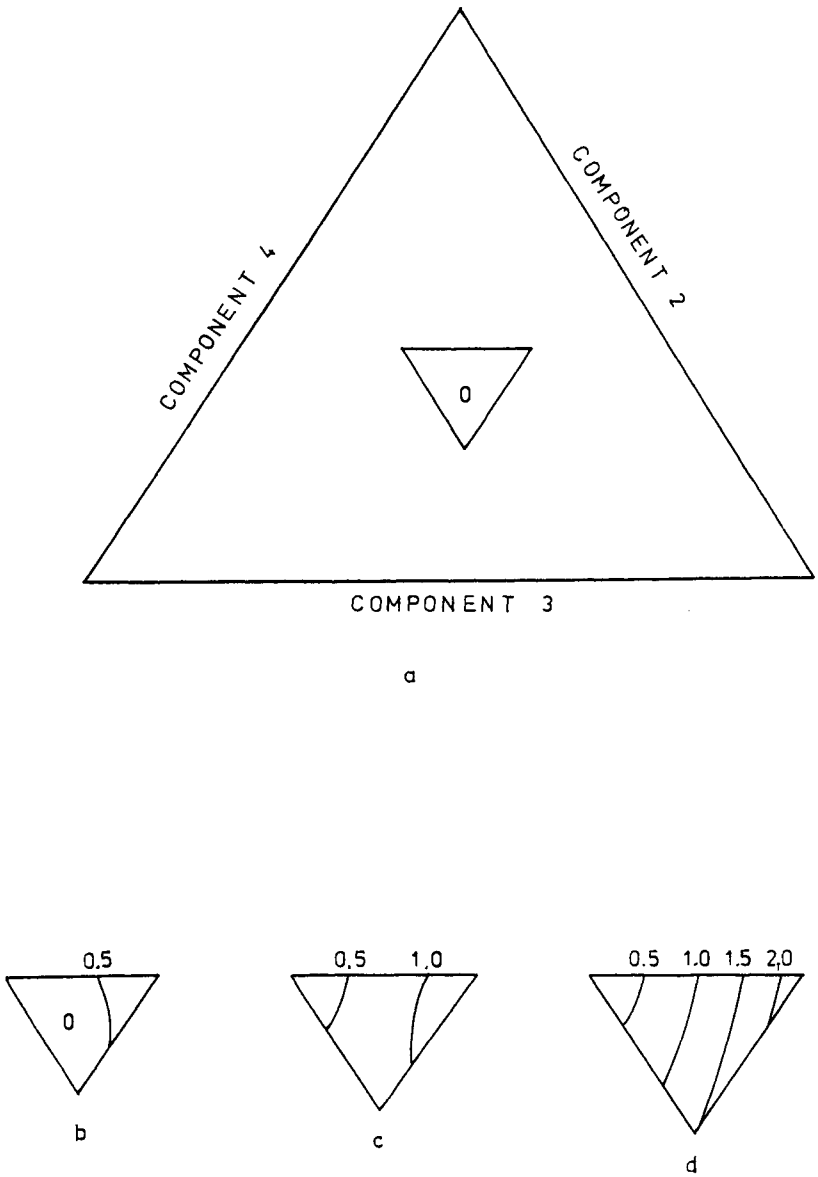


Fig. 25.16. Response surfaces in triangle A ($x_1 = 0$) (a) and slices at successively higher values of x_1 (b, c and d).

observing mixture diagrams between x'_2, x'_3 and x'_4 at given slices of x_1. The results are shown in Fig. 25.16. Hare's article includes several other ways of visualizing information obtained from mixture designs in general.

25.7 Combining mixture and process variables

Sometimes, one needs to combine process variables and mixture variables in one experimental design. When the mixture consists of only two components, this offers no problems. One selects one of the components and treats the fraction of that component in the mixture as a process variable. As soon as there are more than two components, the situation becomes more complex. This can be illustrated with an example from Phan-Tan-Luu and colleagues [15].

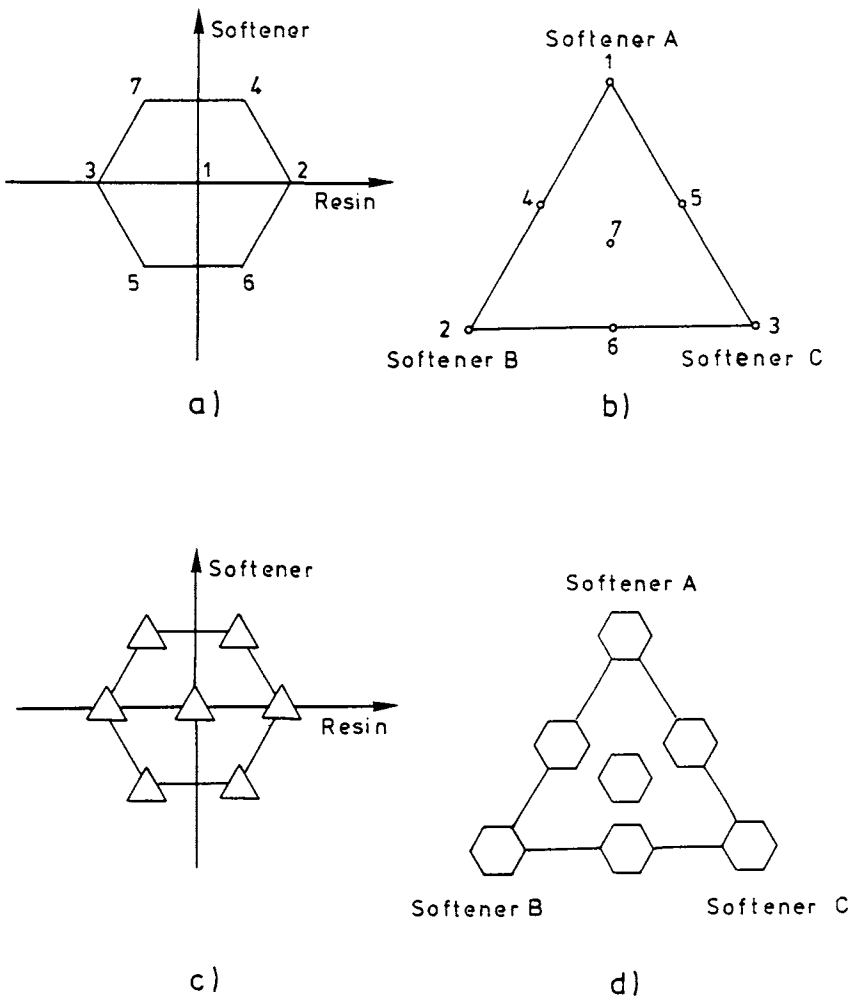


Fig. 25.17. Simultaneous optimization of process and mixture variables for the bed linen example. (a) Doehlert design for the process variables; (b) Scheffé Simplex centroid design for mixture variables; (c) the process/mixture variable design represented in the process variable space; (d) the process/mixture variable design represented in the mixture variable space (from Ref. [15]).

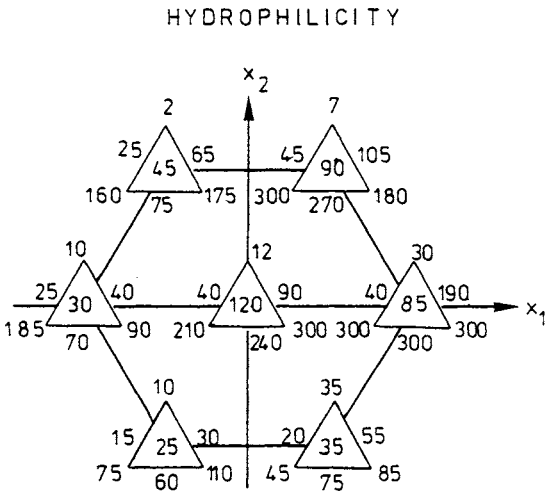


Fig. 25.18. Values for one of the responses in the bed linen example.

The example concerns the development of a finishing product formula for polyester/cotton cloth for use as bed linen. This product formula is based on the use of a resin and that of a softener mixture. The softener mixture consists of three softening products. The questions asked are: (i) how much resin and softener mixture should be used and (ii) what should be the composition of the softener mixture. The first question involves a typical two-factor process variable optimization, while the second concerns the optimization of a mixture. The authors answered the first question by using a Doehlert design (see Chapter 24) and the second by using a mixture design. The two designs are shown in Fig. 25.17a,b. They are combined in such a way that at each of the corners of the Doehlert design, one carries out a mixture design (Fig. 25.17c). Figure 25.18 shows the experimental results for one of the responses studied by the authors. The authors studied in fact several responses. How to treat such multiple response problems is discussed further in Chapter 26.

The combined process variable–mixture designs can always be represented in two ways, namely as a factorial design with in each point a mixture design, or, alternatively as a mixture design with in each point a factorial design. Both are alternative representations of the same experiment. Figure 25.17d shows the bed linen experiment represented as a mixture design, with at each design point a Doehlert design to study the process variables.

The number of experiments required is often high. In our example, it was 7 (for the Doehlert design) $\times$ 7 (for the mixture design). The combination of two-level factorial designs with mixture designs is more common than the use of the Doehlert design, and, in that case, one can apply fractional factorial designs instead of full

factorials. Doornbos and colleagues (see Ref. [16] for a review) have described the use of methods in which both the mixture design and process variable design part are fractionated.

Models for combined mixture and process variables are complex. They contain cross-product terms between the mixture and the process variables (see for instance Gorman and Cornell [17]). Kettaneh-Wold [18] has shown that partial least squares (PLS) (see Chapter 35) as a modelling method offers advantages above the classical regression techniques and this seems to be especially the case for designs that involve both mixtures and process variables.

References

1. D. Feneuille, D. Mathieu and R. Phan-Tan-Luu, Méthodologie de la Recherche Expérimentale. LPRAI, Aix-en-Provence, 1983
2. L. Baert, D. Fanara, J.P. Remon and D.L. Massart, Correlation of extrusion forces, raw materials and sphere characteristics. J. Pharm. Pharmacol., 44 (1992) 676–678.
3. H. Scheffé, Experiments with mixtures. J.R. Stat. Soc., B20 (1958) (2) 344–360.
4. H. Scheffé, Simplex-centroid designs for experiments with mixtures. J.R. Stat. Soc., B25 (1963) (2) 235–263.
5. J.A. Cornell, Experiments with mixtures: a review. Technometrics, 15 (1973) 437.
6. J.A. Cornell, Experiments with mixtures: a review, an update and bibliography. Technometrics, 21 (1979) 95–106.
7. N.R. Draper and W.E. Lawrence, Mixture designs for 3 factors. J. R. Stat. Soc., B27 (1965) (3) 450–465.
8. N.R. Draper and W.E. Lawrence, Mixture designs for 4 factors. J. R. Stat. Soc., B27 (1965) (3) 473–478.
9. D.P. Lambrakis, Experiments with mixtures: an alternative to the simplex-lattice design. J. R. Stat. Soc., B31 (1969) (2) 234–245.
10. D. Mathieu, E. Puech-Costes, M. Maurette and R. Phan-Tan-Luu, The simplex method applied to the detection and following of a discontinuity. Chemom. Intell. Lab. Syst., 20 (1993) 25–34.
11. M. Sergent, E. Meroniand and R. Phan-Tan-Luu, Méthodologie de la recherche expérimentale appliquée à l'étude des assemblages de vins provenant de différents cépages dans les Bouches du Rhône. Rapport Technique 868712, LMRE, Marseille, 1988.
12. R.A. McLean and V.L. Anderson, Extreme vertices design of mixture experiments. Technometrics, 8 (1966) 447–454.
13. R.D. Snee, Experimental designs for quadratic models in constrained mixture spaces. Technometrics, 17 (1975) 149–159.
14. L.B. Hare, Graphical display of the results of mixture experiments, in: Experiments in Industry: Design, Analysis and Interpretation of Results, R.D. Snee, L.B. Hare and J.R. Trout (Eds.). ASQC, 1985, pp. 99–109.
15. J. Chardon, J. Nony, M. Sergent, D. Mathieu and R. Phan-Tan-Luu, Experimental research methodology applied to the development of a formulation for use with textiles. Chemom. Intell. Lab. Syst., 6 (1989) 313–321.
16. D.A. Doornbos and P. de Haan, Optimization techniques in formulation and processing, in: Encyclopedia of Pharmaceutical Technology, Vol. 11. Marcel Dekker, New York, 1995, pp. 77–159.

17. J.W. Gorman and J.A. Cornell, A note on model reduction for experiments in both mixture components and process variables. Technometrics, 24 (1982) 243–247.
18. N. Kettaneh-Wold, Analysis of mixture data with partial least squares. Chemom. Intell. Lab. Syst., 14 (1992) 57–69.

Recommended reading:

J.A. Cornell, Experiments with Mixtures: Designs, Models and the Analysis of Mixture Data, 2nd edn. John Wiley, New York, 1990.

Chapter 26

Other Optimization Methods

26.1 Introduction

In the preceding chapters we have discussed the main methods for optimization. In particular, the response surface methods for process variables (Chapter 24) and mixture variables (Chapter 25) have been applied to a great extent. In this chapter we will explain some further methods that are somewhat less often used.

We will discuss sequential optimization methods (Section 26.2) and mixed sequential–simultaneous approaches (Section 26.3). The same methods can be applied for numerical optimization, i.e. finding optimal values for parameters in, e.g., regression equations. We will also pay some more attention to the optimization criteria. Section 26.4 describes methods that can be applied when more than one optimization criterion has to be considered and Section 26.5 gives a short introduction to Taguchi methodology. This concentrates on finding good responses that are also robust. In other words, both the value of the response and its robustness are considered as criteria.

26.2 Sequential optimization methods

Sequential methods were introduced in Section 21.5. Using these methods, one carries out a very restricted amount of experiments, typically one more than the number of factors. On the basis of these results, one then decides on the next experiment. The result of this experiment and those that were carried out earlier are then used to select the conditions for the next experiment, etc. We will describe two such methods. The first is based on Fibonacci numbers, the next is the Simplex method. Both are used for numerical optimization and the latter also for experimental optimization.

26.2.1 Fibonacci numbers

Fibonacci numbers are called after the 13th century mathematician Leonardo of Pisa, who was also called Fibonacci. Fibonacci numbers are defined by the recursive relationship

$$t_{n+2} = t_n + t_{n+1} \qquad n = 0, 1, 2, \ldots \tag{26.1}$$

with $t_0 = 1$ and $t_1 = 1$. In words, each number of the series is the sum of the two preceding numbers. The Fibonacci series therefore begins as follows: 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, ...

These numbers can be used to direct a *restricted region search*, meaning that the boundaries of the region to be searched are known. The method is valid only for univariate and unimodal situations. The search proceeds by eliminating parts of this region from consideration, thereby narrowing at each cycle the region in which the optimum can be situated.

Consider the case in which the maximum of a response y must be found in a region $[x_A, x_B]$. The function $y = f(x)$, which is unknown to the experimenter, is depicted in Fig. 26.1. The value to be found is x_{opt}. Two experiments are carried out with the variable values x_1 and x_2, chosen in such a way that the distance between x_A and x_1 is equal to that between x_2 and x_B. The resulting y_1 and y_2 values are recorded. In the example it is observed that $y_1 > y_2$. It is therefore possible to conclude that the maximum is not situated in the $[x_2, x_B]$ region, to eliminate this region from further consideration and to concentrate on the $[x_A, x_2]$ region which can be considered in its turn as a restricted region in which a search has to be carried out. In this region there is already one experimental result available, y_1. We can then repeat the strategy of the first cycle by selecting x_3 so that the distance between x_A and x_3 is equal to that between x_1 and x_2. In the present instance, this leads to the elimination of the region $[x_A, x_3]$ and the selection of x_4, so that $x_2 - x_4 = x_3 - x_1$, etc. Fibonacci numbers are used to select x_1, x_2, etc.

First, the experimenter must decide on the width, a, of the optimal region which he will accept compared with the original search region, A. The Fibonacci series indicates which number is the one immediately larger than A/a. If this is the $n+1$th number in the series, then n experiments will be needed. For example, if A/a is 50,

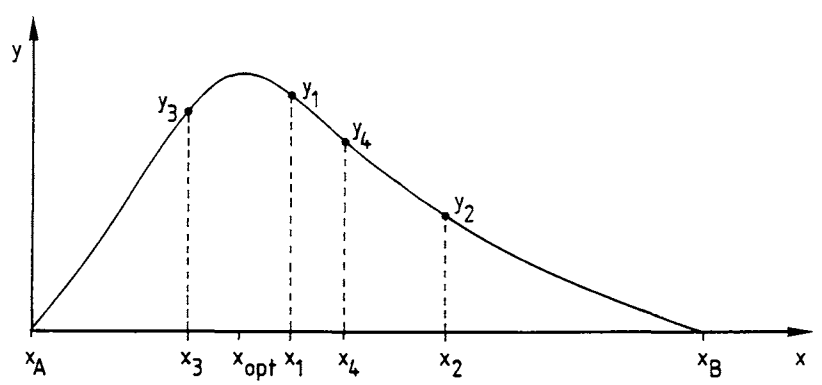


Fig. 26.1. Example of the first stages in a Fibonacci search.

then the smallest Fibonacci number which is higher than A/a is 55. This is the tenth number in the series, t_9, and therefore nine experiments will be necessary.

Let us call the length of the original search region $L_1(L_1 = x_B - x_A)$ and the distances between the experiments x_1 and x_2 and the boundaries, l_1 ($l_1 = x_1 - x_A = x_B - x_2$). The latter is defined as:

$$l_1 = \frac{t_{n-2}}{t_n} L_1 \tag{26.2}$$

For the example

$$l_1 = \frac{21}{55} L_1$$

and therefore $x_1 = x_A + (21/55)\,(x_B - x_A)$ and $x_2 = x_B - (21/55)\,(x_B - x_A)$.

One of the intervals $[x_A, x_1]$ or $[x_2, x_B]$ is eliminated according to whether $y_1 < y_2$ or *vice versa*, as explained earlier. The length of the remaining region is

$$L_2 = L_1 - \frac{t_{n-2}}{t_n} L_1 = \frac{t_{n-1}}{t_n} L_1$$

since

$$t_{n-1} + t_{n-2} = t_n \tag{26.3}$$

Let us suppose that $[x_2, x_B]$ was eliminated; x_1 is retained and we have to determine l_2, so that this is equal to the distances between x_1 and x_2 and between a new experiment x_3 and x_A.

The following general equation is then applied

$$l_k = \frac{t_{n-(k+1)}}{t_{n-(k-1)}} L_k \tag{26.4}$$

so that

$$l_2 = \frac{t_{n-3}}{t_{n-1}} L_2$$

also,

$$L_k = \frac{t_{n-(k-1)}}{t_{n-(k-2)}} L_{k-1} \tag{26.5}$$

so that

$$L_3 = \frac{t_{n-2}}{t_{n-1}} L_2$$

One proceeds in the same way until $n - 1$ experiments have been performed.

For the last experiment

$$\frac{l_{n-1}}{L_{n-1}} = \frac{t_{n-[(n-1)+1]}}{t_{n-[(n-1)-1]}} = \frac{t_0}{t_2} = \frac{1}{2} \tag{26.6}$$

which means that the distance between the last-but-one experiment and the boundary

of the remaining region is half of the length of this region. In other words, the last-but-one experiment is situated at the centre of the remaining search region. The last or nth experiment should also be placed at this point. If the two experiments are carried out with the same x value no new information is gained. Therefore, the last experiment is placed at the smallest distance which is thought to give a measurable difference in response. If $y_n > y_{n-1}$, the optimum is situated in the interval $x_{n-1} - x_{n-2}$; if $y_{n-1} > y_n$, it is to be found in $x_n - x_{n-3}$.

The Fibonacci search can be shown to be very effective, meaning that a very small number of cycles is necessary. This is particularly true when the optimal value must be known very precisely and its effectiveness can be shown by comparing the Fibonacci method with the simplest possible search method, the pre-planned regular interval design. When the optimal region must be one thousandth of the original region, the two methods require 1999 and 16 experiments respectively. In fact, the Fibonacci search method is the best available as far as effectiveness is concerned. When the experimental error is large compared to the slope of the response, this can lead to the exclusion of the wrong region. However, this is no problem in numerical optimization. In the chemometrical literature an example can be found in the linearization of calibration lines for AAS by Wang et al. [1].

26.2.2 The Simplex method

A simplex is a convex geometric figure in the factor space defined by a number of points equal to one more than the number of factors considered in the optimization. For the simplest problem, namely an optimization of two factors, the simplex is therefore a triangle. The simplex has as many dimensions as factors. An example will be used to introduce the technique. Consider the isoresponse surface given in Fig. 26.2, which describes the optimization of a colorimetric determination of sulphur dioxide [2]. The numbers along the isoresponse lines are absorbances and the highest absorbance is considered to be the optimum.

The optimization starts with experiments 1, 2 and 3. The points representing the experiments form an equilateral triangle and point 2 shows the worst response of the three. It is logical to conclude that the response will probably be higher in the direction opposite to this point. Therefore, the triangle is reflected so that point 4 opposite to point 2 is obtained. An experiment is now run with the factor values of point 4. Points 1, 3 and 4 form together a new simplex. The procedure is now repeated.

It appears that point 3 yields the lowest absorbance. Point 3 is therefore rejected and point 5 is obtained. In this way, using successive simplexes, one moves rapidly along the response surface. This procedure is described by the following rule [3]:

Rule 1: the new simplex is formed by rejecting the point with the worst result in the preceding simplex and replacing it with its mirror image across the line defined by the two remaining points.

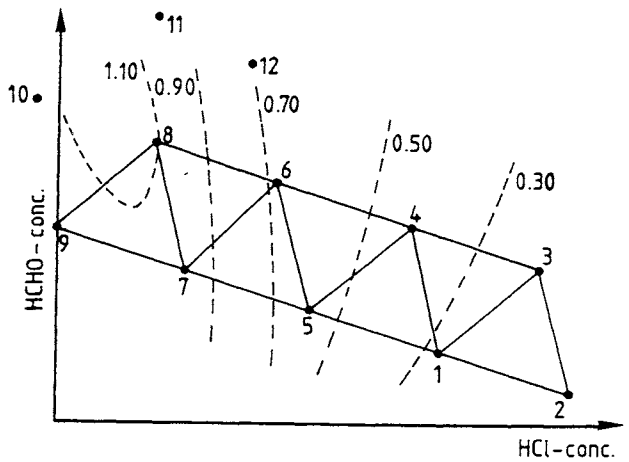


Fig. 26.2. Example of fixed size Simplex optimization (adapted from Ref. [2]).

In the initial stages of an optimization, the new point in a simplex will usually yield a better result than at least one of the two remaining points, because the simplexes will tend to move towards the optimum. When the new point does not cause a move in this general direction a change in the progression axis is necessary. When the new point has the worst response of the simplex, it makes no sense to apply rule 1, as this would lead to reflection back to the point which was itself the worst one in the preceding simplex. For example consider simplex 6, 7 and 8 in Fig. 26.2. Point 6 has the lowest absorbance and is replaced by 9, its mirror image across the line 7–8. Point 9 has the least desirable response in the new simplex. Rule 1 would lead back to point 6, then again to point 9, etc. Therefore, one now applies rule 2.

Rule 2: if the newly obtained point in a simplex has the worst response, do not apply rule 1 but instead eliminate the point with the second lowest response and obtain its mirror image to form the new simplex.

The effect of this rule is to change the direction of progression towards the optimum. This will most often happen in the region of the optimum. If a point is obtained near to it, all of the other new points will overshoot the top of the response curve. A change in direction is then indicated. In the region of the optimum, the effect is that the simplexes circle around the provisional optimal point. For example, in Fig. 26.2 the application of rule 2 would lead to the rejection of the second lowest point, 7. Its reflection yields 10, a point with a negative hydrochloric acid concentration. Let us suppose for the moment that this is possible and that 10 would yield the lowest response. Rule 2 then leads to 11. The response of this point is lower than the response of 8 but better than that of point 10. Point 8 is retained in consecutive simplexes, which is interpreted as indicating that this point is

situated as near to the optimum as one can get with the initially chosen simplex. The situation could also result from an erroneously high response from point 8. To make sure that this is not the case, one applies rule 3.

Rule 3: if one point is retained in three successive simplexes, determine again the response at this point. If it is the highest in the last three simplexes it is considered as the optimum which can be attained with simplexes of the chosen size. If not, an experimental error has been made, the simplex has become trapped at a false maximum and one starts again.

One difficulty which still has to be resolved is what to do in practice when one encounters a situation such as that exemplified by point 10. To avoid it, one identifies the constraints or the boundaries between which the simplex may move. For example, when the factors are concentrations, values lower than 0 are not possible. Once this has been done, one applies rule 4.

Rule 4: if a point falls outside one of the boundaries, assign an artificially low response to it and proceed with rules 1–3.

The effect of applying rule 4 is that the outlying point is automatically rejected without bringing the succession of simplexes to an end.

Let us write these rules in vector notation. The initial simplex is called BNW (Fig. 26.3). In this simplex, the best response is obtained for *vertex* B and the worst for vertex W. The symbol N stands for next best. Let **b**, **n** and **w** be the vectors representing points B, N and W, i.e. $\mathbf{b} = [x_{1b}\ x_{2b}]$, $\mathbf{n} = [x_{1n}\ x_{2n}]$ and $\mathbf{w} = [x_{1w}\ x_{2w}]$. Since W is worst, it is eliminated and the centroid P of the line segment BN is:

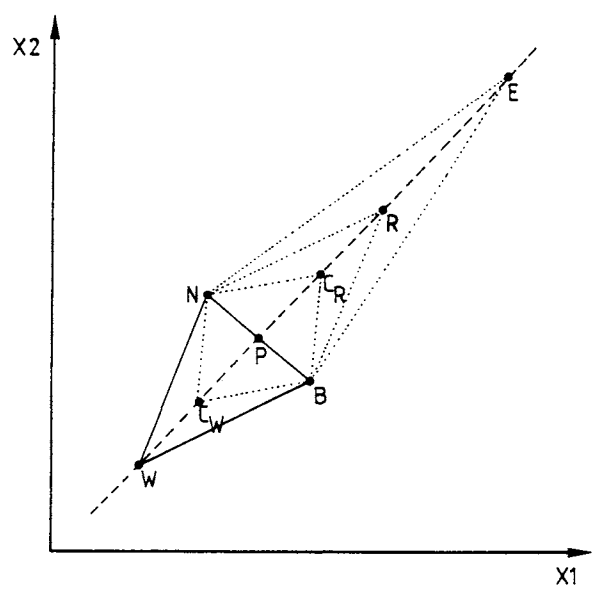


Fig. 26.3. The modified Simplex for two factors x_1 and x_2.

$\mathbf{p} = \frac{1}{2}(\mathbf{n} + \mathbf{b}) = [(x_{1n} + x_{1b}),(x_{2n} + x_{2b})]/2$

the reflected vertex R is obtained by:

$\mathbf{r} = \mathbf{p} + (\mathbf{p} - \mathbf{w})$

Let us now consider the 3 factor simplex, which is a tetrahedron. If we call N_1 and N_2 the points that are neither best nor worst, then:

$$\mathbf{p} = \tfrac{1}{3}(\mathbf{n}_1 + \mathbf{n}_2 + \mathbf{b}) =$$

$$\left[(x_{1n_1} + x_{1n_2} + x_{1b}),\ (x_{2n_1} + x_{2n_2} + x_{2b})\ ,\ (x_{3n_1} + x_{3n_2} + x_{3b})\right]/3$$

and again $\mathbf{r} = \mathbf{p} + (\mathbf{p} - \mathbf{w}) = 2\,\mathbf{p} - \mathbf{w}$.

The two-factor case can be generalized to the *k*-factor case. In words, when the vertex to be rejected has been determined, the coordinates of *k* retained vertices are summed for each factor and multiplied by $2/k$. From the resultant values one subtracts the coordinates of the rejected point. The result yields the coordinates of the new vertex.

The initial simplex in the general case is obtained with the use of the factors in Table 26.1. It is best explained with an example. Suppose the simplex is a tetrahedron, because three factors are optimized. The multiplication factors of Table 26.1 specify the distance of each vertex from the experimental origin. The experimenter has to define the experimental origin and the step size for each factor, i.e. the maximum change that one wants to apply for a certain factor at each step of the procedure. Let the experimental origin be: factor 1 = 10, factor 2 = 100, factor 3 = 20. These are then the coordinates of vertex 1. If the step sizes are respectively 10, 20 and 5, then the other vertices are obtained as follows.

TABLE 26.1

Values of multiplication factors for the calculation of vertices of the initial simplex

Vertex	Factor							
	A	B	C	D	E	F	G	H
1	0	0	0	0	0	0	0	0
2	1.000	0	0	0	0	0	0	0
3	0.500	0.866	0	0	0	0	0	0
4	0.500	0.289	0.817	0	0	0	0	0
5	0.500	0.289	0.204	0.791	0	0	0	0
6	0.500	0.289	0.204	0.158	0.775	0	0	0
7	0.500	0.289	0.204	0.158	0.129	0.764	0	0
8	0.500	0.289	0.204	0.158	0.129	0.109	0.756	0
9	0.500	0.289	0.204	0.158	0.129	0.109	0.094	0.750

Vertex 2:

Factor 1: $10 + (10 \times 1) = 20$
Factor 2: $100 + (20 \times 0) = 100$
Factor 3: $20 + (5 \times 0) = 20$

Vertex 3:

Factor 1: $10 + (10 \times 0.5) = 15$
Factor 2: $100 + (20 \times 0.866) = 117$
Factor 3: $20 + (5 \times 0) = 20$

Vertex 4:

Factor 1: $10 + (10 \times 0.5) = 15$
Factor 2: $100 + (20 \times 0.289) = 106$
Factor 3: $20 + (5 \times 0.817) = 24$

26.2.3 The modified Simplex method

In the original Simplex method the step size is fixed. When it is too small, it takes many experiments to find the optimum; when it is too large, the optimum is determined with insufficient precision. In the latter instance, one can start a new simplex around the provisional optimum with a smaller step size. However, a modified Simplex [4] method in which the step size is variable throughout the whole procedure offers a more elegant (and efficient) solution. The principal disadvantage is that the simplicity of the calculations in the original Simplex method no longer exists. The principles of the method are retained but additionally, provision is made for the expansion or contraction of simplexes.

The simplex search is accelerated by expanding it in directions which seem favourable and slowed down by contracting it in the directions that are unfavourable. This method, which was devised by Nelder and Mead [4], was introduced into chemistry by Morgan and Deming [5]. It is explained here for the two-factor case (Fig. 26.3). This again yields a triangle (which is now no longer necessarily equilateral) as the simplex. Depending on the response in R the following steps are undertaken.

(a) Response at R > response at B.

The simplex seems to move fast in a favourable direction. An expansion is therefore attempted by generating vertex E:

$$E = P + \gamma (P - W)$$

where γ is usually 2. If the response at E is also better than at B, the E is retained and the new simplex is BNE. If not, the expansion is considered to have failed and the new simplex is BNR.

(b) Response at B > response at R > response at N.

The new simplex is BNR. No expansion or contraction is envisaged.

(c) Response at N > response at R.

The simplex has moved too far and it should be contracted. If the response at R is not worse than at W, the new vertex C_R is best situated nearer to R than to W

$$C_R = P + \beta (P - W)$$

where β is usually 0.5. If the response at R is also worse than that at W, the new vertex C_W should be situated nearer to W

$$C_W = P - \beta (P - W)$$

The new simplex is BNC_R or BNC_W. From here on, one proceeds by rejecting in the new simplex the point that was next-to-worst in the old simplex (N).

26.2.4 Advantages and disadvantages of Simplex methods

The Simplex methodology is probably the best known so-called hill-climbing method. It is often used, both in numerical and in experimental optimization [6,7]. Its application for numerical optimization was mentioned in Chapter 11. In experimental optimization, it is very useful to obtain rapid improvement of a single performance criterion. For this purpose, it is probably the best method available. Its main disadvantages are:

– The Simplex method will find the global optimum when there is only one optimum. When there are local optima, the Simplex method will find one of the optima, but not necessarily the best one.

– It does not work with more than one performance criterion. Indeed, these will probably have their maxima at different locations. The Simplex method may be used to find the optimum of a composite response, using utility or Derringer functions (and preferably the latter, see Sections 26.4.3 and 4), but cannot be applied with the other multicriteria methods described in Section 26.4. If utility or Derringer functions are used as criteria, one must bear in mind (see Section 21.4) that composite criteria often lead to local optima and, as explained above, one is not guaranteed to find the global optimum.

– If the imprecision of the measurement is relatively large and the slope of the response surface studied is small, the Simplex method may move in wrong directions. Although this can be corrected in a subsequent move, this makes the method inefficient in such cases.

– The Simplex method gives little information about the response surface. One can use the experimental results obtained to model the surface, but, from a modelling point of view, the points are chosen in a haphazard way. It is, for instance, not evident that the whole experimental area of interest will be mapped and it is probable that the experimental design will not correspond to the optimality criteria of Chapter 24.

In short, the Simplex method is the method of choice for rapid straightforward optimization of single responses or of composite responses when one knows there can only be one optimum. It is also very useful when improvement, but not necessarily optimization is wanted.

26.3 Steepest ascent methods

Instead of using a simplex, we can apply local factorial designs as hill-climbing methods. Consider four experiments, constituting a 2^2 factorial experiment (see Fig.

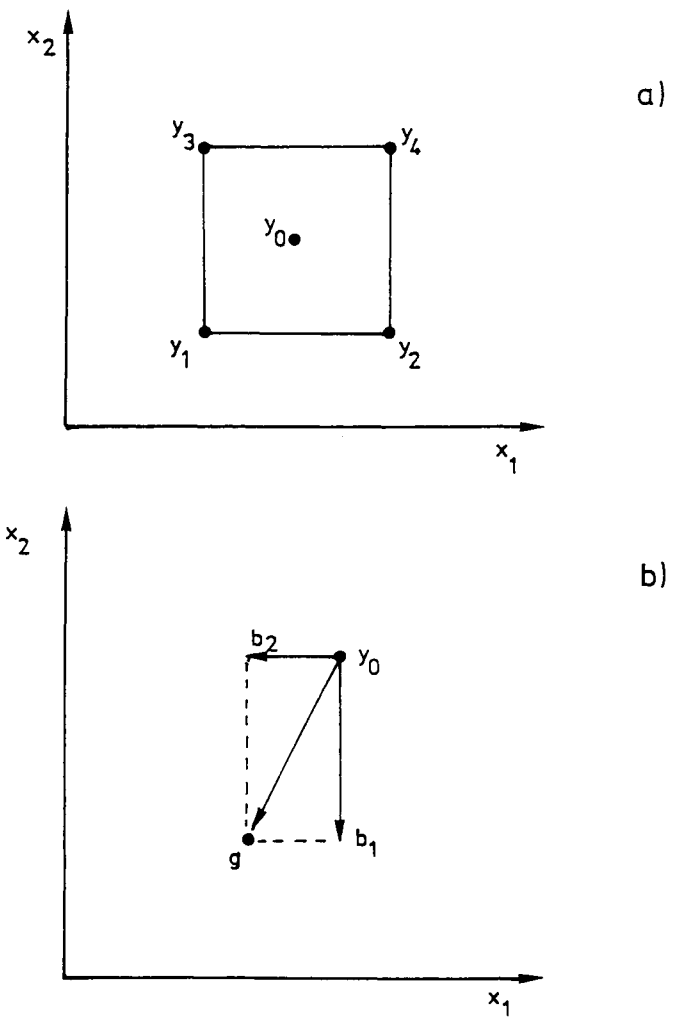


Fig. 26.4. (a) A 2^2 full factorial with centre point. (b) Additional experiment in a steepest ascent procedure.

26.4a). The response y can be described as a function of factors x_1 and x_2, yielding equation

$$y = b_0 + b_1x_1 + b_2x_2 \tag{26.7}$$

For the sake of simplicity, we suppose that it has been decided that the interaction term can be neglected. The centre point of the design has also been measured and that measure has been replicated (yielding y_{01}, y_{02}, y_{03}). As we have seen in Section 22.9.3, this allows us to carry out a curvature check. If there is no curvature, then eq. (26.7) describes the response well. The response surface defined in this way is a plane and the optimal (supposedly highest) value must be situated on the boundary defined by the first four experiments.

Suppose that both b_1 and b_2 are negative. It is then possible that there is an optimum outside the experimental domain in the direction of lower values of the x_1 and x_2 variables. It would be logical to do an experiment in that direction. The path of steepest ascent is given by a direction such that for every unit of movement in the direction of x_1, one should move b_2/b_1 units in the direction of x_2 (Fig. 26.4b).

A simple (synthetic) example is given in Table 26.2 and Fig. 26.5. A more realistic example can be found in the book by Box, Hunter and Hunter [8]. A first 2^2 factorial design is carried out around the starting point (100, 48). The design consists of experiments 1 to 4 and the centre point is replicated 3 times (expts. 0.1, 0.2 and 0.3). It is found that $b_1 = -1$ and $b_2 = -2$ in scaled units. Therefore one

TABLE 26.2
Steepest ascent example

Expts.	y	x_1	x_2	x_1	x_2
		original		scaled	
0.1	45	100	48	0	0
0.2	44.4	100	48	0	0
0.3	45.5	100	48	0	0
1	46	102	44	+1	−1
2	42	102	52	+1	+1
3	48	98	44	−1	−1
4	44	98	52	−1	+1
5	50	98	40		
6	55	96	32	0	0
7	51	94	24		
8	52	98	28	+1	−1
9	52	94	28	−1	−1
10	53	94	36	−1	+1
11	51	98	36	+1	+1
12	54.5	96	32	0	0
13	55.5	96	32	0	0

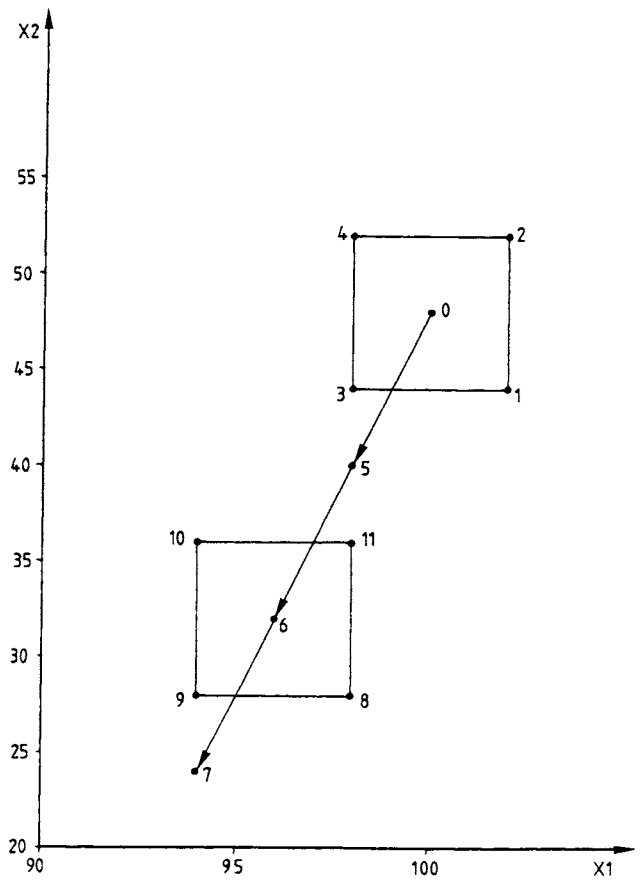


Fig. 26.5. The steepest ascent method for the data of Table 26.2.

obtains the steepest ascent path given in Fig. 26.5. Along it one selects a number of additional experiments. The length of the step can be based on intuition or, as proposed by Brooks [9], $2\sqrt{(b_1^2 + b_2^2)}$ (in scaled units). In our case we choose to carry out experiment 5. Since this yields clearly better results than for the experiments 0 to 4 one continues in the same direction with experiment 6 and 7. The response of experiment 7 seems to indicate that one has gone beyond the optimum. Therefore, one carries out a new 2^2 factorial experiment around the provisional optimum of experiment 6. Replication of point 6 (experiments 12 and 13) and a curvature check show that the surface is curved in this region. One concludes that the optimum is situated in this neighbourhood. If one wants to know more exactly where it is situated, one could build a central composite design (see Chapter 24) using points 6 (and its replicates 12, 13), 8, 9, 10, 11 as central, respectively factorial points and adding four star points.

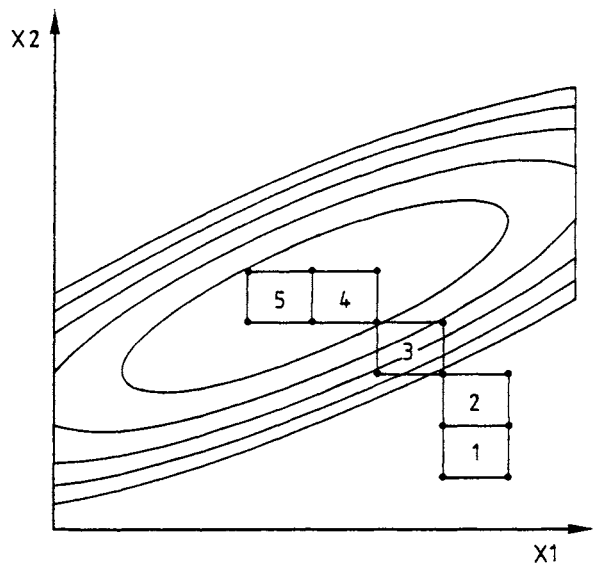


Fig. 26.6. Box-type factorial evolutionary operation.

Another simple procedure, called *evolutionary operation* (EVOP) was proposed first by Box [10]. The method has mainly been used for the optimization of industrial processes. It is illustrated by Fig. 26.6. The principle is that one describes the region around the starting point by a factorial design (here a 2^2 design). This factorial design is used as in the steepest ascent method to determine the coefficients of the model given by eq. (26.7). From this model one can then derive the direction of steepest ascent. Instead of carrying out experiments along the line of steepest ascent one carries out a new factorial experiment, which makes use of some of the experiments constituting the first design (see Fig. 26.6). If one prefers the use of second degree equations, the Doehlert design can be used in a similar way (see Section 24.3.4).

26.4 Multicriteria decision making

26.4.1 Window programming

Multicriteria decision making (MCDM) is applied when more than one response has to be taken into account. Often, this requires finding optimal compromises. *Window programming* is applied in some specific cases, such as the optimization of selectivity in chromatography. Although it can be applied in multidimensional situations, it is often used in univariate optimization when there are several

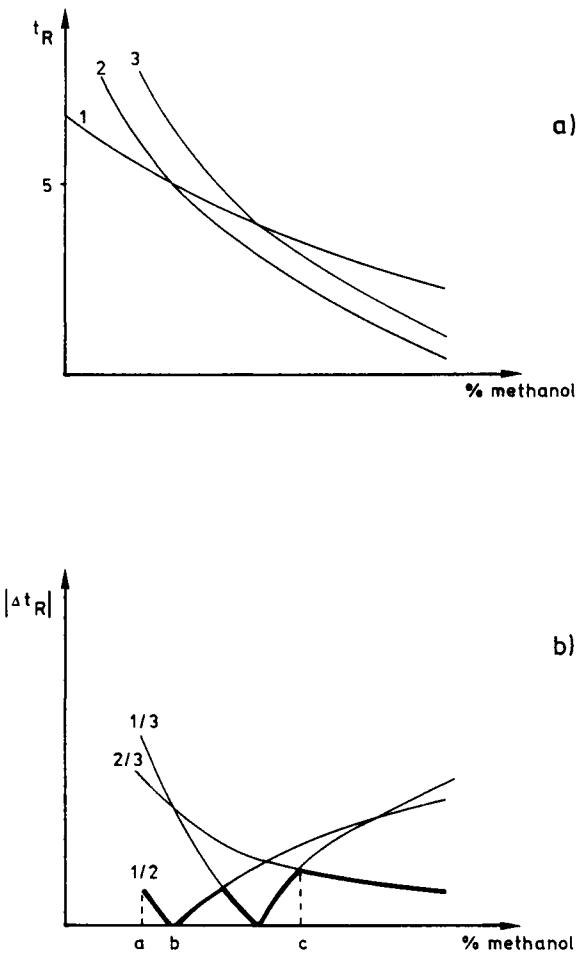


Fig. 26.7. Window programming: (a) retention times t_R of substances 1, 2 and 3 in function of % methanol; (b) $|\Delta t_R|$ for the pairs of substances 1/2, 1/3 and 2/3 in function of % methanol. The fat line is min $|\Delta t_R|$. Point c yields the optimal result.

responses to be optimized. It can be described as a minimax approach. One determines for all possible experimental conditions, the response which gives the worst result. The value of this response for the set of experimental conditions i is $y_{\min,i}$. Then, for all possible experimental conditions ($i = 1, n$), one determines that with the highest $y_{\min}$. The criterion is therefore $\max_{i=1,n} (y_{\min,i})$.

For univariate optimization in chromatography, this method, pioneered by Laub and Purnell [11], is simple but powerful. Let us consider Fig. 26.7. In Fig. 26.7a t_R, the retention time, of three hypothetical substances is given as a function of one

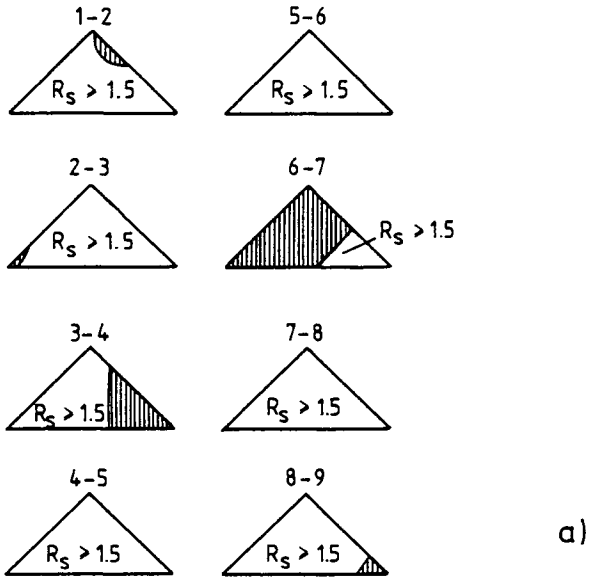
variable, the percentage methanol in the solvent. For three substances as in Fig. 26.7 this is a three-criterion problem. Indeed, one must consider the quality of the three binary separations (substances 1 and 2, 1 and 3, and 2 and 3). For simplicity, we suppose that the quality of the separation between two substances is given by the absolute difference in retention times $|\Delta t_R|$. We do this to be able to explain the method in a simple way. In fact, the criteria in chromatography are more complex and in real applications one uses criteria such as resolution to express how good a separation between two substances is.

In Fig. 26.7b we plot $|\Delta t_R|$ as a function of percentage methanol for each pair of substances. These lines cannot be described easily by a simple linear or quadratic function because some of them have an intermediate value of 0 where the order of elution of the two substances changes (a cross-over). Incidentally this is why one should be careful in modelling complex criteria (see warning in Section 21.4). $|\Delta t_R|$ cannot be properly modeled, but one can first model t_{R1} and t_{R2} separately, in this case with a quadratic function, and, afterwards, compute for each value of the dependent variable the composite criterion $|\Delta t_R|$.

In point a, substances 1 and 3 and 2 and 3 are separated much better than 1 and 2. One can conclude that in that point the global quality is determined by the quality of the separation of the worst separated pair. This means that in each point our criterion for the whole chromatogram is min $|\Delta t_R|$. In point b, for instance, min $|\Delta t_R|$ is zero, because 1 and 2 are not separated at all. The thick line is the one describing min $|\Delta t_R|$ ($= y_{min}$) as a function of % methanol. This thick line describes a series of windows, which explains the name of the technique. The highest point on that thick line or highest point on any window is c. This is the point where min $|\Delta t_R|$ is largest $\left(= \max_{i=1,n}(y_{min,i})\right)$ and therefore this is the optimum. In practice one would probably prefer to work at slightly higher % methanol than c because the chromatogram is then more robust (see Section 26.5).

26.4.2 Threshold approaches

Threshold approaches are widely used in experimental design. One does not try to find the optimum, but a region where all responses have acceptable values. The value to be reached is called the threshold. The most common way of finding the region is to use maps, either an isocontour map for two variables or a mixture triangle. One makes these maps first for each response separately and crosses out those areas where the threshold is not reached. Then one superimposes all those maps. The area which has not been crossed out in any of the maps is then the acceptable area. If necessary, one can work in stages. In a second iteration, one can give more desirable values to one, more or all of the responses and thereby pinpoint still better conditions.



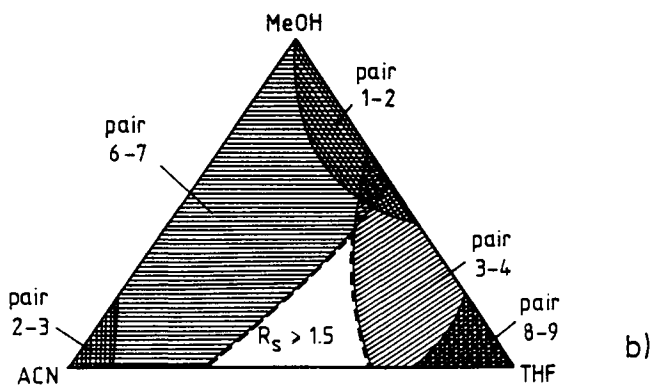


Fig. 26.8. Overlapping resolution map in reversed-phase chromatography (adapted from [12]). (a) Domains, in which separation is possible for successive pairs of substances, are left blank; (b) superposition of the maps in a.

An example of this method comes from chromatography [12] where the method is known as the *overlapping resolution map method.* It is used to find an acceptable solvent in HPLC. The mobile phase is a mixture of three solvents. In a first map (see Fig. 26.8) one crosses out all areas where the first and the second peak are not

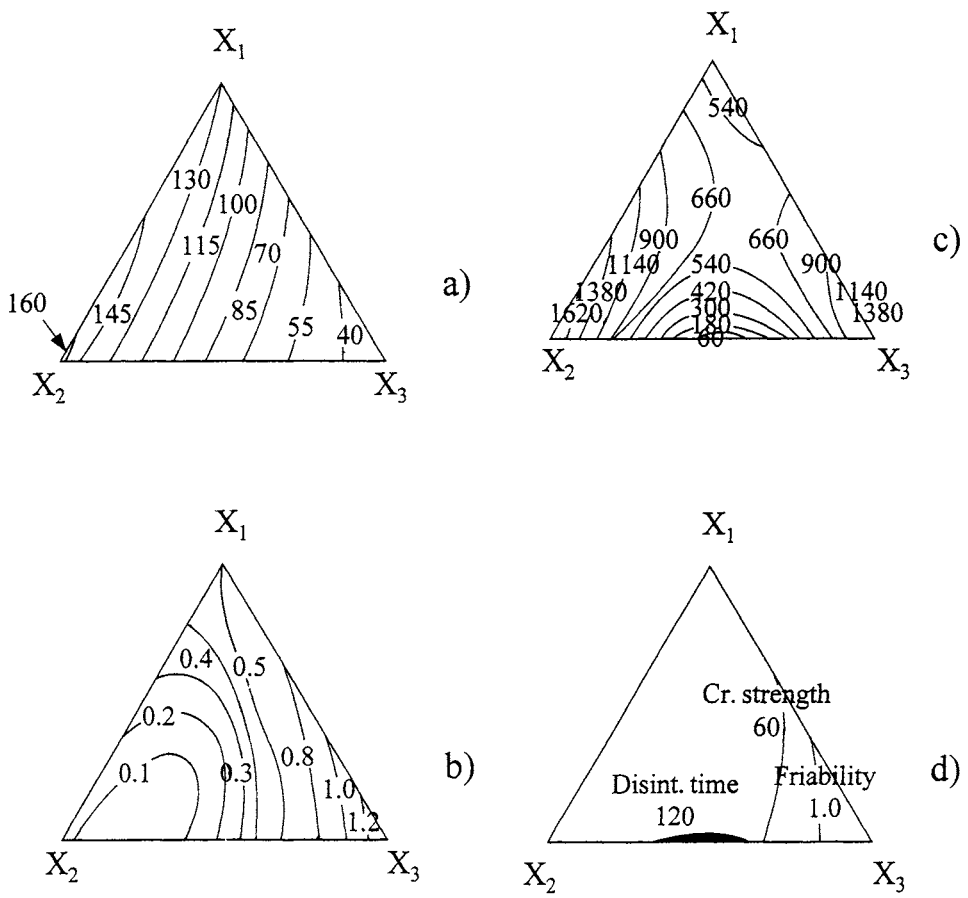


Fig. 26.9. Combined contour plot for a formulation example. x_1 = anhydrous lactose, x_2 = avicel, x_3 = α-lactose. (a) Contour plot for crushing strength; (b) contour plot for friability; (c) contour plot for disintegration time; (d) combined contour plot for the limits given in the figure. The black area is the acceptable area.

sufficiently separated (for instance $R_S < 1.5$). To do that one models (see Chapter 25) the retention of the relevant peaks and computes the resolution over the whole area from those models. In the second map this is done for the second and third peak, etc. Then the overlapping map is obtained by superposition of all the binary separation maps and the separation conditions retained as acceptable are those in the blank area of the overlapping resolution map.

Another example, from pharmaceutical technology [13] this time, is shown in Fig. 26.9. A tablet formulation is optimized for three responses (crushing strength, friability and disintegration time). The factors are the relative amounts of the excipients anhydrous lactose (x_1), avicel (x_2) and α-lactose (x_3) (in fact, they are

pseudocomponents — see Chapter 25). The contour plots for the individual criteria and a *combined contour map* with as constraints crushing strength >60 N, friability <1%, disintegration time <120 s are shown. The black area in Fig. 26.9d shows the domain in which all these criteria are obeyed. A formulation with 50% x_2 and 50% x_3 is preferred.

26.4.3 Utility functions

Suppose there are m criteria $y_j(y_1, ..., y_m)$ to be optimized. For experiment i, they take the values y_{1i}, ..., y_{ji}, ..., y_{mi}. Suppose also that it is possible to express numerically the importance of the criteria by weights $w_j(w_1, ..., w_m)$, then one obtains a *utility function*:

$$U_i = \sum_{j=1}^{m} w_j y_{ji} \tag{26.8}$$

The multicriteria problem is then reduced to the single criterion problem of optimizing U_i. Unfortunately, this approach has some important disadvantages, namely:

(a) it is very difficult to give *a priori* weights for all the criteria;

(b) it is possible that the multicriteria optimum found leads to an unacceptable value of one or more of the criteria: it can happen that very good solutions are found for one of the criteria with high weight, so that the bad results for some of the other criteria are compensated.

This method has nevertheless been used extensively in chromatography. An example is the CRF (chromatographic response function) [7]:

$$\mathrm{CRF} = \sum_{j=1}^{m-1} \mathrm{R_s}(j,j+1)\, n^a + b(t_{\max} - t_m) - c(t_{\min} - t_1) \tag{26.9}$$

where m = number of peaks expected; n = number of peaks detected; $\mathrm{R_s}(j,j+1)$ = resolution between peaks j and $j+1$; t_1, t_m = times for first and last peaks; $t_{\min}$, $t_{\max}$ = times desired for first and last peak; a, b and c = weights.

The first term sums $m-1$ criteria (resolutions), while the two last terms consider as criteria how close t_1 and t_m are to the ideal situation. The difficulty, of course, is to determine sensible values for a, b and c.

26.4.4 Derringer functions

Instead of summing the criteria, one can multiply them. Harrington [14] proposed to scale the values of the criteria between 0 (unacceptable) and 1 (optimal). These values are then called *desirabilities*. For instance, if resolution in chromatography should be at least 1 and, if 2 is the target to be reached, then

all resolution values equal to or smaller than 1 would have a desirability $d = 0$, all resolution values equal to or higher than 2 a desirability of 1 and a resolution of 1.6 would then have desirability $(1.6 - 1)/(2 - 1) = 0.6$. Derringer and Suich [15] adapted this as follows:

$$\begin{aligned} d &= 0 && \text{for } y \leq y_{min} \\ d &= 1 && \text{for } y \geq y_{max} \\ d &= [(y - y_{min})/(y_{max} - y_{min})]^r && \text{for } y_{min} < y < y_{max} \end{aligned} \tag{26.10}$$

where d needs to be maximized, y_{min} is the minimum acceptable value of y, y_{max} the value beyond which improvement is of no further interest and r is a coefficient to be determined by the user. The effect of using different values of r is shown in Fig. 26.10. The authors give also transformations that can be used when a certain value of y is wanted and both a lower and a higher value are less desirable.

The global desirability D is then obtained as

$$D = \Pi d_j \tag{26.11}$$

where d_j are the values for individual criteria.

This procedure has the advantage that if one of the criteria is unacceptable (i.e. = 0), then the global desirability is also 0. With the weighted sum of eq. (26.8), this is not the case. As explained, a very good value for one criterion can mask an unacceptable value for another criterion when utility functions are used. The use

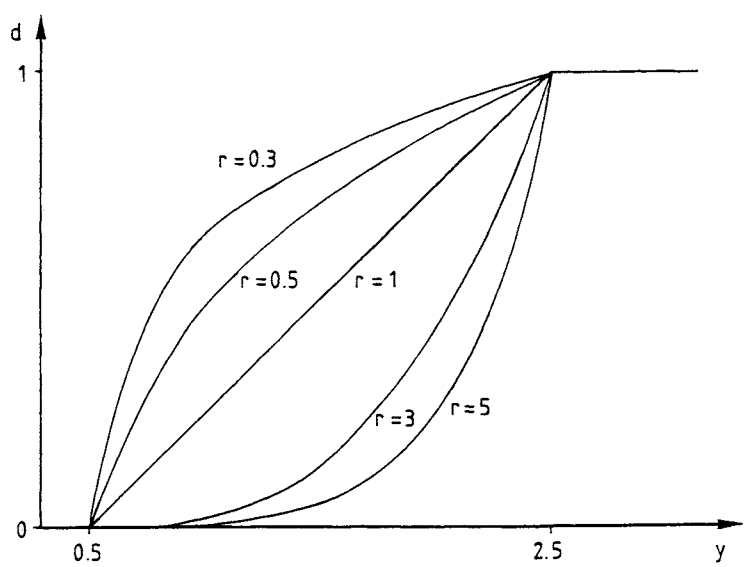


Fig. 26.10. Values of Derringer and Suich's d for different values of r, $y_{min} = 0.5$, $y_{max} = 2.5$.

of Derringer's method in chromatography was introduced by Bourguignon and Massart [16].

26.4.5 Pareto-optimality methods

The Pareto-optimality principle is due to Vilfredo Pareto (see also Section 2.3.4). It was introduced in chemometrics by Smilde et al. [17]. In this section, we will follow the argumentation of Keller et al. [18]. Pareto-optimality is defined as follows.

An experiment is Pareto-optimal if there is no other experiment which has a better result on one criterion without having a worse result on another.

Let us suppose that there are two criteria, y_1 and y_2, to be optimized and that optimization for both of them means maximization. Experiment 1 has yielded $y_1 = 100$ and $y_2 = 60$ and experiment 2 $y_1 = 100$ and $y_2 = 50$. Experiment 1 is better for y_2 than experiment 2, without being worse for criterion y_1. Experiment 1 is Pareto-optimal and is said to dominate experiment 2. Suppose now that experiment 1 would have yielded $y_1 = 60$ and $y_2 = 100$. In that case experiment 1 is still better for y_2 but now it is worse for y_1. It is now impossible to decide on a pure numerical basis. The numerical difference for y_2 (50–100) is larger in absolute value than that for y_1 (100–60), but it may well be that y_1 is much more important for practical reasons: only an expert in the field studied can decide whether experiment 1 is to be preferred to experiment 2.

The Pareto-optimality concept can be used with more than two responses but with two it is most attractive, because of the graphical interpretation (Figs. 26.11 and 26.12). When there are many responses, it becomes increasingly unlikely that one experiment will dominate another for all responses and the method is less useful. Let us consider therefore the two-dimensional response case of Table 26.3. Consider for example experiments 5 and 4. The latter is dominated by the first. If one considers experiments 8 and 7, the latter is also dominated by the former. Experiments 5 and 8 both lie on the border of the cloud of points in Fig. 26.11 in the direction of high y_1 and high y_2 (we are maximizing both). The line drawn through the border points (5, 6, 8, 9 and 10) connects experiments that dominate all the experiments below it. This is so for the actual experiments, but also for potential experiments, that were however not carried out. An expert will have to decide which, among the points along the line offers the best compromise between y_1 and y_2. It may well be that he decides that this is situated half-way between experiments 6 and 8. In that case, it will be necessary to decide which experimental conditions would yield such values for y_1 and y_2.

Suppose that y_2 is to be maximized, while y_1 is minimized. This would be typical for chromatography, where one wants to maximize separation quality (for instance, resolution) and minimize time. In that case, one would decide that the borderline constituted by points 1, 2 and 5 links the Pareto-optimal points (Fig. 26.12).

TABLE 26.3

Data set for explaining Pareto-optimality

Experiment	y_1	y_2
1	1	4
2	2	9
3	3	3
4	4	8
5	4	10
6	7	9
7	7	7
8	9	7
9	10	3
10	11	1

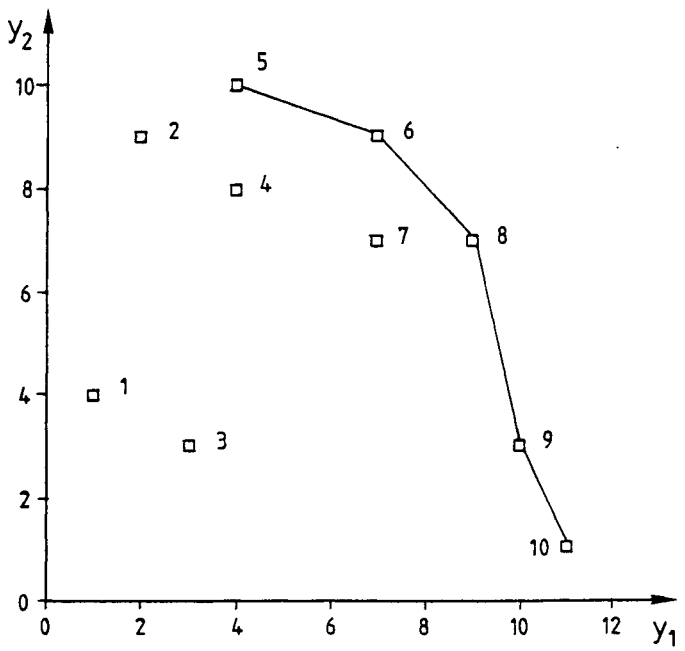


Fig. 26.11. Pareto-optimality. Responses y_1 and y_2 must be maximized. All Pareto-optimal solutions are found on the solid line.

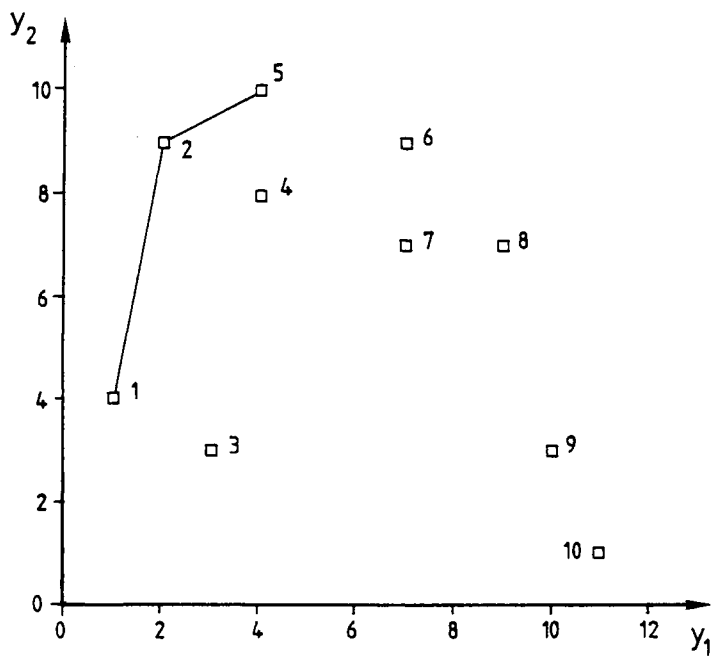


Fig. 26.12. Pareto-optimality. Response y_1 is to be minimized, while y_2 is maximized. All Pareto-optimal solutions are found on the solid line.

26.4.6 Electre outranking relationships

In the Pareto-optimality procedure the relation between two experiments can be described only in two ways. Either one experiment dominates another one and it is clear which one is the better, or else the two cannot be compared. This may be too simple for some situations. Consider for instance the following situations described by Brans and Mareschal [19].

Situation 1: experiment A scores 100 on criteria y_1 and y_2, B scores 30 and 20; A is Pareto-optimal compared to B because it dominates B.

Situation 2: experiment A scores 100 on y_1 and 20 on y_2 and B scores 30 on y_1 and 100 on y_2; A and B are incomparable.

Situation 3: A scores 100 on y_1 and 99 on y_2 and B scores 1 on y_1 and 100 on y_2. In the Pareto sense, the two experiments are incomparable. However, the difference in y_2 is so small, that one might reason that it is much less important than the difference in y_1. In that case one would prefer A to B.

Electre [20], and also Promethee (see next section) can work with preferences. To explain the method, we will apply it to data extracted from a study by Chardon et al. [18,21]. The original study consisted of the optimization of properties of bedsheets in function of resin, softener content and softener composition (see also

TABLE 26.4

Table of criteria for the evaluation of bedsheets [18,21]

j	Criterion	Weights	Possible criterion values
1	Smoothness	1	6, 8, 9
2	Hand	1	8, 9, 10
3	Hydrophilicity	1	45, 40, 35, 30, 20
4	Soil release	1	60, 65, 70, 80

TABLE 26.5

Criterion values for some experiments [18,21] to optimize the properties of bedsheets

Criterion	Optimization goal	Evaluation of bedsheets					
		Experiment					
		21	24	44	61	62	64
1	max	6	6	9	8	8	8
2	max	10	10	8	9	9	9
3	min	30	40	45	35	45	20
4	max	60	80	70	65	80	80

Chapter 25). The whole design required 49 experiments and 5 responses were measured. Of the 49 experiments only 6 fulfilled the minimum requirements for all criteria. For these six experiments, one criterion gave the same value for all experiments and therefore only four criteria are considered further.

Table 26.4 gives the criteria j, the weights w_j assigned to these criteria (here we considered all criteria to be equally weighted) and the values the criteria can take. In Table 26.5 the experimental values obtained and the optimization goals are given. One now compares each experiment with each other experiment. This is done in two steps. In the first step, one notes in what respect the experiments differ. Suppose we compare experiments 21 and 24, then 21 is better according to criterion 3, and worse according to criterion 4. The results are given in Table 26.6.

In the second step, one takes into account the weights in order to arrive at a numerical expression. The preference ratio of experiment A over B is given by

$$P = \frac{\sum_{j \in N^+} w_j}{\sum_{j \in N^-} w_j} \tag{26.12}$$

TABLE 26.6

Comparison of the experiments of Table 26.5 according to the criterion values. Bold numbers indicate criteria for which the row experiment is better than the column experiment, italic numbers indicate those criteria for which the opposite is true

	21	24	44	61	62	64
21	–	**3**, *4*	**2,3**, *1,4*	**2,3**, *1,4*	**2,3**, *1,4*	**2**, *1,3,4*
24	–	–	**2,3,4**, *1*	**2,4**, *1*,3,	**2,3**, *1*	**2**, *1,3*
44	–	–	–	**1,4**, *2,3*	**1**, *2,4*	**1**, *2,3,4*
61	–	–	–	–	**3**, *4*	*3,4*
62	–	–	–	–	–	*3*
64	–	–	–	–	–	–

TABLE 26.7

Numerical comparison of the experiments of Table 26.5. Preference ratios are given for experiments in column (1) over those in columns (2) to (7)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	21	24	44	61	62	64
21	–	–	–	–	–	–
24	–	–	3	–	2	–
44	–	–	–	–	–	–
61	–	–	–	–	–	–
62	–	–	2	–	–	–
64	3	2	3	∞	∞	–

where N^+ symbolizes the criteria for which A is better than B and N^- those for which the inverse is true. For example, for the preference of 21 over 64, this becomes

$$P = 1/3 = 0.33$$

We can now construct Table 26.7, where only the values that exceed 1 are given. To return to the example, as $P = 0.33$ for the preference of 21 over 64, then $P = 1/0.33 = 3$ for the preference of 64 over 21.

Until now, we have taken into account only the fact that one experiment has a better value for some criterion or not. It is possible that one experiment is so much worse according to one criterion that, even when it is better in all other respects, we do not wish to conclude that it is better. To do this, we can add discrepancy conditions. In the present example we have not included this possibility.

At this stage, a dominance threshold, T, can also be introduced which must be at least 1 and is usually higher. The idea is that it is preferable not to judge one

experiment to be better than the other when only a slight difference between both is obtained. In this way, the uncertainty involved in choosing the w_j values is taken into account and the fact that some other criteria may have been overlooked. In the present example, T is considered to be 1.33 and all values that do not exceed this threshold are eliminated. In our simple example no experiments are eliminated and Table 26.8 is obtained, in which there is now a summary of those instances where one procedure is clearly better than another. These procedures dominate the others (symbol D). For example, in Table 26.8 it can be seen that experiment 64 dominates all others, while 24 dominates 62 and 44. We also say that 64 *outranks* all other experiments, while 24 outranks 62 and 44. From this table, a *dominance graph* is constructed, where $1 \rightarrow 2$ means that 1 dominates 2. This graph is shown in Fig. 26.13.

TABLE 26.8

Dominance table for the experiments of Table 26.5. Dominance is given for experiments in column (1) over those in columns (2) to (7).

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	21	24	44	61	62	64
21	–	–	–	–	–	–
24	–	–	D	–	D	–
44	–	–	–	–	–	–
61	–	–	–	–	–	–
62	–	–	D	–	–	–
64	D	D	D	D	D	–

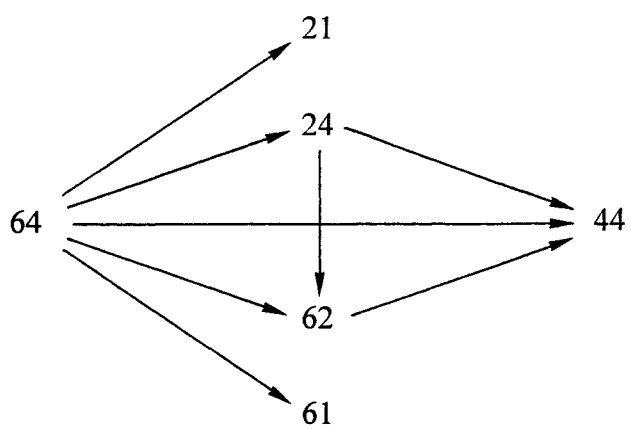


Fig. 26.13. Dominance graph for the example of Tables 26.4–26.8. Experiment 64 dominates all others.

26.4.7 Promethee

Promethee, in the same way as Electre, aims to obtain outranking relationships, but is more refined than Electre. The latter simply counts the number of criteria for which one solution is better than another, but Promethee quantifies the degree of preference of one solution compared with another for each criterion. Promethee was originally developed by Brans et al. [19] for location problems, i.e. for the selection of one location for a factory, warehouse, etc. from several alternatives. The method was introduced into chemometrics by Keller et al. [18]. The first step of the procedure is to define preference functions for each criterion. Promethee includes many types of functions. For instance, the function of Fig. 26.14 can be used which is characterized by the following equations:

$$\begin{aligned} P(\mathrm{A,B}) &= 0 \qquad \text{for } d \leq 0 \\ P(\mathrm{A,B}) &= d/z \qquad \text{for } 0 < \mathrm{d} \leq z \\ P(\mathrm{A,B}) &= 1 \qquad \text{for } d > z \end{aligned} \tag{26.13}$$

where $d = y_A - y_B$.

P(A,B) characterizes the preference of solution A over B. This means there is also a P(B,A) describing the preference of B over A. The criterion is the value of d. When $d \leq 0$, A is not preferred over B and P(A,B) = 0. When $d > z$, A is so strongly preferred over B that a further increase in d is considered to have no further influence in the preference intensity. When d is situated between 0 and z, the preference depends linearly on d. In the next step, one sums the P_j(A,B) values for the m criteria j and for each pair of solutions A,B and B,A. This sum can be

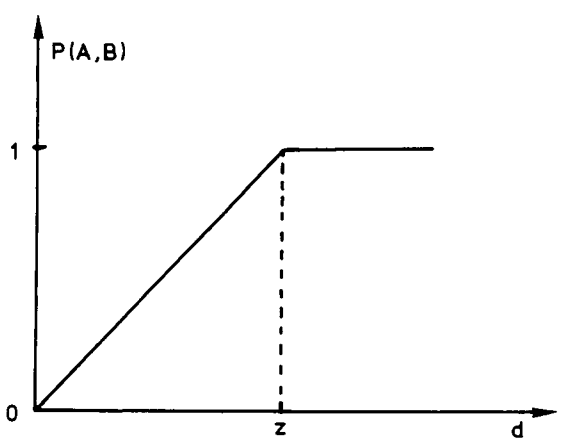


Fig. 26.14. A preference function describing the preference P(A,B) for experiment A over experiment B in function of the criterion value d using eq. (26.13).

TABLE 26.9

A two-criterion example (from Ref. [18])

Experiment	y_1	y_2
K	4	100
L	7	90
M	8	20

TABLE 26.10

Preference table for the data of Table 26.9

Comparison	d_1	d_2	P_1	P_2	Π
K,L	–3	10	0	0.5	0.25
L,K	3	–10	1	0	0.50
K,M	–4	80	0	1	0.50
M,K	4	–80	1	0	0.50
L,M	–1	70	0	1	0.50
M,L	1	–70	0.5	0	0.25

weighted with weights w_j for criterion j. In that case the weights are chosen such that their sum over all m criteria is 1.

$$\Pi(\mathrm{A,B}) = \sum_{j=1}^{m} w_j P_j(\mathrm{A,B}) \tag{26.14}$$

where

$$\sum_{j=1}^{m} w_j = 1$$

Let us consider a very simple example where 3 experiments K, L and M were carried out for 2 criteria, y_1 and y_2 (Table 26.9). Suppose that $z_1 = 2$ and $z_2 = 20$. Then for y_2, $d_2(\mathrm{K,L}) = 100–90 = 10$ and $P_2(\mathrm{K,L}) = 10/20 = 0.5$. The values for d, P and Π are given in Table 26.10 for the case of equal weights $w_1 = w_2 = 0.5$. In a third step we compute so-called outranking flows.

$$\Phi^{+}(\mathrm{A}) = \sum_{i \in n} \Pi(\mathrm{A},i) \tag{26.15}$$

$$\Phi^{-}(\mathrm{A}) = \sum_{i \in n} \Pi(i,\mathrm{A})$$

TABLE 26.11

Outranking flows for the data of Table 26.9

Experiment	Φ^+	Φ^-
K	0.75	1.00
L	1.00	0.50
M	0.75	1.00

$\Phi^+(A)$ is the positive outranking flow, which describes to what extent experiment A outranks all other experiments i, while the negative outranking flow $\Phi^-(A)$ is the extent to which A is outranked by other experiments. The results for our simple example are shown in Table 26.11.

We now carry out pairwise comparisons, using the following set of rules.
A outranks B if:

$$\Phi^+(A) > \Phi^+(B) \text{ and } \Phi^-(A) < \Phi^-(B)$$

or

$$\Phi^+(A) > \Phi^+(B) \text{ and } \Phi^-(A) = \Phi^-(B) \tag{26.16}$$

or

$$\Phi^+(A) = \Phi^+(B) \text{ and } \Phi^-(A) < \Phi^-(B)$$

A is indifferent to B if:

$$\Phi^+(A) = \Phi^+(B) \text{ and } \Phi^-(A) = \Phi^-(B)$$

Otherwise A is incomparable with B.

The net flow is defined by

$$\Phi(A) = \Phi^+(A) - \Phi^-(A)$$

In the above example, L outranks K and M, and K and M are indifferent. One should prefer L. If, for some reason, this becomes impossible, one can prefer either M or K.

In this example, we have applied equal weights for the criteria, but this is not necessary. The Promethee procedure includes a sensitivity analysis called Gaia [19], which allows us to investigate the effect of the weights through a principal component analysis procedure on the net flows. An application to the data of Table 26.5 is given in [18].

26.5 Taguchi methods

In Section 2.3 we discussed process capability indices and concluded that both deviation from target value and dispersion around target values should be small enough. Taguchi [22] addressed the question of how to use experimental design to achieve minimum dispersion around the target value and thereby increase the robustness of a product. We have already introduced the notion of robustness in chemical analysis in Chapter 13 and have shown that in chemical analysis experimental design can be applied to identify sources of non-robustness (Chapter 23). Taguchi's ideas have had a profound effect on the thinking about quality. In this section we will describe only the ideas that have a direct bearing on chemometrics. They consist essentially of two elements, namely a response that also includes dispersion and experimental designs that allow us to optimize (in fact, maximize) that response [22–26].

26.5.1 Signal-to-noise ratios

A response is characterized both by its value y and by the dispersion of the values when the experiment is replicated. In that case, the value of the response is given by $\bar{y} = \sum y_i / n$ and the dispersion by $s^2 = \sum(y_i - \bar{y})^2 / (n - 1)$. In this context $\bar{y}$ is called the signal and s the noise and Taguchi's proposal was to maximize the *signal-to-noise ratio* given as

$$Z = 10 \log(\bar{y}^2 / s^2) \tag{26.17}$$

This is equivalent to minimizing the relative standard deviation.

Taguchi divides the factors that have an influence on the responses in two categories, namely:

(1) *control factors*, that influence the robustness of the process as measured by Z;

(2) *signal* or *adjustment factors*, that do not influence Z, but have an effect on $\bar{y}$. In a statistical process control (SPC) setting these factors can be used to set the process such that the mean value is on target.

The control and signal factors together are called the *design factors*.

26.5.2 Inner and outer designs

Taguchi proposes to investigate the effect of the factors on the response, $\bar{y}$, and on the variance, s^2, by carrying out an experimental design, usually a two-level (full or fractional) factorial or a three-level factorial (orthogonal array) design. This design is called the *inner design* and Taguchi also uses his own terminology to describe the designs. For instance, his L8 orthogonal array turns out to be a 2^{7-4} (III) design. To obtain an estimate of the variance, in the simplest case the

TABLE 26.12

Factors studied in the leaf spring example (from Ref. [26])

Factor symbol	Factor	Low level	High level
B	High heat temperature (°F)	1840	1880
C	Heating time (seconds)	25	23
D	Transfer time (seconds)	12	10
E	Hold down time (seconds)	2	3
O	Quench oil temperature (°F)	130–150	150–170

experiments are replicated. If we want to measure the robustness of the response at the experimental conditions studied, we can instead in each point carry out a design (often a screening design — see Chapter 23), which is called the *outer design*. This design describes the effect of small changes in all or some of the factors that were also studied in the inner design or of other factors. The variance of the responses measured in each outer design is then the s^2 of the inner design.

An example can be found in the article of Pignatiello and Ramberg [26]. They applied a Taguchi design to optimize a product (a leaf spring) with a target value of 8 inches. Deviations in both directions were undesirable. They wanted to know which factors could be controlled in view of obtaining the desired 8 inches. Four possible such factors (B, C, D and E — see Table 26.12) were identified. Another factor (O) was considered difficult to control and is therefore treated as an adjustment factor.

A 2^{4-1} fractional factorial design was carried out (see Table 26.13) as the inner design. The outer design consists of three replicates at the O^- and three at the O^+ level. The engineers in charge of the project were interested not only in the main

TABLE 26.13

Results (in inches) obtained in the Taguchi experiment for the leaf spring example

Run	Factor				Results (y)					
	B	C	D	E	O^-			O^+		
1	–	–	–	–	7.78	7.78	7.81	7.50	7.25	7.12
2	+	–	–	+	8.15	8.18	7.88	7.88	7.88	7.44
3	–	+	–	+	7.50	7.56	7.50	7.50	7.56	7.50
4	+	+	–	–	7.59	7.56	7.75	7.63	7.75	7.56
5	–	–	+	+	7.94	8.00	7.88	7.32	7.44	7.44
6	+	–	+	–	7.69	8.09	8.06	7.56	7.69	7.62
7	–	+	+	–	7.56	7.62	7.44	7.18	7.18	7.25
8	+	+	+	+	7.56	7.81	7.69	7.81	7.50	7.59

TABLE 26.14

Interaction effects and Taguchi statistics for the data of Table 26.13

Run	Factor							Statistics		
	B	C	D	E	BC	BD	CD	$\bar{y}$	s^2	Z
1	–	–	–	–	+	+	+	7.54	0.09	28.00
2	+	–	–	+	–	–	+	7.90	0.07	29.46
3	–	+	–	+	–	+	–	7.52	0.001	47.70
4	+	+	–	–	+	–	–	7.64	0.01	38.68
5	–	–	+	+	+	–	–	7.67	0.09	28.11
6	+	–	+	–	–	+	–	7.79	0.05	30.59
7	–	+	+	–	–	–	+	7.37	0.04	31.55
8	+	+	+	+	+	+	+	7.66	0.02	35.31

effects of the factors named, but also in the interactions BC, BD and CD. It was considered that other interactions were not to be expected. Using the techniques described in Chapter 23, we can verify that the main effects are confounded with three-factor interactions and the possible interactions are not confounded with each other (see Table 26.14).

The Taguchi statistics are also given in Table 26.14. As an example, we compute these values for run 1:

$$\bar{y} = (7.78 + 7.78 + 7.81 + 7.50 + 7.25 + 7.12)/6 = 7.54$$

$$s^2 = [(7.78 - 7.54)^2 + (7.78 - 7.54)^2 + (7.81 - 7.54)^2 + (7.50 - 7.54)^2 + (7.25 - 7.54)^2 + (7.12 - 7.54)^2] / (6 - 1) = 0.09$$

$$Z = 10 \log_{10}(7.54^2/0.09) = 28.00$$

The effects (on Z) are given in Table 26.15. It turns out that the largest effect is due to C followed by the CD interaction. The Z-values are best at the highest C level and at the lower level of D and it was therefore concluded that these factors should be kept at these levels. The other factors are less important, although, due to the low number of degrees of freedom it is not clear whether they are significant or not. It was decided to investigate whether one might use them to fine-tune the $\bar{y}$ values obtained and achieve results as close as possible to the target of 8 inches. For that reason an ANOVA was carried out on the experimental data, i.e. on the y values of Table 26.13. Indeed, until now the analysis has indicated at which levels of certain factors to work to obtain a good Z-value, but it does not tell us whether this is due to the $\bar{y}$ or to the s^2 term. In this ANOVA, one now includes O as a factor. Table 26.16 gives the results. The supposed adjustment factor O has an effect on $\bar{y}$ and should be kept at the lower level. The controlled factors B and E have an

TABLE 26.15

Effects (on Z) for the data of Table 26.14

Factor	Effect
B (+ CDE)	–0.33
C (+ BDE)	9.27
D (+ BCE)	–4.56
E (+ BCD)	2.94
BC (+ DE)	–2.30
BD (+ CE)	3.45
CD (+ BE)	–5.19

TABLE 26.16

Analysis of variance of the y-values of Table 26.13 (adapted from Ref. [26])

Source	df	SS	MS	F	P
B + CDE	1	0.587	0.587	35.44	0.000001
C + BDE	1	0.373	0.373	22.52	0.0004
BC + DE	1	0.004	0.004	–	–
D + BCE	1	0.010	0.010	–	–
BD + CE	1	0.005	0.005	–	–
CD + BE	1	0.015	0.015	–	–
E + BCD	1	0.129	0.129	7.79	0.01
O + BCDEO	1	0.809	0.809	48.85	<0.000001
BO + CDEO	1	0.086	0.086	5.19	0.03
CO + BDEO	1	0.328	0.328	19.80	0.0001
BCO + DEO	1	0.001	0.001	–	–
DO + BCEO	1	0.035	0.035	2.11	0.16
BDO + CEO	1	0.020	0.020	1.21	0.26
CDO + BEO	1	0.027	0.027	1.63	0.21
EO + BCDO	1	0.009	0.009	–	–
Residue	32	0.530	0.017		

Significant effects:
B (+ CDE): 0.221
C (+ BDE): –0.176
E (+ BCD): 0.104
O (+BCDEO): –0.260
BO (+ CDEO): 0.085
CO (+ BDEO): 0.165

effect on $\bar{y}$ and can be fine-tuned to achieve the closest value to 8 possible. C also has an effect on $\bar{y}$, but should not be used as a tuning factor, since it is fixed at the + level to maximize Z.

References

1. X. Wang, J. Smeyers-Verbeke and D.L. Massart, Linearization of atomic absorption calibration curves. Analusis, 20 (1992) 209–215.
2. D.E. Long, Simplex optimisation of the response from chemical systems. Anal. Chim. Acta, 46 (1969) 193–206.
3. W. Spendley, G.R. Hext and F.R. Himsworth, Sequential application of Simplex designs in optimization and evolutionary operations. Technometrics, 4 (1962) 441–461.
4. J.A. Nelder and R. Mead, A Simplex method for function optimization. Computer J., 7 (1965) 308–313.
5. S.L. Morgan and S.N. Deming, Optimization strategies for the development of gas-liquid chromatographic methods. J. Chromatogr., 112 (1975) 267–285.
6. K.W.C. Burton and G. Nickless, Optimization via Simplex. Part I. Background, definitions and a simple application, Chemom. Intell. Lab. Systems, 1 (1987) 135–149.
7. J.C. Berridge, Techniques for the Automated Optimization of HPLC Separations. Wiley, New York, 1985.
8. G.E.P. Box, W.G. Hunter and J.S. Hunter, Statistics for Experimenters. Wiley, New York, 1978.
9. S.H. Brooks, A comparison of maximum-seeking methods. Oper. Res., 7 (1959) 430–457.
10. G.E.P. Box, Evolutionary operation, A method for increasing industrial productivity. Appl. Stat., 6 (1957) 81–91.
11. R.J. Laub and J.H. Purnell, Criteria for the use of mixed solvents in gas-liquid chromatography. J. Chromatogr., 112 (1975) 71–79.
12. J.L. Glajch, J.J. Kirkland, K.M. Squire and J.M. Minor, Optimization of solvent strength and selectivity for reversed-phase liquid chromatography using an interactive mixture-design statistical technique. J. Chromatogr., 199 (1980) 57–79.
13. G.K. Bolhuis, C.A.A. Duineveld, J.H. de Boer and P.M.J. Coenegracht, Simultaneous optimization of multiple criteria in tablet formulation: Part I, Pharmaceutical Technology EUROPE, June 1995, pp. 42–49.
14. E.C. Harrington jr., The desirability function. Industrial Qual. Control, 21 (1965) 494–498.
15. G. Derringer and R. Suich, Simultaneous optimization of several response variables. J. Qual. Technol., 12 (1980) 214–219.
16. B. Bourguignon and D.L. Massart, Simultaneous optimization of several chromatographic performance goals using Derringer's desirability function. J. Chromatogr., 586 (1991) 11–20.
17. A.K. Smilde, A. Knevelman and P.M.J. Coenegracht, Introduction of multi-criteria decision making in optimization procedures for high-performance liquid chromatographic separations. J. Chromatogr., 369 (1986) 1–10.
18. H.R. Keller, D.L. Massart and J.P. Brans, Multicriteria decision making: a case study. Chemom. Intell. Lab. Systems, 11 (1991) 175–189.
19. J.P. Brans and B. Mareschal, PROMETHEE — GAIA Visual Interactive Modelling for Multicriteria Location Problems, Internal Report University of Brussels, STOOTW/244, 1989.
20. B. Roy, Electre. Revue METRA, 11 (1972) 121–131.
21. J. Chardon, J. Nony, M. Sergent, D. Mathieu and R. Phan-Tan-Luu, Experimental research methodology applied to the development of a formulation for use with textiles. Chemom. Intell. Lab. Systems, 6 (1989) 313–321.
22. G. Taguchi, Introduction to Quality Engineering: Designing Quality into Products and Processes. Kraus International Publication, White Plains, NY, 1986.

23. R.N. Kackar, Taguchi's quality philosophy: analysis and commentary. Quality Progress, December 1986, 21–24.
24. G.E.P. Box, Signal-to-noise ratios, performance criteria, and transformations. Technometrics, 30 (1988) 1–17.
25. R.N. Kackar, Off-line quality control, parameter design and the Taguchi method. J. Qual. Technol., 17 (1985) 176–188.
26. J.J. Pignatiello jr and J.S. Ramberg, Discussion. J. Qual. Technol., 17 (1985) 198–206.

Additional reading:

M.M.W.B. Hendriks, J.H. de Boer, A.K. Smilde and D.A. Doornbos, Multicriteria decision making. Chemom. Intell. Lab. Systems, 16 (1992) 175–191.

Chapter 27

Genetic Algorithms and Other Global Search Strategies

27.1 Introduction

In the late 1980s a new class of computational methods, the so-called natural computation methods were introduced in chemometrics. As their name suggests, these methods are inspired by a natural process. *Genetic algorithms* (GAs), *artificial neural networks* (ANN) and *simulated annealing* (SA) are the most prominent examples that have been studied in chemometrics. GAs are numerical optimization methods which simulate biological evolution. Simulated annealing (see Section 27.9) is an optimization method which simulates the gradual cooling of a solid to overcome local energy minima. Artificial neural networks are based on a model of the working of the brain and are treated in Chapter 44. In this chapter we focus on the principles and mechanisms of GAs and briefly describe simulated annealing (Section 27.9) and tabu search (Section 27.10).

In general, one can state that all search or optimization strategies are based on some assumptions about the search space. The weakest possible assumption is that each candidate solution can be evaluated. Examples of strategies that make use of this assumption only are:

- enumerative search (a systematic scan of the search space);
- random search.

Such methods are generally applicable but inefficient.

Other search strategies make strong assumptions about the response surface and are in fact tailored to the problem. They try to find an optimum as quickly as possible and search only in a local area of the search space. Such search strategies are called local. Examples are the gradient search methods, and Simplex optimization. Local methods are efficient but not very robust. They only find the optimum in the search space closest to the starting point, and hence are easily trapped in local optima. These methods are best suited for optimization problems for which enough prior knowledge is available.

Genetic algorithms are situated somewhere in between these two extremes. They are based on weak or moderate domain assumptions. The extent of the assumptions, e.g. the amount of heuristic knowledge about the problem can be adjusted for each case to the desired level. When the target problem is too complex

to be tackled by methods based on strong domain assumptions, genetic algorithms can still produce reliable results. With GAs it is e.g. possible to overcome to a great extent the disadvantages of the optimization methods such as Simplex, described in Chapter 26. They are able to search large problem spaces with a much smaller risk to be trapped in local optima. This advantage is paid for by a computational effort, which may be large. The recent interest of many researchers in these methods can be largely ascribed to the development of faster (e.g. parallel) and cheaper computer systems. Therefore GAs have become very useful in computational optimization problems. When the optimization procedure requires time consuming or expensive experiments it is better to use local optimization methods that focus on a minimal number of experiments.

27.2 Application scope

Genetic algorithms are useful to treat numerical optimization problems that are difficult to solve with classical methods. The degree of difficulty is mainly determined by two factors: the size of the search space and the presence of suboptima. Although the applications of GAs are situated in numerous scientific fields, they can be approximately subdivided into three major categories: numerical problems, sequencing problems and subset selection problems.

Problems in the first category consist of solving numerical calculation models. An example in chemistry is curve fitting of, e.g., an IR spectrum using a number of Gaussian peaks. Especially when the number of peaks is not known, this becomes a huge search space with many different local optima.

The typical example of sequencing problems is the travelling salesperson (TSP) problem. The problem consists of finding the shortest route for a person to visit a number of cities, such that every city is visited only once. While this problem is extremely simple to state it is still a major challenge to solve. To visit N cities there are $(N - 1)!$ (more than 10^{10} for $N = 15$) possibilities, with many almost equally good solutions. This type of optimization problems is not so common in chemistry.

An example from the third category is feature selection; e.g. the selection of a number of wavelengths from a spectrum for multicomponent calibration. When the number of features to select from is large and the optimal number of features to select is not known, the search space becomes huge and contains many local optima.

In the following sections the basic principles of GAs are explained with examples from each of the previous categories.

27.3 Principle of genetic algorithms

Biological evolution has been particularly successful in the design and creation of amazingly complex organisms. The following mechanisms are important for biological evolution [1]:

- it is based on Darwin's classical rules about natural evolution: struggle for life (competition rule) and survival of the fittest (selection rule);
- it acts on an encoded form of life, the chromosomes, rather than on the living beings themselves. Random changes are introduced by natural mutation.

This has stimulated several scientists to simulate this process on computers in order to solve complex computational problems. Holland [2] was one of the pioneers in the development of genetic algorithms.

To simulate biological evolution in order to perform optimization tasks, suitable evolutionary components must be designed:

- an encoding technique for the candidate solutions;
- competition between these candidate solutions;
- recombination between surviving solutions so that a new generation of possibly better solutions can be produced that replace the existing one;
- mutation: random changes.

We focus in this chapter on the traditional genetic algorithm, the so-called Simple Genetic Algorithm (SGA), as introduced by Holland [2]. We will, however, mention some important variants and extensions that may be useful for solving chemical problems [2–4].

27.3.1 Candidate solutions: representation

The candidate solutions of the problem must be computer coded. In the traditional genetic algorithm these are bit strings, containing zeros and ones, that are manipulated further by the algorithm [5]. We will explain this by means of some examples. The first example is a simplification of an application described in the literature [6]. It concerns the optimization of the pH and the composition of the mobile phase for an HPLC separation. The capacity factor, k, of the different solutes can be calculated by means of a non-linear model. From the k-values the separation quality can be derived by means of, e.g., calculating the resolution. This simplification is artificial because this problem can easily be solved by conventional optimization and experimental design methods. However, it allows us to explain in a simple way the principles of the method.

Candidate solutions are all valid combinations of the parameters to be optimized. In this example they can be represented as a two-dimensional vector, encoded as a string in the computer. The collection of the candidate solutions constitutes in this case a two-dimensional search space (see Fig. 27.1).

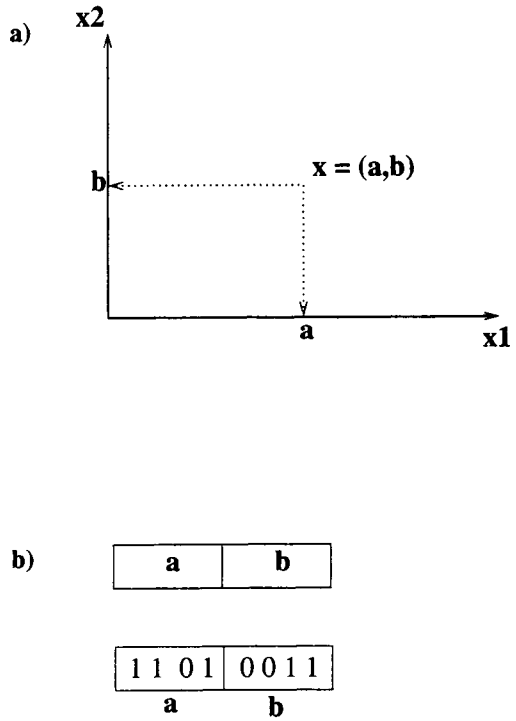


Fig. 27.1. (a) Graphical representation of a candidate solution. $x1$ represents the pH and $x2$ the concentration. $x = (a,b)$ represents a candidate solution. (b) String representation of the candidate solution.

There are many possibilities to represent the candidate solutions. The parameter values may be scaled and can be encoded as integer values or real values. Qualitative parameters can be encoded as integer values with a restricted number of valid values. An often used representation is the binary encoding of the parameter values as in the traditional GA. In this encoding method the characters in the bit string are zero and one. In general all bit strings (candidate solutions) have the same length. Each bit string is divided in sub-parts. Each sub-part represents a solution parameter and has a minimum length of one bit (see Fig. 27.2).

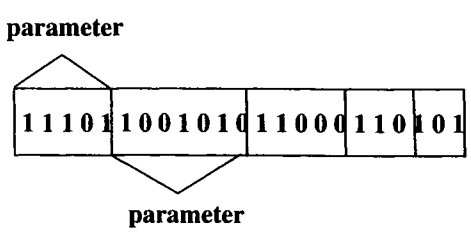


Fig. 27.2. Example of a binary representation of a candidate solution consisting of five parameters.

Suppose a parameter has the integer value $V = 5$. With a three-bit string representation $2^3 = 8$ different (0–7) values can be represented. The binary representation of the integer 5 is then 101 ($V = 1\times2^0 + 0\times2^1 + 1\times2^2 = 5$). When the parameter has the real value R, that must be encoded in a binary number (BN) of B bits, first a certain range $[l,u]$ is defined and divided into 2^B equidistant levels. Then the appropriate integer value, V, must be calculated. It is the integer number, closest to the value V':

$$V' = \frac{R - l}{u - l}(2^B - 1) \tag{27.1}$$

This integer value can then be represented in a binary number of B bits. The factor $2^B - 1$ determines the total number of different values that can represented between l and u. With a larger B, more values can be represented and thus a higher resolution can be obtained. The appropriate B for a specific problem can be derived from the required resolution. The binary parameter can be translated into its real value by:

$$R = l + (u - l)\frac{1}{2^B - 1}V \tag{27.2}$$

V is the integer value corresponding to the binary string.

Example:

A parameter that can have a real value between $l = 0.0$ and $u = 10.0$ must be represented in a binary form by a four-bit string. For the value $R = 3.3$, e.g., the binary representation is obtained as follows:

$$V' = \frac{3.3 - 0.0}{10.0 - 0.0}15 = 4.95$$

$$V = 5$$

$$\text{BN} = 0101$$

The binary string 1011 ($V = 11$) then represents the value: $R = 0.0 + (10.0 - 0.0)\,{}^{11}/_{15} = 7.3$. The string 1100 ($V = 12$) represents the real value 8.0. A four-bit representation corresponds thus with a resolution of 0.7. To obtain for this example a resolution of 0.1 it must be possible to represent 101 different values. To ensure that the factor $2^B - 1$ is larger than 101, the number of bits, B, must be at least 7.

In our chemical optimization example the string consists of two sub-parts, one representing the pH and the other the concentration. In Fig. 27.3 some encoding types are given for a string, representing a candidate solution: pH = 7.4 and concentration = 10.0; the range for the pH value is [0.0–10.0] and for the concentration it is [0.0–30.0].

The encoding with real numbers looks, from the human point of view, the most familiar since we think of the solution in real numbers. This makes it easier to represent the problem in a sensible way. Most GA experts, however, do not advise

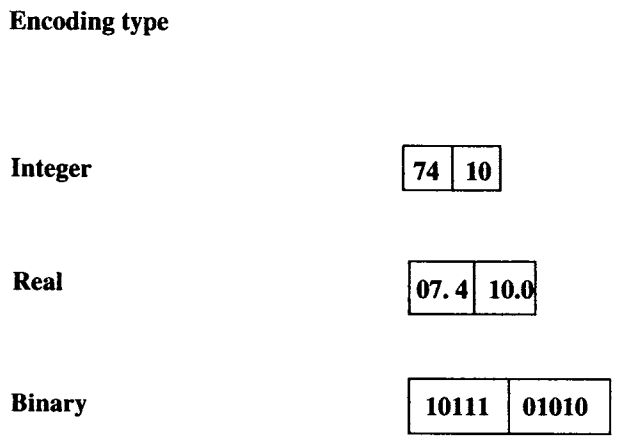


Fig. 27.3. Some encoding types for the candidate solution (pH = 7.4; concentration = 10 M).

the real encoding technique (see Section 27.3.2.4) [7]. Another frequently used technique is Gray coding. A Gray-code represents integers $(0,1,...,2^{N-1})$ as a binary string of length N in such a way that the Gray-code representation of adjacent integers differs in only one bit position. Walking through the integer sequence therefore requires flipping just one bit at a time.

Example: the binary coding of the integers (0,1,...,7) is (000 001 010 011 100 101 110 111) for $N = 3$. A possible Gray coding is (000 001 011 010 110 111 101 100). There exist multiple Gray codings of any N.

A more realistic example is to find the 3D configuration of a (bio)-macromolecule that is compatible with an experimentally obtained NMR spectrum. Since different configurations are caused by different torsional rotations around bonds it is straightforward to represent the 3D configuration by means of the torsion angles in the molecule. Consequently, the candidate solutions consist of a number of values, representing the different torsion angles. For example, the string:

150.0 060.0 300.0 000.0 090.0

represents a candidate solution involving 5 torsion angles, each of these, represented as a real value. When it is sensible to represent the angles as a multiple of, e.g., 30° it is possible to represent the torsion angles as a multiple of 30. The candidate solution is then represented as: "5 2 10 0 3". These integers can also be encoded as a string of binary numbers: "0101 0010 1010 0000 0011". The choice of representation determines the mesh size at which the problem space is searched. In the first decimal representation it is 0.1° while for the other representations it is 30°.

Before the candidate solutions can be encoded it must first be decided how the (chemical) problem can be represented. In some cases this is straightforward, such as in the first example. In the second example, however, there are different ways

of representing a chemical 3D structure. An alternative to the use of torsion angles is the use of 3D coordinates.

A third example concerns curve fitting. As outlined in Chapter 11, the only way to fit complex functions (that are non-linear in the parameters) to the data is to use an iterative strategy. However, this strategy requires much domain knowledge about the parameter values. If the initial values of the parameters are not close enough to the correct ones, the method is likely to end up in a local optimum resulting in a bad fit. GAs allow to use much less knowledge about the parameter values. A candidate solution consists of a string containing all parameters of the function, which has to be estimated. Again the representation defines the mesh size of the grid on the search space, evaluated by the GA.

A typical example of sequence optimization is the travelling salesperson problem. Given 5 cities that have to be visited by the salesperson a candidate solution can be represented as a string of integers, each representing a particular city. The sequence in the string determines the visiting sequence. The string: "1 3 2 4 5" means that the cities are visited in the order 1; 3; 2; 4; and 5. A chemical example concerns the industrial separation of a multicomponent flow into a number of multicomponent products of specified composition by means of a sequence of distillation columns. The problem is then to synthesize an optimal distillation sequence that separates the single multicomponent feed into several multicomponent products at a minimal cost [8].

In variable or feature selection the problem consists of selecting a subset of k variables from a set of K variables (see e.g. Chapter 10). To represent this problem, all variables in the source set of K variables must be uniquely labelled and represented as a string. For instance the string "1 2 3 4 5 6 7 8 9 10" represents a source-set with $K = 10$ variables. A candidate solution can be represented as a bit string of length K. A zero or one in a specific position determines the absence or presence of the specific variable:

"0 1 0 0 1 1 0 1 1 0"

represents a subset consisting of $k = 5$ variables (2,5,6,8,9). We can distinguish a fixed-size subset selection and a variable-size subset selection. For the fixed-size selection k is considered constant for all candidate solutions. The sum of the bits in the string must thus equal k for all candidate solutions. There is no such constraint for variable sized subset selection. In this case the number k must also be optimized.

27.3.2 Flow chart of genetic algorithms

27.3.2.1 Initiation

The first step in the genetic algorithm (see Fig. 27.4 for a flowchart) is the creation of a number of candidate solutions. Without prior (heuristic) knowledge

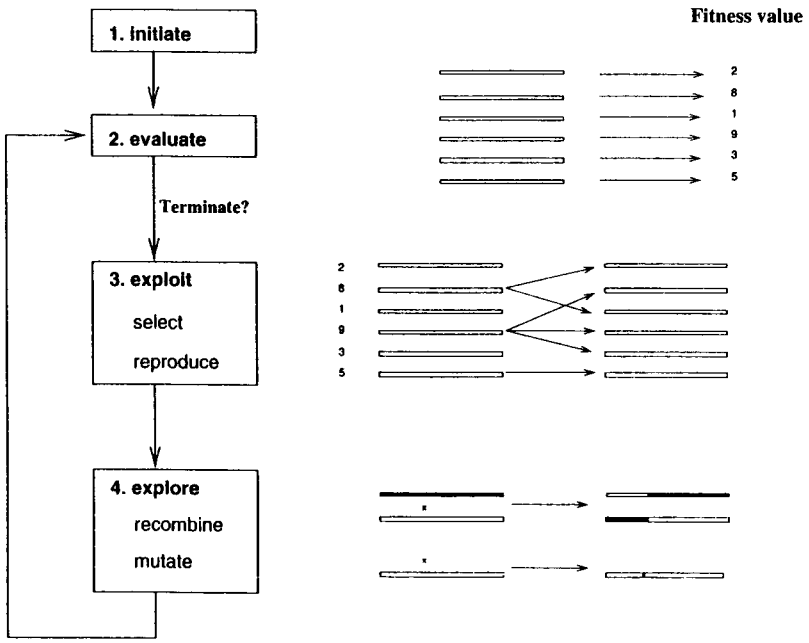


Fig. 27.4. Flow chart of a basic GA.

the candidate solutions are randomly chosen. In this way the whole search space is equally well spanned. To avoid that the search space becomes unnecessarily large, it can be useful to use prior knowledge to select the initial candidate solutions in a more deterministic way [9,10].

The number of candidate solutions, N_p, is usually chosen in the range 50 to 100 for most practical situations. This set of candidate solutions (or *population*) constitutes the first *generation* [11]. From this first generation subsequent generations are created. The idea is that eventually, in the final generation, a candidate solution emerges that is the optimal solution of the problem.

27.3.2.2 Evaluation and termination

The next step in the algorithm consists of evaluating the quality of the candidate solutions. This requires an evaluation criterion or, in genetic algorithms terminology, a *fitness value*. The optimal or target value of the evaluation criterion must also be defined (note that the value for the real optimum is often unknown). Since judging the quality of a solution requires domain knowledge, this step is the most domain dependent of the whole algorithm.

For the NMR example it is possible to calculate the NMR spectrum belonging to each candidate solution. The smaller the difference between the calculated and the experimental spectrum, the closer the configuration is to the true one. The mean

squared difference between the two spectra can be used as an evaluation criterion. The target value for the fitness value is in this case zero or a small value depending on the experimental error.

In general, the fitness value is obtained by means of an *objective* or *evaluation function*: $y = f(\boldsymbol{x})$, where $\boldsymbol{x}$ is the string representation of the candidate solution and y is the fitness value of that particular candidate solution. The evaluation function thus has a string as input and returns a value indicating the performance of the candidate solution with respect to the problem considered. The fitness values can be used as such or may be scaled (e.g. between 0 and 100). The calculation of the fitness value is usually the most computationally intensive step.

For the curve fitting example the candidate solutions are evaluated by the Root Mean Square Error (RMSE, see Chapter 10) of the resulting fit. The fitness value can thus be defined as the RMSE between the experimental data and the values, predicted with the non-linear function, using the parameter values belonging to the candidate solution.

In the evaluation step it must also be verified whether the strings meet all constraints of the solution space. When illegal strings emerge, several strategies can be followed:

- Rejection: the strings that do not fulfil the constraints are rejected.
- Penalization: the strings that do not meet all constraints are penalized by lowering the fitness value.
- Repairing: the strings are repaired according to a certain strategy so that they become legal. When e.g. a candidate solution emerges that is situated outside the boundaries of the allowed domain, the parameter values can be changed so that it falls inside the domain.

Constraint handling in GAs is still a major point of research and a full description is outside the scope of this chapter [2,3,12].

The next step is to check whether the algorithm can be terminated. This is the case when there is a candidate solution fulfilling the target criterion for the fitness value. At the first generation it is unlikely that such a solution is among the candidate solutions since these are randomly selected. Usually, a convergence termination criterion is also included. When the candidate solutions of the new generations are not significantly better, the execution of the algorithm can be stopped. At this point other search heuristics (e.g. the Simplex method) can be included to improve the population of solutions. Another simple termination criterion is a maximum number of generations. It prevents the from algorithm going on indefinitely when no convergence can be achieved.

27.3.2.3 Selection phase

The next step is inspired by Darwin's selection rule (survival of the fittest). It is also called the *exploitation stage* of the algorithm. A new set of candidate solutions

is created from the current population. These new solutions constitute a new, temporary population. In the simplest case this temporary population (sometimes called the *breeding population*) has the same size (N') as the original population (N_p). This new population is created as follows:

- Select a candidate solution from the current population, according to a predefined selection strategy (see further). Make a copy of this string and put it aside as first member of the new population.
- Repeat this N' times until the new population is complete. Note that selected candidate solutions are not excluded in the subsequent selections and can thus be selected several times.

When the selection strategy favours the solutions with a higher fitness value, the new population will contain more of the better solutions of the original population. This is also called *selection pressure*. However, no new solutions are created yet at this stage. The new population is expected to show, on average, a higher quality than the previous one. A common selection strategy is to make the sampling chance proportional to the fitness value. There are several variants of this principle [3,5,13]:

- *Roulette selection*: This is the most common, although not the best strategy. The selection chance is proportional to the (scaled) fitness value. Scaling is used to prevent that a few solutions showing a fitness value which is much larger (10–100 times larger) than the average dominate the selection procedure and cause premature convergence. Scaling also prevents that all candidate solutions obtain about the same fitness value. The most usual scaling in this procedure is the linear scaling: scaled fitness value = $a \times$ fitness value + b. The scaling factors a and b are selected according to two criteria: (i) all fitness values must be positive and (ii) the best candidate solution should have a scaled fitness value that is about three to four times the average fitness value.

The procedure is called roulette wheel selection since can be imagined as a special kind of roulette wheel (see Fig. 27.5). This procedure is used in most applications and a convenient way to implement it is as follows:

1. Put the strings in a random sequence.
2. Calculate the cumulative fitness value for each string in the selected sequence.
3. Generate a random number between 0 and this total fitness.
4. Select the first string whose cumulative fitness value is greater than or equal to the random number.

An example of roulette wheel selection of 5 strings is shown in Table 27.1. This procedure ensures that the probability of each string to be selected is proportional to its (scaled) fitness value.

- *Linear selection*: the selection probability is based on the fitness rank instead of the fitness value.
- *Threshold selection*: This procedure ensures that the worst solutions do not enter the new population. First the N_w worst solutions are removed from the

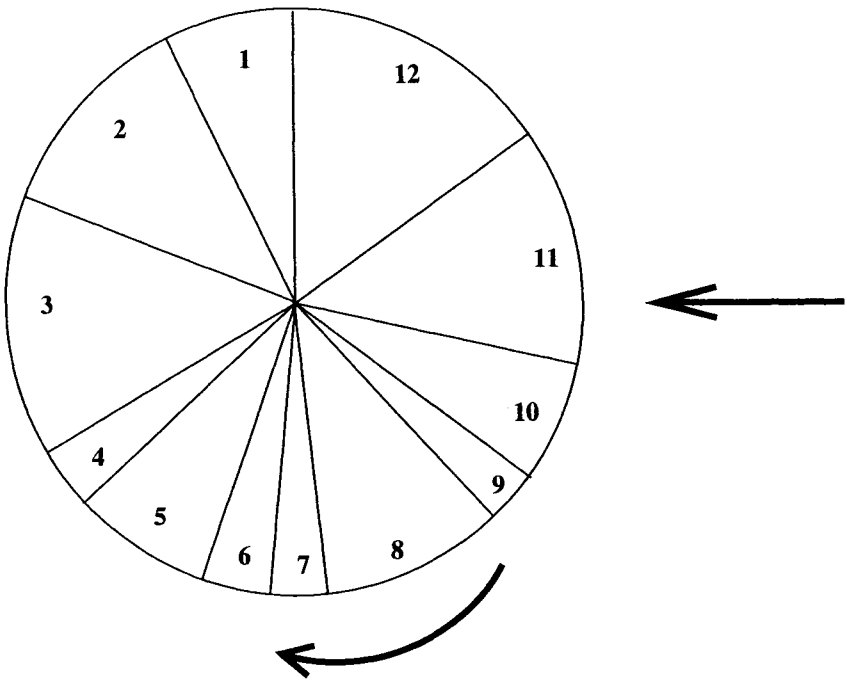


Fig. 27.5. Roulette wheel selection: all candidate solutions (the numbers on the wheel) are assigned a pie piece on the wheel, proportional to their scaled fitness values.

TABLE 27.1

Example of selection by means of the roulette wheel strategy

String	1	2	3	4	5	6
Fitness	3	5	15	34	2	20
Cumulative fitness value	3	8	23	57	59	79
	Random numbers:		Selected strings			
	32		4			
	6		2			
	18		3			
	44		4			
	77		6			

original population. From the remaining number ($N_p - N_w$) of candidate solutions strings are randomly selected. In this way the N_w worst solutions are never transferred to the next generation. N_w is a predefined integer with a value between 0 and ($N_p - 2$). $N_w = 0$ results in random selection.

- *Tournament selection*: From the initial population N_l candidate solutions are randomly selected (with replacement). From this set the best string is selected and transferred to the new population. N_t is an integer value between 2 and $N_p - 1$. Binary tournament selection ($N_t = 2$) is the most usual. $N_t = N_p$ is excluded, since then the next population would consist of N' identical strings (the best one) and would not have any diversity.
- *Elitist selection*: This is a procedure with a deterministic part to guarantee that the best candidate solution(s) are transferred to the next population. In a first step the N_e best solutions are unconditionally selected. The other members of the population are selected in a probabilistic way, e.g. according to one of the previously discussed strategies. Combined with tournament or threshold selection this procedure performs best in practice.

A non-random selection procedure results in a population which is on average, better than the previous one. In this way the available information is exploited. It is important to realize that bad solutions have a lower chance to be transferred to the new population, but are not excluded and may be selected and participate with the other candidate solutions in the procedure. This prevents premature convergence.

27.3.2.4 Recombination and mutation

The selection pressure, described in the previous step is essential but not sufficient to search the problem space efficiently. All that has been accomplished is that in future populations the best solutions of the original population will predominate but no new solutions have been found yet. To achieve this, variations must be brought into the population. At the same time important information, already present in the population, must be preserved. This can be achieved by means of so-called genetic operators: *recombination* or *cross-over* and *mutation*.

Cross-over (recombination)

The inspiration for this operator comes from the biological crossover between chromosomes. Fractions of two candidate solutions (parent strings) are recombined so that two new solutions (child strings) are obtained. The simplest type of cross-over is the one-point cross-over as illustrated in Fig. 27.6. Two strings are split into two parts at a randomly selected point. The different parts are then recombined. The idea is that better solutions may be obtained by this type of recombination. Important information already present is usually preserved and passed on to the child strings. This is illustrated in Fig. 27.7. Suppose that the overall solution (consisting of two parameters a and b) of the problem is known and is represented as string S in Fig. 27.7. The substrings a and b represent the two parameters of the solution. The two parent strings are two candidate solutions to be recombined. The bits in bold represent 'correct' bits. The cross-over operation allows that pieces of well performing bit strings, called *building blocks*, are exchanged by the parent strings to reproduce better children.

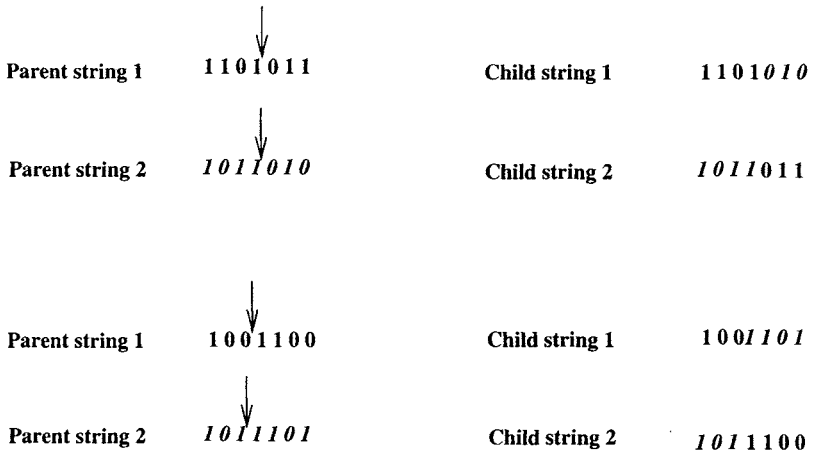


Fig. 27.6. Example of the one-point cross-over operator.

The recombination procedure is carried out as follows: the strings of the temporary population are put in pairs. In the most common case the pairing of the strings is done in a random way. This may however result in unproductive recombinations. It is indeed possible that due to the selection procedure of the previous step, some duplicate strings may emerge in the new population. Pairing and recombination of these strings results of course in exactly the same strings. To avoid this, pairing can be done using more prior knowledge. It is reasonable to assume that diversity is most enhanced when very dissimilar strings are paired. A pairing procedure that favours the combination of dissimilar strings is therefore sometimes used. The difference between the fitness values of the strings can be used as a measure of dissimilarity. It can indeed be assumed that strings with different fitness value are dissimilar. Strings with equal fitness value however are not necessarily the same, since distinct solutions of the same quality may exist. An alternative measure of dissimilarity is the Hamming distance (in the encoded binary space) or Euclidean distance (in the real problem space) between the strings (see also Chapter 30).

A cross-over point is selected randomly for each pair of strings.

At this point the parent strings can be broken and recombined to form the child-strings. This recombination of the pairs is performed with a probability P_r, which is commonly chosen between 0.5 and 1.0.

Obviously, the recombination operator introduces new candidate solutions in the GA. Many GA practitioners claim that the cross-over operator is what distinguishes the GA from all other optimization techniques [7].

In Fig. 27.7a a situation is shown in which the correct combination can never be obtained in one step with a simple one-point cross-over, even when all correct

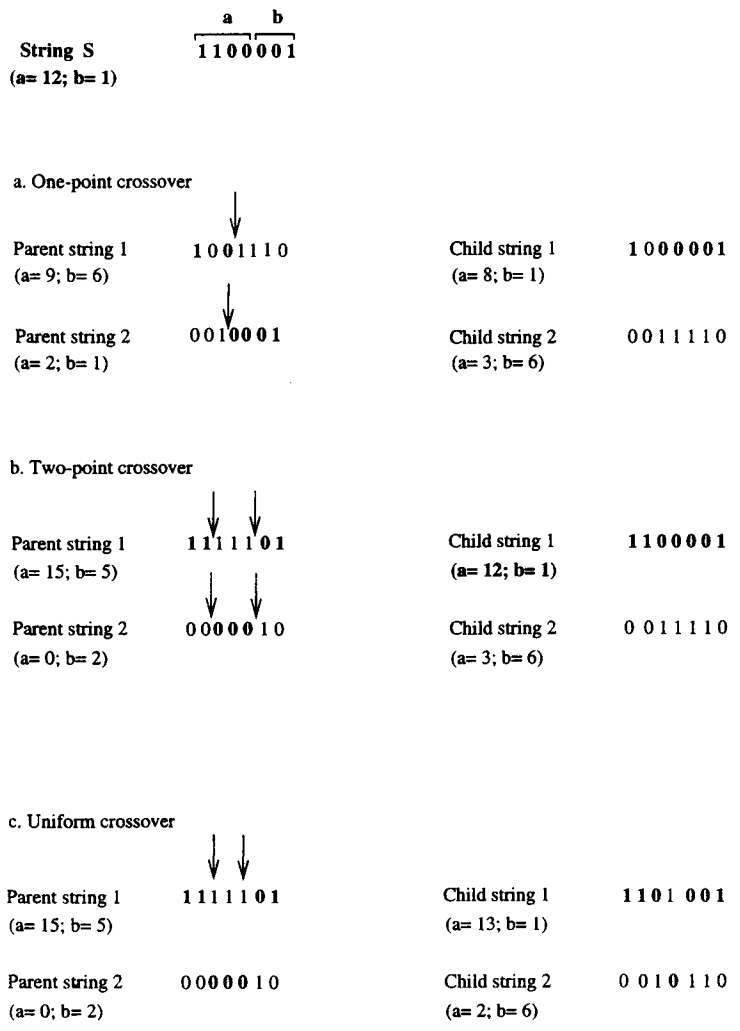


Fig. 27.7. Example of two-point cross-over. String S represents the global solution. The bits in bold have the correct value.

information is present in the parent strings. With a two-point cross-over, two cut-points are randomly selected and the bitstring between the two cut-points is interchanged between the parent strings. This cross-over allows to combine the parent strings in Fig. 27.7b to form a correct child string. A GA, using a one-point cross-over requires two or more generations to obtain that result. This type of cross-over operator is in general called the *n*-point cross-over operator. The cross-over operator that performs best in practice is the uniform cross-over type. This operator interchanges a number of randomly selected bits between the parent strings (see Fig. 27.7c).

Note that the cut-point(s) can be situated between any two bits of the string and is not restricted to the places between the subparts of the string representing the different parameters. This feature makes it possible that the new candidate solutions obtain values for the parameters that were not present in the initial population. This is also illustrated in Fig. 27.7; the integer values, represented by the parent and child strings are given. It can be seen that, e.g. for the parameter a, the values for the child strings (12 and 3) are different from those of the parent strings (15 and 0). It also shows that although the correct values of the parameters are not present in the parent strings, it is possible that the optimal value emerges in the child strings. This allows a powerful search behaviour [3,14,15].

When the real number encoding technique is used, the cut-point can, in contrast to the binary representation, only be set between the subparts of the string, representing the parameters. This results in a much more restricted search behaviour of the GA. The new solutions are only different combinations of the parent strings. Therefore variants of the classical cross-over operator are developed in which some numerical features are incorporated. One example is the average cross-over where the child strings are formed by averaging different subparts of the parent string [16–18].

It should be noted that whatever type of cross-over is used there always is a danger that invalid solutions are created. This can be handled in the evaluation phase, by checking that the candidate solution is within all constraints. As a rule, however, it can be stated that when many illegal solutions are found it means that the representation and the cross-over operator do not combine well. Usually an alternative representation or cross-over type is tried.

In some cases problem-specific cross-over operators must be designed. A typical example is when a GA is used to solve a problem in which a sequence must be optimized, such as the industrial separation problem. Suppose that in a chemical plant there are six distillation columns with different separation properties available. In Fig. 27.8 a typical cross-over is shown. When cross-over operators, as described earlier, are used it is clear that more invalid (i.e. one collum appears twice in a sequence) than valid child strings will emerge. New cross-over operators have been developed especially to handle this kind of sequence problems [7].

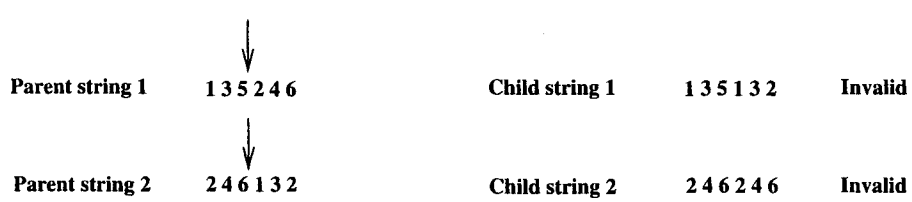


Fig. 27.8. Example of invalid strings that are produced by the one-point cross-over operator.

Mutation

In one particular situation cross-over is not able to introduce the necessary variation. When one bit is the same in all the strings, this bit cannot be varied by means of the cross-over operator. Consider e.g. the following population of strings:

0 1 1 0 0 1 0 1

1 0 0 1 1 1 0 1

0 1 1 0 1 1 1 1

1 1 1 0 1 0 0 1

0 0 1 1 1 1 1 1

The last bit in all these strings equals 1. This means that only odd numbers are present in the population and it is not possible to obtain an even number by means of the cross-over operation alone. By this mechanism the GA may be trapped in a local optimum of the search space. To overcome this problem, another type of operator was introduced: the mutation operator. Mutation is a random local change in the bit string. Any position in any string is subject to mutation with a probability P_m, which is usually chosen below 0.05. With a binary representation mutation causes those bits that are selected to flip from zero to one or *vice versa*. The effect of the operator is shown in Fig. 27.9. Mutation ensures additional diversity in the population and helps to avoid premature convergence. The change of one bit can indeed move the string to a totally different place in the solution space. The mutation operator as described here cannot be applied to strings that are encoded as real numbers. The analogue for real number encoded strings can be the replacement of a parameter value by a random value. When real number encoding is used mutation can however be made more intelligent. It can be used to force already well-performing strings, that are situated on a hill in the solution space, to explore the hill. This can be accomplished by changing with a small value the current parameter values. Using binary coding, this effect can only be achieved by first decoding the string to a real number and reconverting it back afterwards.

Old string	**Mutate ?**	**New string**
1001100	**nnnnnyn**	**1001110**
1011101	**nnnynnn**	**1010101**

Fig. 27.9. The mutation operator.

Mutation was initially assigned a background role in the GA since high mutation rate would damage useful information already present in the population. Therefore a genetic algorithm that uses a low mutation probability performs better than one that uses a high mutation probability. A high mutation probability with no cross-over at all is similar to a random search. A mutation rate of zero may cause the GA to be trapped in a local optimum as explained before. Recently however the role of mutation has been revised [19,20]. In some types of problem the influence of the mutation operator on the performance of the GA seems to be higher than that of the cross-over operator. A high mutation rate and a low cross-over rate would be preferable in such cases. Future research will have to bring clarity in this situation.

27.3.2.5 Replace and continue

After the application of cross-over and mutation the current population can be replaced by the temporary new generation population. When the original and the temporary population are equally large ($N' = N_p$), the replacement procedure is straightforward. When the temporary population is chosen to be smaller than the original population (i.e. $N' < N_p$) the replacement procedure requires some attention. From the N_p strings of the original population N' are replaced by the strings of the temporary population. This strategy is called a steady state algorithm with a generation gap, $G = N'/N_p$. The selection of the N' can be biased towards the worst strings of the original population. The same selection principles as applied in the selection procedure (step 2) can be applied. This step completes one cycle or one *generation*. The procedure is repeated many times. Depending on the nature of the problem (number of parameters, degree of complexity) some hundreds, even up to many thousands of generations are necessary before a sufficient degree of convergence is achieved.

27.3.2.6 Performance measure of a generation

The performance of a generation can be measured in different ways. There is, however, no unique best way to measure the performance of a particular GA. We can compute the mean of the fitness values of all members of the population, the mean of the K best members of a generation, where K is a predefined integer number, or the best solution of each generation or a combination of these possibilities etcetera. Usually the mean or median of the population together with the 'best ever string' is plotted as function of the generation number. A typical performance plot of a GA-run is given in Fig. 27.10.

27.4 Configuration of genetic algorithms

In the previous section many configurational aspects of GAs have been discussed. The most important ones are summarized in Table 27.2. All these possible

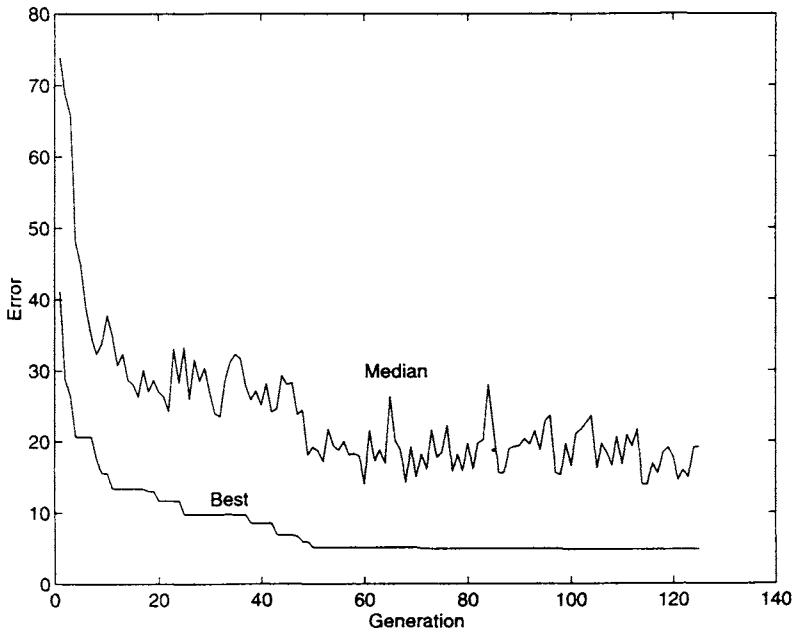


Fig. 27.10. A typical performance plot of a GA. The best ever solution together with the median of each generation are plotted.

TABLE 27.2

The most important configuration parameters of a GA

Choice of representation of the solutions;	
Choice of strategy for:	initialization of the algorithm; selection of strings (and selection of the appropriate parameters); recombination of strings/mutation; replacement of strings.
N_p, the size of the initial population;	
N', the size of the temporary population;	
P_r, the recombination probability;	
P_m, the mutation probability.	

configuration settings make the GA very versatile. The experienced user has many possibilities to tailor the GA to the requirements and characteristics of a specific application. There is however up to now no standard methodology available to find

optimal settings for a specific problem. This implies that experience and some rules of thumb will have to guide the user. It should be stressed that a reasonable understanding of the search mechanisms and of the influence of the different parameters is a prerequisite.

The configuration optimization starts with a good initial guess for the configuration parameter settings from where further improvements can be made. Different strategies can be used to optimize the configuration. Fortunately, however, it turns out that GAs are reasonably robust to a suboptimal configuration. The GA will usually still find the global optimum in the search space, but with non-optimal settings it requires much more time before convergence is reached. A much more difficult situation occurs however when the GA is not converging. It is then very hard to find out which settings have to be changed in order to obtain convergence.

The experienced user of GA can apply an interactive optimization strategy. Given a good starting point for the configuration settings, this initial estimate is refined by stepwise adjustments of the settings. It is of course possible to apply a more systematic approach using, e.g., factorial designs. A drawback of this approach is that it requires a large amount of computing time so that in practice such an approach is hardly feasible. Schaffer et al. [21] applied full factorial designs to find the optimal parameter setting for a set of numerical function optimization problems. They came to the conclusion that the optimal settings vary from problem to problem. Moreover, these experiments were conducted using only one type of encoding (Gray encoding).

The use of GAs to optimize the configuration of GAs has been reported in the literature [7]. To remain practical, the number of control parameters to be optimized should remain limited.

Previously we implicitly assumed that the optimal settings to start a genetic algorithm run remain optimal during the whole run of the GA. However, it is more reasonable to assume that the optimal parameters of the GA should evolve during the run. This then requires dynamic parameter settings, but the approach is more complicated.

27.5 Search behaviour of genetic algorithms

27.5.1 Search accuracy and precision

It is important to realize that the performance of a GA is based on a population of solutions rather than on one specific solution. In that context a *search accuracy* can be defined. By accuracy we mean that the correct optimum in the solution space is approached by a sufficient number of strings in the population.

In general it is observed that GAs show a very good search accuracy in the sense that the correct hill in the fitness landscape is often reached. Due to the probabilistic nature of the operators in the GA it is much more difficult to reach the actual top of the hill. We say that GAs show a poor *search precision.*

Note that these properties are opposite to the features of a deterministic optimization strategy such as a Simplex optimization or other hill climbing techniques. These techniques show a poor accuracy, because they always reach the hill in the fitness landscape that is closest to their starting point. Whether this is the highest hill depends entirely on the initialization point. Consequently, these methods are not suitable to search rough response surfaces. However, once they have found the correct hill it is almost certain that they will reach the top of that hill in a short time: they show a good search precision.

27.5.2 Behaviour of GA in the presence of multiple optima

The fact that there are many strings in a population competing with each other allows the GA to find more than one high hill in the fitness landscape. When several optima of similar quality are present the strings are divided over these optima during the evolution of the GA. This is an interesting feature of the GA since it allows different solutions of similar quality and this in turn can throw new light on the problem under investigation. An example is the search for the 3D structure of (bio)-macromolecules. Here it is of utmost importance to find all structures that are compatible with the experimentally obtained NMR spectrum of a molecule [e.g. 22].

When the different optima in the fitness landscape are not equally good the GA mechanisms will eventually make the strings converge to the best optimum. When one is interested in finding the different optima some additional parameters may need to be introduced in the GA. One possibility is to incorporate a sharing penalty in the fitness value: when too many strings are situated around the same optimum the fitness value of these strings is diminished with a certain value. This will favour the strings to spread on different optima if they are present.

27.6 Hybridization of GAs

Because of the specific properties of GAs it is profitable to combine them with other techniques. In this section some of these techniques are discussed (see e.g. Refs. [7,13,23]).

Post hybridization. One obvious combination is the use of a hill climbing technique (see Chapters 11 and 26) after a GA-run. Since GAs are powerful to find

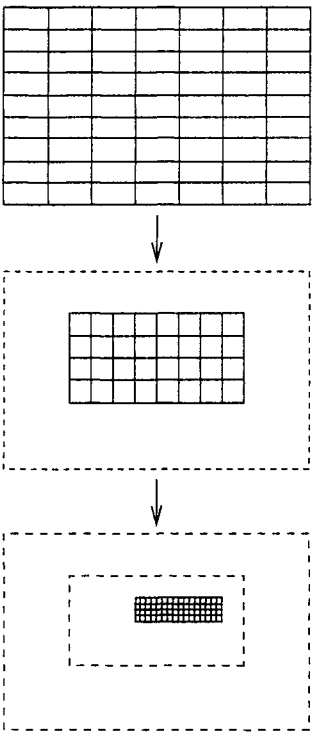


Fig. 27.11. Self-hybridization of GAs. The mesh size is gradually decreased in subsequent GA runs.

the highest hill but are bad at finding its top one can increase the precision by applying a hill climbing search strategy with the best strings. The last generations of a GA are not really efficient for finding the final solutions, since the strings are merely circling around the hill and will only find its top by accident. This phenomenon is also called *genetic drift*. Another possibility is to combine cluster analysis of the strings during the GA run which allows the identification of multiple optima in the fitness landscape.

Self-hybridization. It is possible to run several GAs in sequence with a different configuration. This results in a chain of GAs in which each passes results to the next GA. An example is the iterative search volume contraction. The first GA is meant to search the space with a large mesh size. The mesh size can be controlled by the representation and encoding. After a number of generations the GA is stopped and a new GA is started with the current population that searches a subspace of the solution space with a smaller mesh size. This procedure is referred to as self hybridization and makes it possible to search large spaces in a more efficient way. The procedure is schematically shown in Fig. 27.11.

Pre-hybridization. Other techniques can be applied to incorporate domain knowledge in the initialization step. De Weijer et al. [9] use a neural network to estimate the number of peaks in a curve fitting problem. This approach usually results in a much more efficient optimization procedure than blind (random) initialization.

Parallel-hybridization. It is possible to run several GAs in parallel, e.g. each with a different configuration or population initialization. After a number of generations information is exchanged between these different GAs. The way to do this is still subject to research. A possibility is to define a migration operator. This operator exchanges some of the strings between the different populations according to a certain strategy. This is quite a complex procedure and many variants exists. The description of these procedures is outside the scope of this chapter.

27.7 Example

In this section we follow step by step a run of a GA to solve a (relatively) simple problem. We will try to find the optimum of a function of two parameters ($p1$, $p2$).

$$f(p1,p2) = e^{-7|(p2-0.5)^2+(p1+1)^2|} + 0.5\, e^{-0.4(p2+1)^2 - 0.2(p1-0.5)^2}$$

In Fig. 27.12 the function is given together with the contour plot of the function values in the parameter space. The problem is to find the parameter combination ($p1$,$p2$) yielding the highest peak. We will apply a basic GA to solve this problem. The configuration parameters are selected as follows:

Population size, $N_p = 20$;

Cross-over operator: one-point cross-over;

Cross-over probability, $P_r = 0.75$;

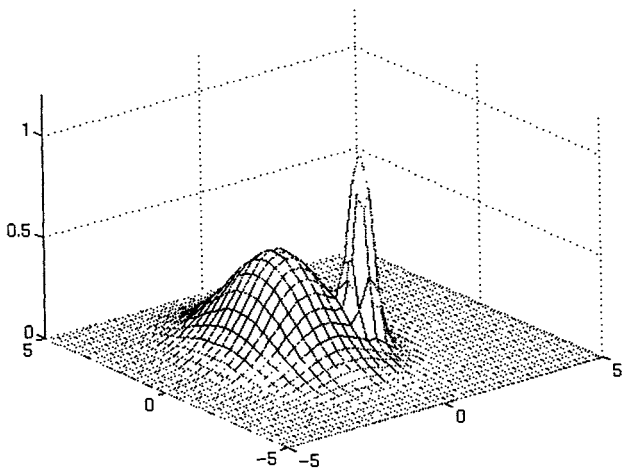
Mutation probability, $P_m = 0.05$;

Stopping criterion: maximum 30 generations.

The first step is to select a representation of the candidate solutions. We selected a binary coding of the parameters and defined the range of the parameters to be between –3.33 and +3.33. Each parameter is coded by 10 bits; this allows us to represent 2^{10} levels between –3.33 and +3.33. A candidate solution thus consists of the two 10-bit binary encoded values of the parameters $p1$ and $p2$ (see Fig. 27.13). According to eq. (27.2) the string 0100101111, e.g., represents a real number value of:

$$R = -3.3 + (6.66)\frac{1}{2^{10}-1}303 = -1.33$$

(a)



(b)

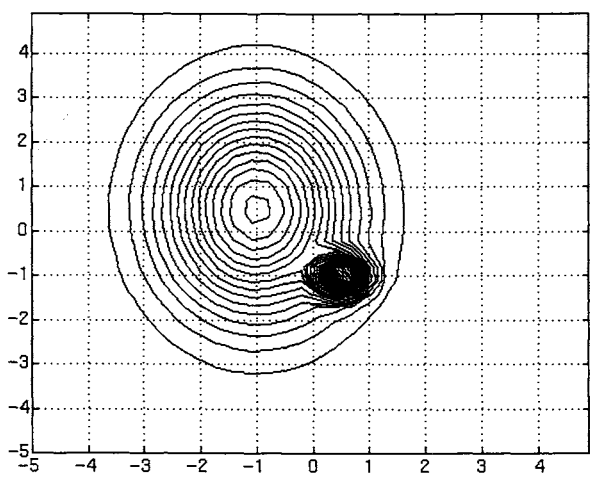


Fig. 27.12. Graphical representation (a) and contour plot (b) of the function:
$f(p1,p2) = e^{-7[(p2-0.5)^2 + (p1+1)^2]} + 0.5\ e^{-0.4(p2+1)^2 - 0.2(p1-0.5)^2}$.

P1	P2
0 1 0 0 1 0 1 1 1 1	1 0 0 0 1 1 1 1 0 0

P1 = -1.33 **P2= 0.42**

Fig. 27.13. Binary representation of the candidate solution ($p1 = -1.33$; $p2 = 0.42$).

TABLE 27.3

Initial population for the example of Section 27.7

	String	$p2$	$p1$	Fitness value
1	10000000100100001000	0.016	−1.6	0.4100
2	10100000110100011010	0.85	−1.5	0.4418
3	11000100000011010010	1.8	−2.0	0.2492
4	00110010011110000000	−2.0	2.5	0.0010
5	00001011001001111111	−3.0	0.83	0.0106
6	00100111010011000100	−2.3	−2.1	0.0658
7	00010100000000111001	−2.8	−3.0	0.0119
8	11100011110001111100	2.6	−2.5	0.0817
9	01001010011000011110	−1.4	0.20	0.3078
10	10100100001011001011	0.94	1.3	0.0557
11	10101001110000001010	1.1	−3.3	0.0596
12	01010000001001110100	−1.3	0.76	0.4883
13	01011111011011001101	−0.85	1.3	0.0457
14	00111001001010000100	−1.8	0.86	0.0442
15	01111010000100100111	−0.61	−1.4	0.4285
16	10100111101111100111	1.0	3.2	0.0004
17	00000000111010010110	−3.3	0.98	0.0057
18	10001110101011111111	0.38	1.7	0.0294
19	10101010001011110101	1.1	1.6	0.0315
20	01100100010110100111	−0.72	−0.58	0.3456

Initial population summary:
Maximum fitness, 0.4883
Average fitness 0.1557.

TABLE 27.4

Recombination and mutation step for the example of Section 27.7

String	Cut-point*	Cross-over	Mutation
01100100010110100111	5	10100000110100000111	10101000110100000111
10100000110100011010		01100100010110111010	00100100010111111010
00111001001010000100	10	10000000101010000100	10000000101010100100
10000000100100001000		00111001000100001000	00111001000100001000
00111001001010000100	13	10101011001010000100	10101011001010001110
10101010001011110101		00111000001011110101	00111000001011110101
10101010001011110101	2	10100000110100011001	10100010110110011001
10100000110100011010		10101010001011110110	10111010001011110110

(continued opposite)

TABLE 27.4 (*continuation*)

String	Cut-point*	Cross-over	Mutation
10100000110100011010	4	01100100010110101010	01100100010110101010
01100100010110100111		10100000110100010111	10100000110100010111
10000000100100001000	13	01001010100100001000	01001011100100000000
01001010011000011110		10000000011000011110	10000000011000010110
11000100000011010010	13	10000000000011010010	10000000000011010010
10000000100100001000		11000100100100001000	11000100100101011000
01100100010110100111	14	01111000010110100111	01111000010110101111
01111010000100100111		01100110000100100111	01100110000100100111
01100100010110100111	—	01100100010110100111	01000100010110110111
10001110101011111111		10001110101011111111	10001110111011111111
01001010011000011110	—	01001010011000011110	01000010011000011110
01100100010110100111		01100100010110100111	01100100010010100111

*Starting from the right.

Since the problem is relatively simple a small population is selected ($N_p = 20$) and kept constant. The initial generation is selected randomly and is given in Table 27.3. For each of these candidate solutions the fitness value (here the function value) is calculated. From this population a temporary population is selected using the roulette wheel selection strategy on the fitness values, as explained in Section 27.3.2. The selected strings are:

1,3,20,9,18,2,14,20,14,9,1,20,19,2,2,20,15,1,20,19.

The next step is to apply recombination and mutation. The cross-over procedure is started by pairing randomly the selected strings. For the one-point cross-over one cutpoint must then be randomly selected for each pair. The result is shown in Table 27.4. In the first column the pairs and the selected cutpoints are shown. On each pair the cross-over is applied with a probability of 0.75. The result is shown in the second column of Table 27.4. Finally the mutation operator acts on each bit with a probability of 0.05. The result is shown in the third column. This temporary population replaces the initial population and the fitness of each candidate solution of this new population is calculated (see Table 27.5). This completes one cycle. As can be seen the results of the new generation are not better than the initial population. It is on average even a little worse. This is due to the stochastic properties of the algorithm. In Fig. 27.14 the position of the candidate solutions of

TABLE 27.5

Generation 1 for the example of Section 27.7

	String	$p2$	$p1$	Fitness value
1	10101000110100000111	−1.6	−2.4	0.1735
2	00100100010111111010	−3.6	−0.059	0.0125
3	10000000101010100100	0.02	1.6	0.0139
4	00111001000100001000	−2.8	−2.4	0.0261
5	10101011001010001110	1.7	1.4	0.0388
6	00111000001011110101	−2.8	2.4	0.0006
7	10100010110110011001	1.4	−1.0	0.4316
8	10111010001011110110	2.3	2.4	0.0026
9	01100100010110101010	−1.1	−0.84	0.2996
10	10100000110100010111	1.3	−2.3	0.2310
11	01001011100100000000	−2.1	−2.5	0.0553
12	10000000011000010110	0.0098	0.21	0.2645
13	10000000000011010010	0.0	−2.9	0.1040
14	11000100100101011000	2.7	−1.6	0.1646
15	01111000010110101111	−0.30	−0.79	0.4319
16	01100110000100100111	−1.0	−2.1	0.1914
17	01000100010110110111	−2.3	−0.71	0.0971
18	10001110111011111111	0.58	2.5	0.0038
19	01000010011000011110	−2.4	0.29	0.0470
20	01100100010010100111	−1.1	−3.4	0.0321

Generation 1 summary:
Maximum fitness, 0.4319.
Average fitness, 0.1320.
Best ever, 0.4883.

some generations are shown. As the procedure is repeated several times, it can be seen that the results are improving and the optimum is found around the tenth generation. As the algorithm proceeds, the solutions tend to cluster around the maximum in the solution space. The fact that not all candidate solutions are close to the optimum illustrates the poor search precision of the algorithm. In this example the only stopping criterion is the maximum number of generations which is set to 30.

It can be seen that although the algorithm converged around the tenth generation, due to the stochastic properties of the algorithm further scattering can occur. The performance plot is shown in Fig. 27.15.

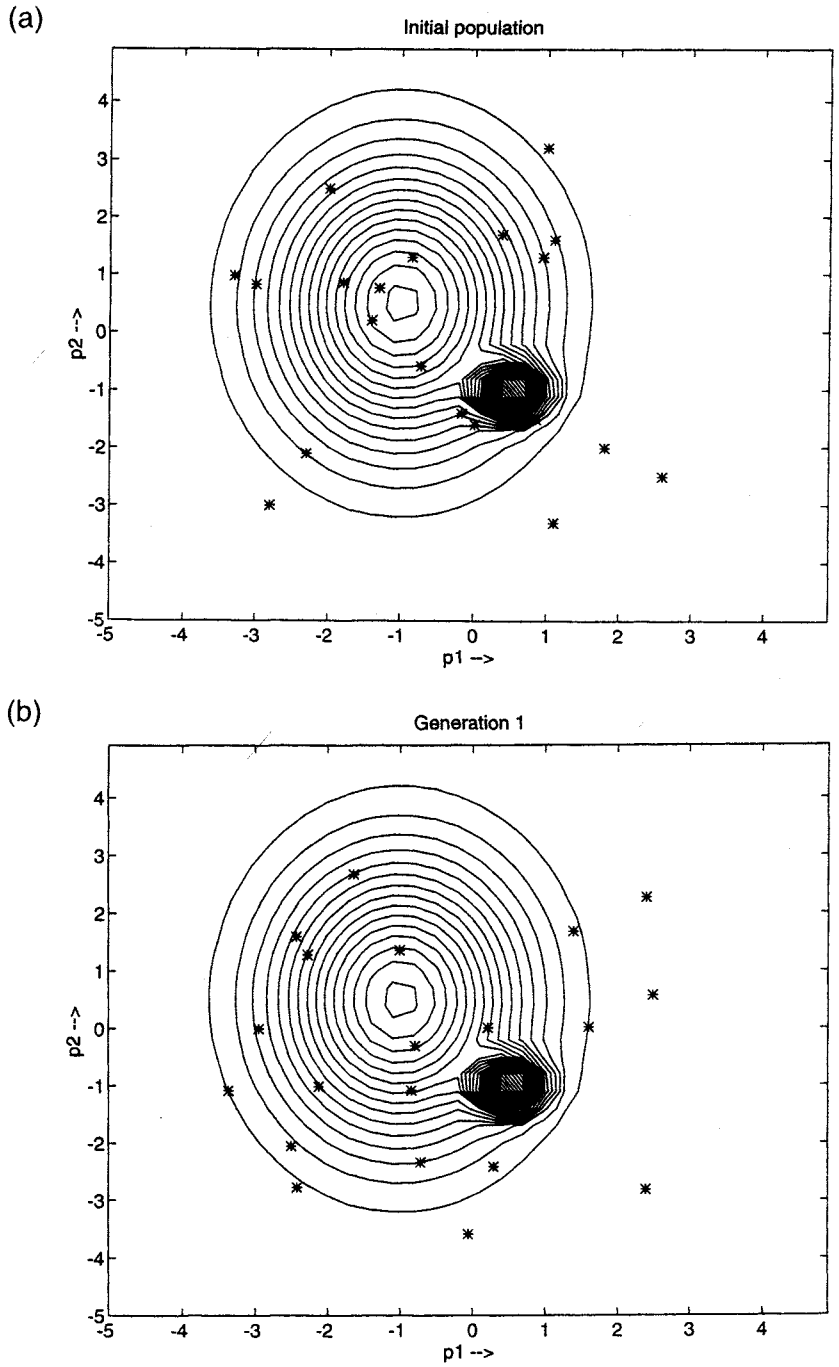
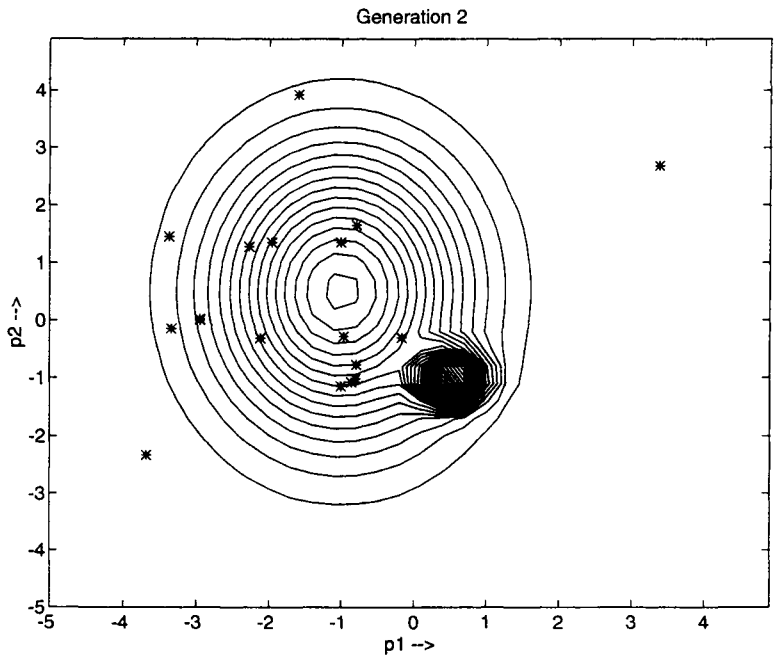


Fig. 27.14. Scatterplot of the initial population (a) and of subsequent populations from generations 1 to 15 (b–p) on the contour plot of the function to be optimized.

(c)



(d)

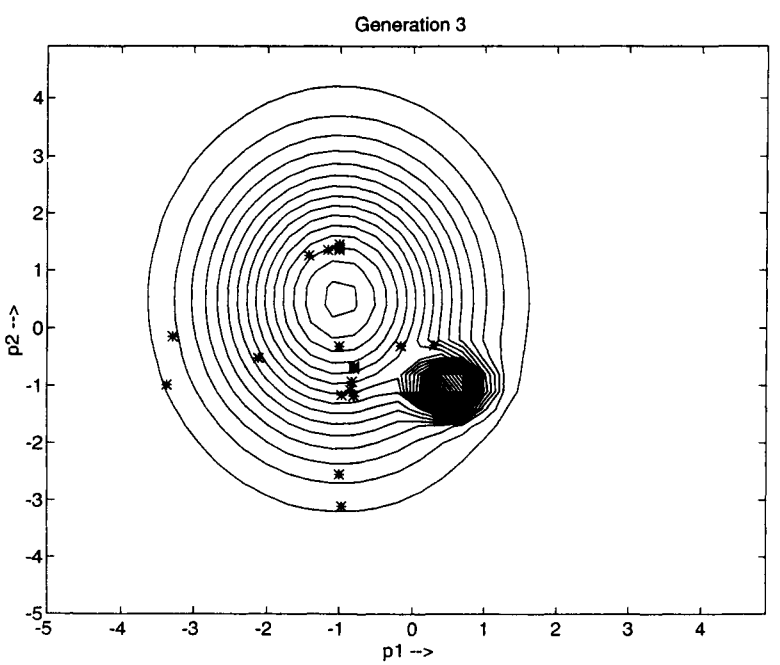
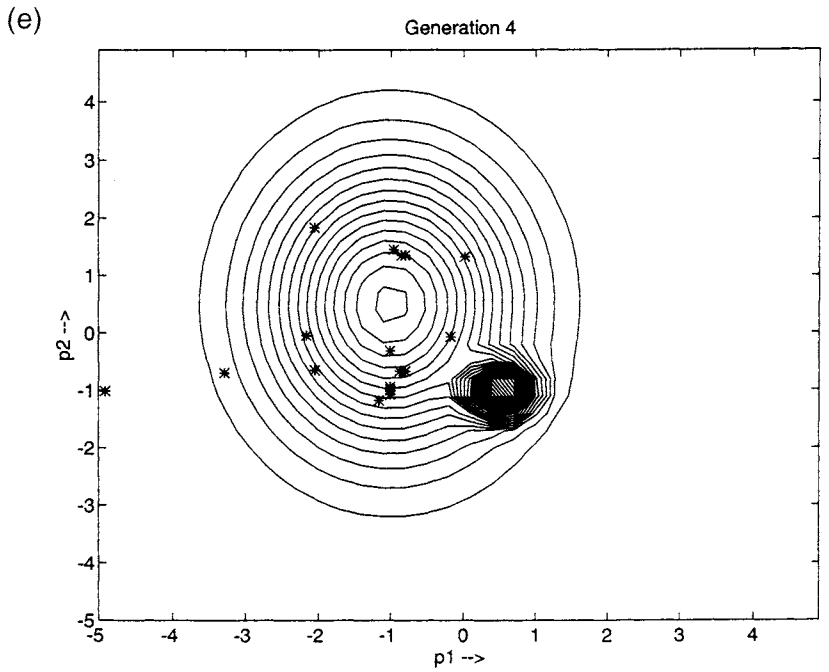


Fig. 27.14 continued.



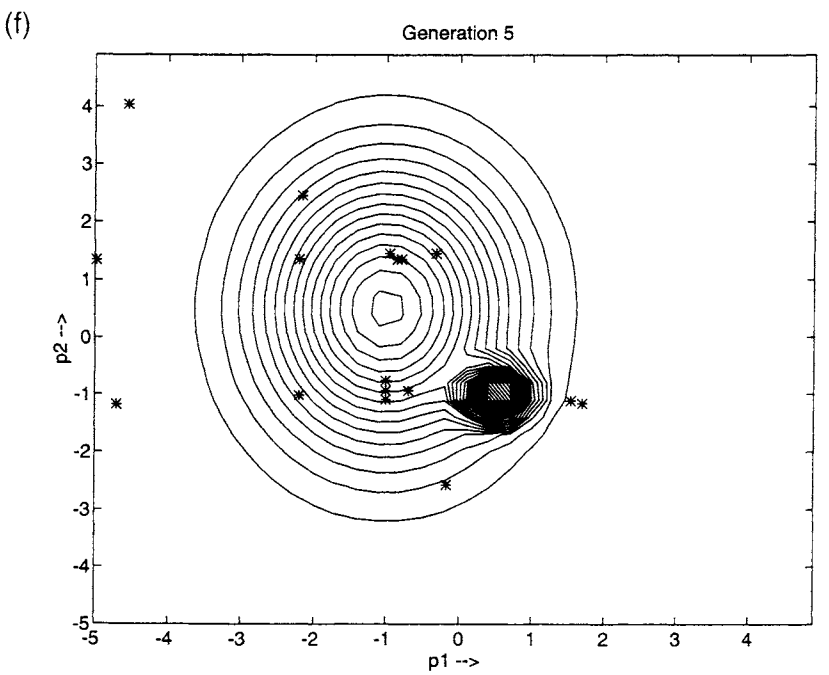
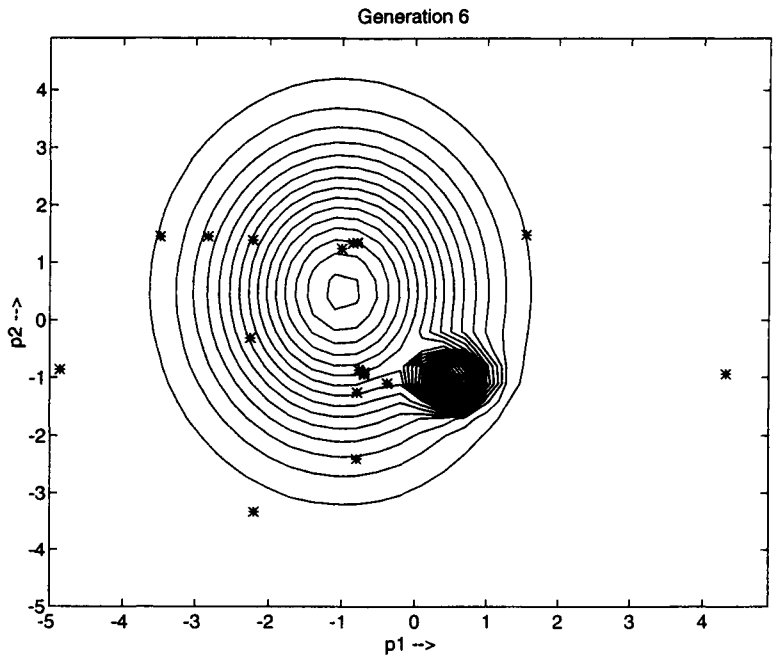


Fig. 27.14 continued.

(g)



(h)

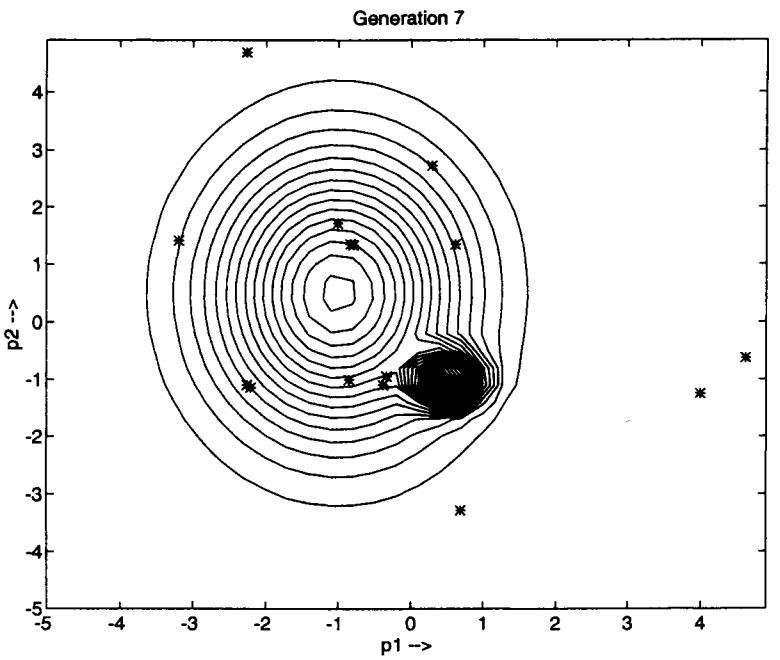
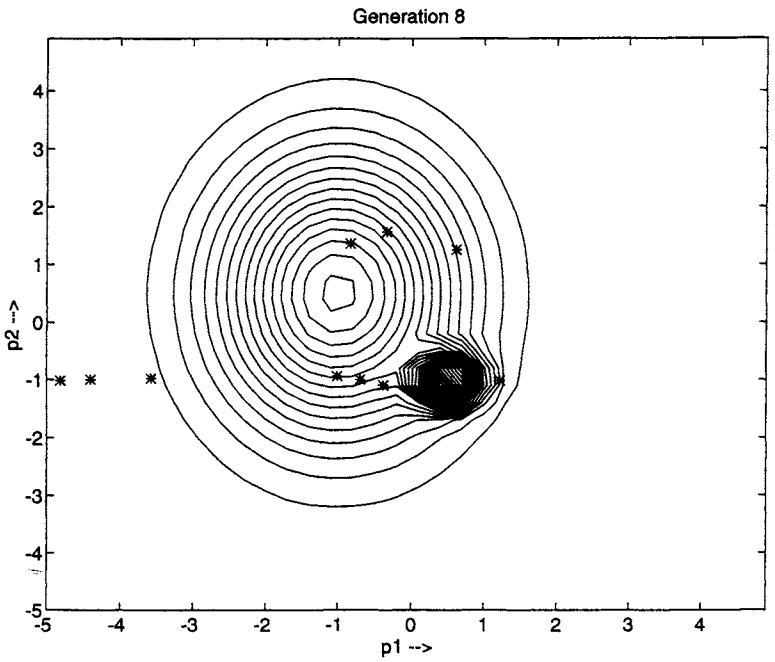


Fig. 27.14 continued.

(i)



(j)

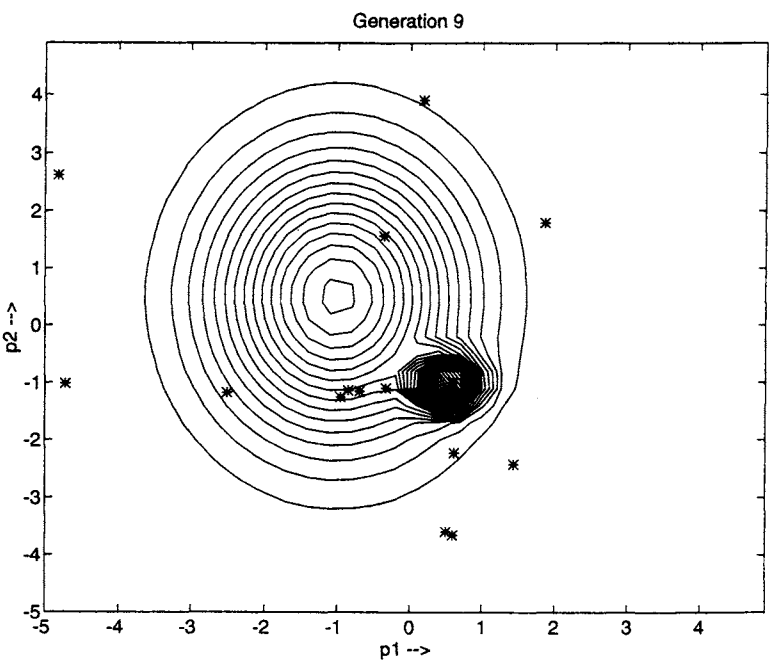
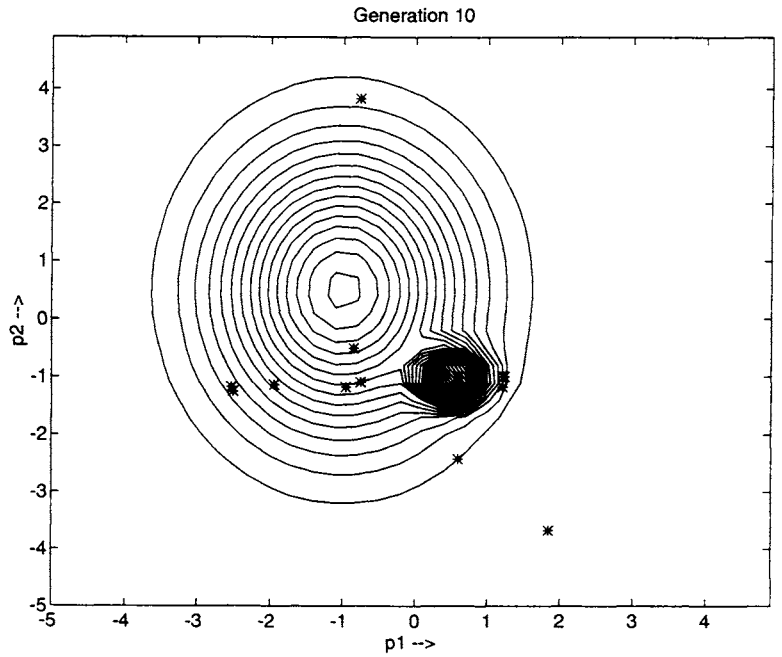


Fig. 27.14 continued.

(k)



(l)

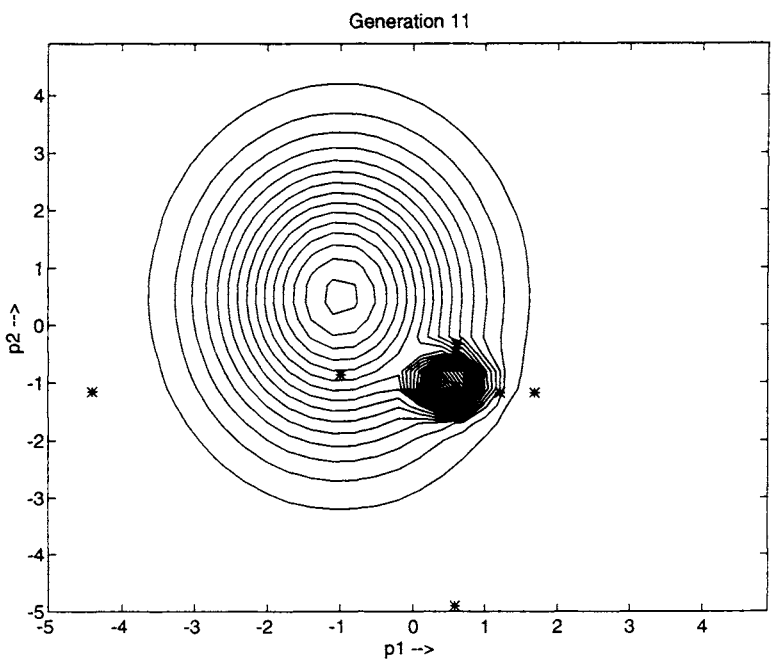
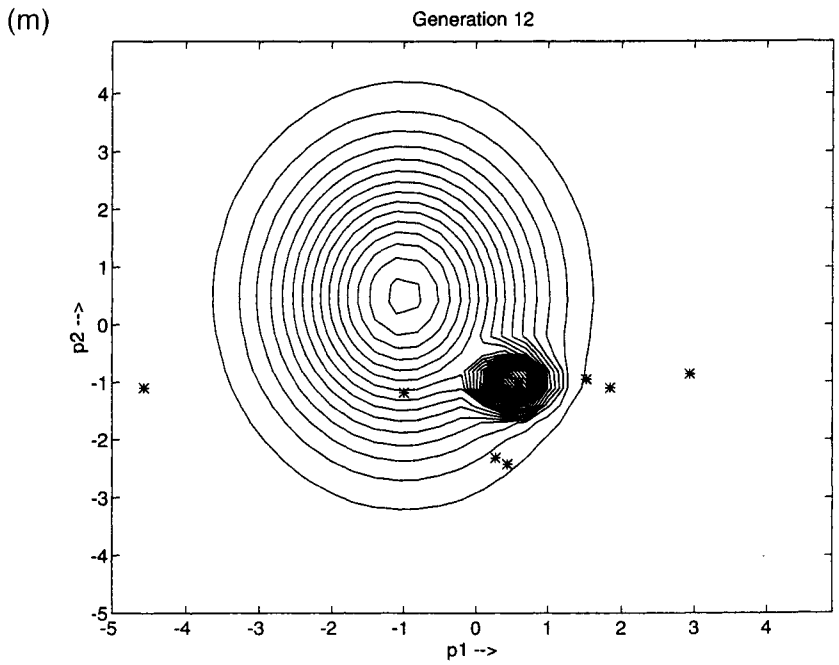


Fig. 27.14 continued.



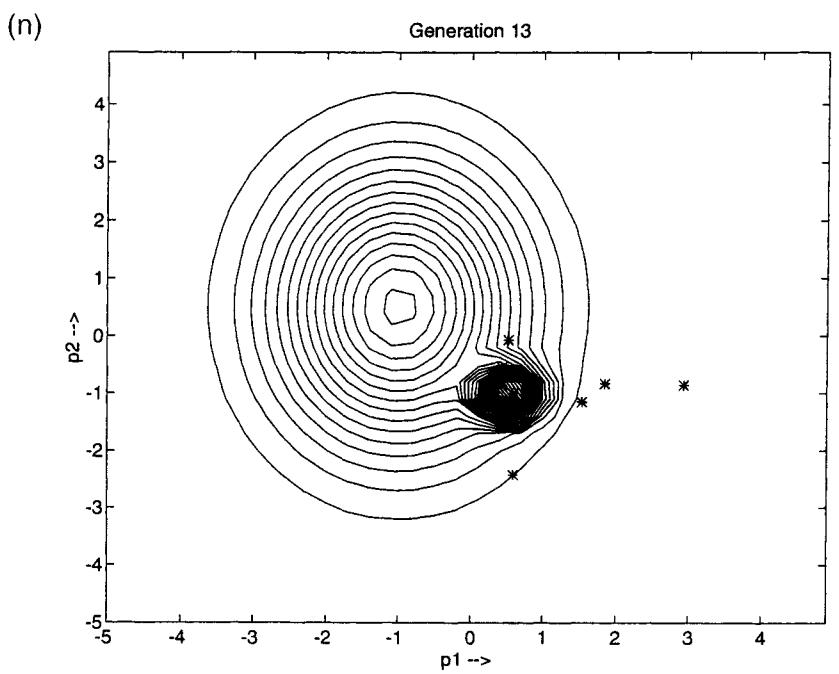
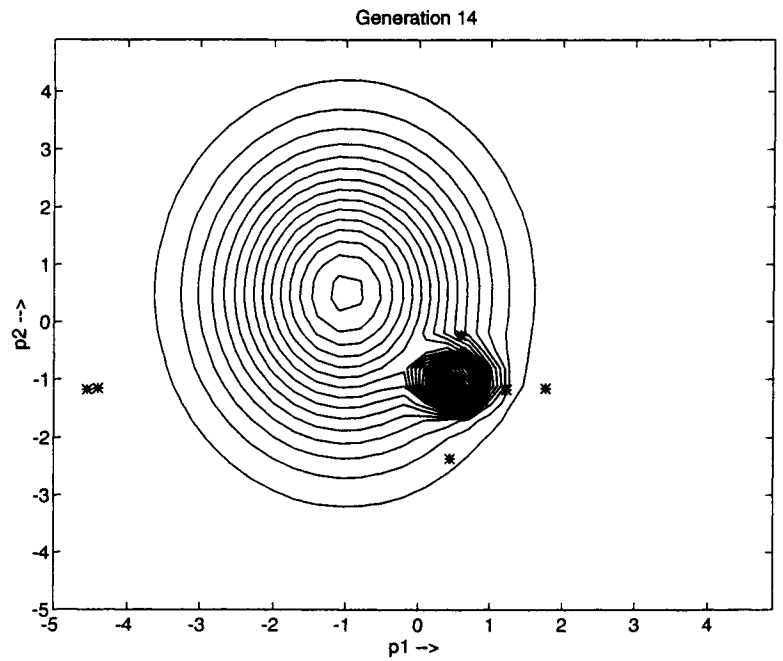


Fig. 27.14 continued.

(o)



(p)

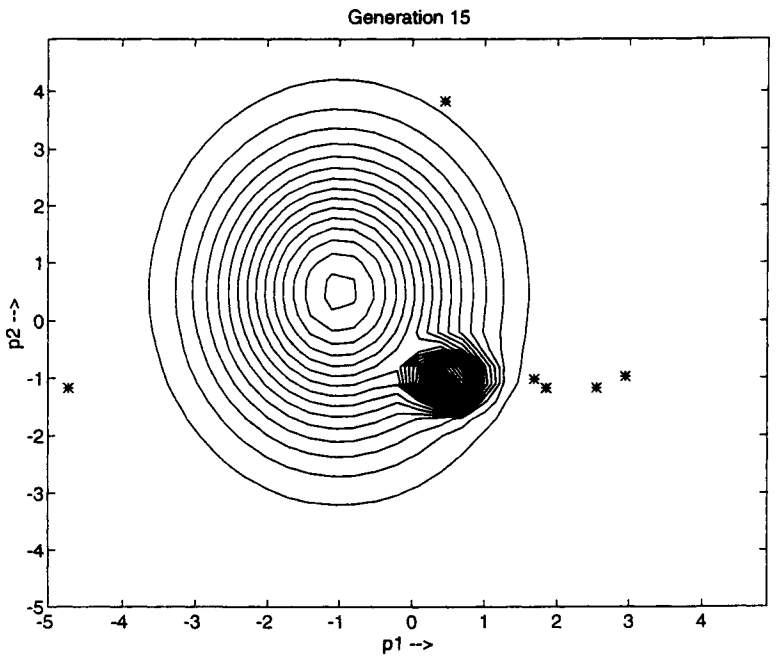


Fig. 27.14 continued.

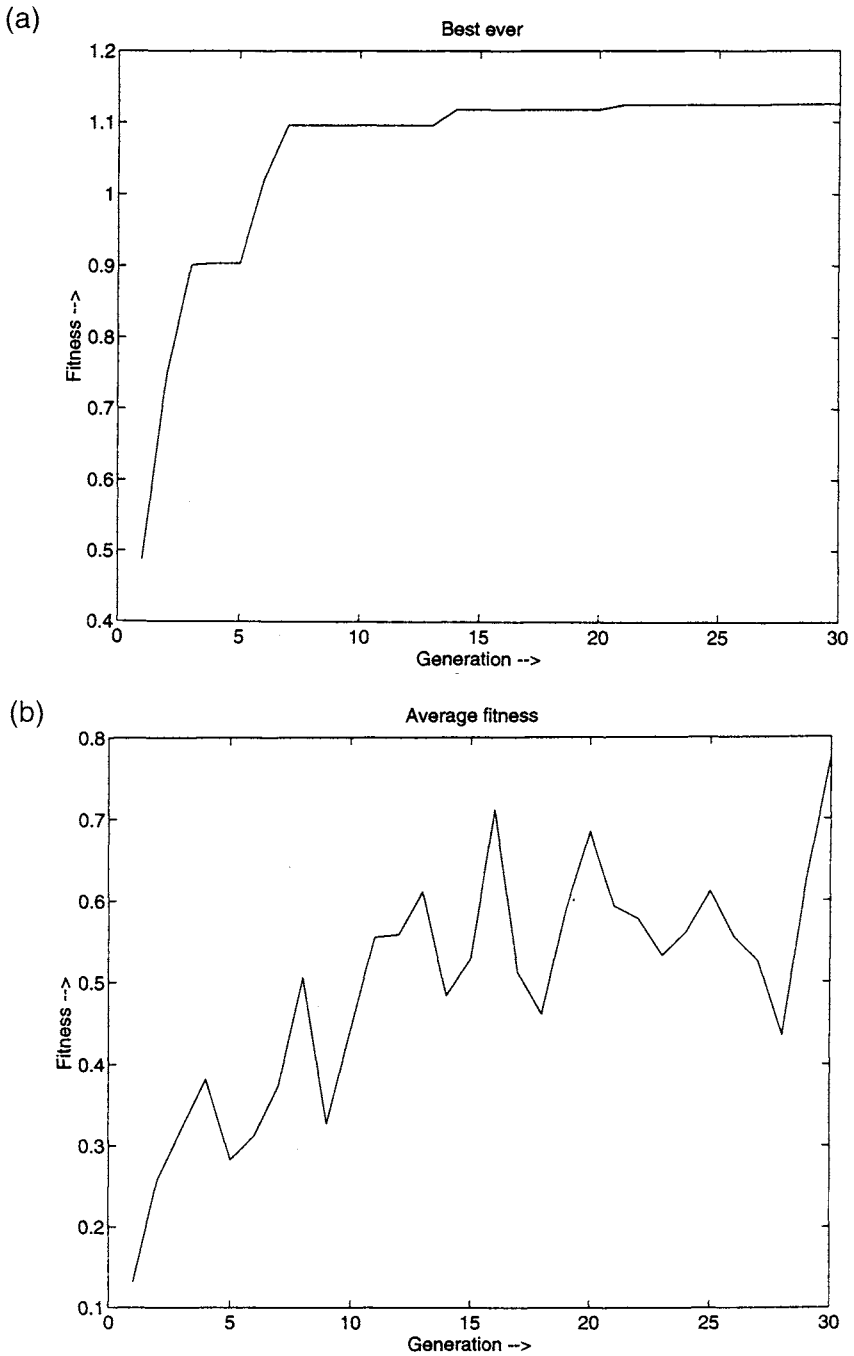


Fig. 27.15. Performance plot of the GA of Section 27.7. Plot of the best ever solution (a) and of the average of each generation (b).

27.8 Applications

The field of GAs is relatively young. The first applications were reported around 1960 when Holland introduced the method [24]. It is however only since the 1980s that the number of publications increases exponentially, mainly due to advances in the computer technology. In 1985 the first conference on GAs took place and the first textbook appeared in 1989 [3]. An extensive bibliography was made by Alander [25]. He compiled more than 3700 references on GAs and conferences from many different fields of science. He also made a separate bibliography on GAs in chemistry and physics. This bibliography contains more than 300 articles from 100 journals from which about 200 articles are on chemistry.

Examples in chemistry can be found in the three different application types for GAs (numerical, sequencing and subset selection problems). In Table 27.6 some examples and references are given in the three application fields.

TABLE 27.6

Applications of GAs in chemistry

Numerical Problems:
Estimation of model parameters;
– fitting IR spectra of PET yarns [9];
– nonlinear chromatographic behaviour [6];
– kinetic parameters [26];
– LMS regression parameters [27];
Optimization of statistical quality control parameters [28];
Multicriteria optimization [29];
PLS, calibration [30, 31];
Quantitative structure activity relationships [32];
Molecular modelling, molecular design [22,33–43];
Neural network training [44];
Display of chemical structures [45].
Sequencing problems:
Sequential assignment of NMR spectra of proteins [40];
Design of sequence of distillation columns [8].
Subset selection problems:
Wavelength selection for multivariate calibration [46–48];
Filter selection for multivariate image analysis [49];
Clustering of large datasets [50].

It is clear from these examples that there are many possibilities for GAs in chemistry and probably many more applications and useful results will emerge during the coming years. An important aspect that will need proper attention for certain applications is the validation of the result found with the GA. The application of wavelength selection for multivariate calibration is such an example. Many combinations are tried out by the GA and the probability to find chance correlations is considerable [51].

The major drawback and the bottleneck of the use of GAs is the configuration problem. For the moment no good solutions are available to overcome this problem. It is partly due to this drawback that interest in other global search strategies, such as simulated annealing and tabu search is increasing. These methods pose much less configuration problems.

27.9 Simulated annealing

The basic idea for the simulated annealing (SA) algorithm was provided by Metropolis et al. in 1953 [52]. They published an algorithm that simulates the controlled cooling of material in a heat bath to a low-energy state, a process which is known as annealing. If the cooling process is too fast, a number of imperfections arise in the solid state. Kirckpatrick et al. [53] found thirty years later that the same idea could be used to solve general optimization problems.

27.9.1 Principle of simulated annealing

The approach resembles classical optimization (e.g. steepest descent or Simplex), but has the ability to overcome its major difficulty, namely to be trapped in local optima. As mentioned earlier, the solutions found by local optimization are determined by the starting position. This is caused by the fact that they always move in the direction that optimizes the objective function. In this way they end up in the optimum closest to the starting point. A way to alter this behaviour is to allow also steps in a direction that yields inferior solutions. Since the overall objective is to find the optimal solutions, these steps in the 'wrong' direction must be taken carefully. In the simulating annealing method they are allowed, but controlled by a probability function that has its roots in Metropolis' work in statistical thermodynamics.

According to the laws of thermodynamics the probability that a system moves to a state with a higher energy is given by:

$$p(\delta E) = e^{-\frac{\delta E}{kT}}$$

$p(\delta E)$ is the probability that at a temperature, T, the system moves to a state with an energy that is a value δE higher; k is the Boltzman constant. In the Metropolis

algorithm a perturbation is induced in the system and the associated energy change is calculated. When the energy is lower the perturbation is accepted; when the energy is higher the perturbation is accepted with the probability given above. This process is iterated while the temperature decreases until the system has cooled down into a frozen state, i.e. no changes that yield an energy increase are accepted.

The different states of the system represent the different candidate solutions. The energy of the state can be seen as the value of the objective function for the specific solution. The perturbations to move into another state can be compared with moving to a neighbour candidate solution. The frozen state corresponds with the final solution, found by the algorithm. The temperature is then a control parameter. In simulated annealing terminology, the control parameter that determines the probability of accepting inferior solutions is still called the temperature. The Boltzman constant is not retained in the terminology, instead a cooling parameter α, between zero and one is defined to allow the temperature to decrease as the algorithm progresses (see further).

Any local optimization method can be converted into a simulated annealing strategy, by allowing steps in an inferior direction according to a certain probability. A classical local optimization (minimization) procedure can be summarized as:

1. Select a starting position: solution S_0.
2. Evaluate S_0 by calculating the objective function, $f(S_0)$.
3. Search for a neighbouring solution, S, using a predefined stepsize such that $f(S) < f(S_0)$ by a suitable method (e.g. steepest descent).
4. Replace S_0 by S.
5. Repeat the last two steps until no better neighbouring solutions can be found.

A simulated annealing procedure can be summarized as:

1. Select a starting position, S_0.
2. Evaluate S_0 by calculating $f(S_0)$.
3. Select an initial value of the control parameter, the temperature, T.
4. Select a value of the control parameter α, to reduce the temperature.
5. Select randomly a neighbouring solution S.
6. Calculate $\delta = f(S) - f(S_0)$.
7. If $\delta < 0$ take S as the new solution.
8. If $\delta > 0$ accept S with a probability, $\exp(-\delta/T)$; keep S_0 with a probability, $1 - \exp(-\delta/T)$.
9. Repeat the steps 5–8 a number of times, n_{it}.
10. Set the control parameter $T = \alpha T$.
11. Repeat steps 5–10 until convergence.

A convenient method to implement step 8 is to select a random number in the interval (0,1). When this random number is smaller than the value of $\exp(-\delta/T)$ the attempted candidate solution is accepted; when it is larger, the attempted solution is not accepted.

27.9.2 Configuration parameters for the simulated annealing algorithm

From the previous summary it is clear that some decisions must be taken before applying simulated annealing. They concern a number of problem specific aspects, such as the cooling scheme.

The cooling scheme is defined by the initial temperature and the rate at which the temperature decreases. This rate is determined by the number of states, n_{it}, that are investigated at each temperature and by the control parameter α. When the initial temperature is too high, almost all inferior solutions will be accepted and the search becomes a random search. When the temperature is too low, no inferior solutions will be accepted, and the search reduces to a slow version of a classical local search such as Simplex or hill climbing. The optimal value is problem dependent. In general it is assumed that a probability between 0.5 and 0.9 for accepting inferior solutions is a good guess. This means that a suitable initial temperature depends on the average difference in the value for the objective functions for neighbouring solutions, i.e. the form of the objective function. When not enough prior knowledge is available the temperature can initially be increased. The probability of accepting inferior solutions is determined experimentally. When the temperature that yields a suitable probability is reached, the cooling process can be initiated. This process can be compared to the process of first heating the system until it is melted before starting the cooling process.

According to theoretical considerations, the value of α should be high (0.8–0.95) which implies a slow cooling process. Other, more complicated, cooling schemes are also possible. The number of repetitions at each temperature, n_{it}, depends on the size of the neighbourhood and may be different for different temperatures. It is, e.g., important that the algorithm spends enough time in the neighbourhood of a (local) optimum, to explore it fully. Therefore n_{it} is usually increased when the program progresses (e.g. depending on the ratio of superior versus inferior solutions). Another approach to determine n_{it} is to change the temperature when a certain number of attempted solutions are accepted. This implies that at the beginning of the procedure the temperature changes faster than at lower temperatures.

As well as the previous generic decisions, some problem-specific decisions should also be considered. It has been implicitly accepted in the previous algorithm that the neighbourhood candidate solutions can be easily defined and remain unchanged during the process. In practice the neighbourhood definition, i.e. the step size, may require a considerable amount of prior knowledge. It may also be desirable to decrease the step size when approaching an optimum.

Many modifications to the basic algorithm have been proposed and tried out. The most important concern different cooling schemes and the use of alternative functions to determine the probability to accept the inferior solutions. A full

explanation of these modifications is outside the scope of this book. Moreover, in practice the basic algorithm seems to perform best. The interested reader is referred to the books by Reeves [54] and Aarts and Korst [55].

27.9.3 Applications

Since Kirckpatrick showed the possibilities of the simulated annealing algorithm of Metropolis, numerous applications have been published in different fields of sciences. A good overview can be found in Aarts and Korst [55]. In analytical chemistry Kalivas et al. [56] successfully applied simulated annealing for selecting an optimal calibration set for NIR determinations. They also compared the SA with Simplex optimization and found the SA superior to Simplex for wavelength selection in UV-VIS. Lucasius et al. [48] performed a comparative study involving genetic algorithms and simulated annealing for wavelength selection in multicomponent analysis. Different optimization criteria were tested. No general conclusions could be drawn since the performance depends on the domain characteristics which determine the fitness landscape and on the configuration settings of the algorithms. Other comparative studies have been performed in various domains [39,57].

27.10 Tabu search

Tabu search (TS) is yet another strategy to solve difficult optimization problems with many local optima. It was introduced first in 1977 by Glover [58]. The first publications on the theory of TS appeared in 1986 [59–61] and the first tutorial in 1989 [62]. Together with simulated annealing and GAs, TS has been selected as 'extremely promising' for the future treatment of practical applications [54]. Just like simulated annealing, it is basically an enhancement of a local optimization strategy. It first searches an optimum. The position of the optimum is 'remembered' and is avoided in future searches.

In the same way as in a local search method, the TS searches the environment of the initial selected candidate solution. The definition of this neighbourhood is an important, problem specific decision. All valid candidate solutions are evaluated and the best one is selected. In a separate list (the tabu list) this solution is saved. When the procedure reaches a (local) optimum a classical local search strategy would stop here but TS continues the search. It evaluates again the environment of the local optimum and the best solution in it is selected. It allows thus, such as in simulated annealing that worse solutions are accepted. The only restriction for a candidate in the environment to be selected as the next solution is that it is not listed on the tabu list. This means that the procedure will escape from the local optimum but in a different direction as it reached it. There are two

drawbacks to the procedure as described above. Firstly, as the procedure continues the tabu list grows and becomes too large to be practically applicable. Secondly, the demand of never ever visiting a location twice is too restrictive. It is possible that a certain position must be passed twice to allow exploration of another direction. Therefore in practice the tabu list is limited to about 10 steps. The most recently visited position replaces the oldest.

Improvements and alterations have been proposed. One improvement is to introduce long-term and intermediate-term memory tabu lists, in addition to the short-term memory tabu list as described above. Tabu search has been hybridized with steepest descent methods and with GAs. Applications of tabu search have been reported for scheduling problems, layout and circuit design, clustering, neural network training and sequencing problems. To the authors' knowledge no applications have been reported in chemistry.

References

1. C. Darwin, On the Origin of Species. John Murray, London, 1959.
2. J.H. Holland, Adaption in natural and artificial systems. University of Michigan Press, Ann Arbor, MI, 1975, Revised Print: MIT Press, Cambridge, MA, 1992.
3. D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Massachusetts, 1989.
4. L. Davis (Ed.), Handbook of Genetic Algorithms. Van Nostrand Reinhold, New York, 1991.
5. D.E. Goldberg and K. Deb, A comparative analysis of selection schemes used in genetic algorithms, in: Foundations of Genetic Algorithms, G.J.E. Rawlins (Ed.), Morgan Kaufmann, San Mateo, 1991, pp. 69–93.
6. R.M. Lopes Marques, P.J. Schoenmakers, C.B. Lucasius and L. Buydens, Modelling chromatographic behaviour as a function of pH and solvent composition in RPLC. Chromatographia, 36 (1993) 83–95.
7. L. Davis, A genetic algorithms tutorial, in: Handbook of Genetic Algorithms, L. Davis (Ed.), Van Nostrand Reinhold, New York, 1991, pp. 1–101.
8. C.A. Floudas and P.M. Pardolas, A collection of test problems for constrained global optimization algorithms, G. Groos and J. Hartmanis (Eds.), Springer-Verlag, Berlin, 1990.
9. A.P. de Weijer, C.B. Lucasius, L.M.C. Buydens, G. Kateman, H.M. Heuvel and H. Mannee, Curve-fitting using natural computation. Anal. Chem., 66 (1994) 23–31.
10. R. Nakana, Y. Davidor and T. Yamada, Optimal population size under constant computation cost in parallel problem solving from nature 3, Y. Davidor, H.P. Schwefel and R. Manner (Eds), Springer Verlag, Berlin, 1994, pp. 130–138.
11. D.E. Goldberg, K. Deb and J.H. Clark, Accounting for noise in the sizing of populations in foundations of genetic algorithms 2, L.D. Whitley (Ed.), Morgan Kaufmann, San Mateo, 1993, pp. 127–140.
12. A.H.C. van Kampen, C.S Strom and L.M.C. Buydens, Lethalization, penalty and repair functions for constraint handling in the genetic algorithm methodology. Chemom. Intell. Lab. Syst., 34 (1996) 55–68.
13. C.B. Lucasius and G. Kateman, Understanding and using genetic algorithms. Part 2. Representation, configuration and hybridization. Chemom. Intell. Lab. Syst., 25 (2) (1994) 99–146.

14. W.M. Spears and K.A. de Jong, An analysis of multi-point crossover, in: Foundations of Genetic Algorithms, G.J.E. Rawlins (Ed.), Morgan Kaufmann, San Mateo, 1991, pp. 310–315.
15. G. Syswerda, Uniform cross-over, in: Proceedings of the Third International Conference on Genetic Algorithms. Morgan Kaufmann, San Mateo, 1989, pp. 2–9.
16. A. Wright, Genetic algorithms for real parameter optimization, in: Foundations of Genetic Algorithms, G.J.E Rawlins (Ed.). Morgan Kaufmann, San Mateo, 1989, pp. 205–218.
17. L.J. Eschelman and J.D. Schafer, Real-coded genetic algorithms and interval schemata, in: Foundations of Genetic Algorithms 2, L.D. Whitley (Ed.). Morgan Kaufmann, San Mateo, 1993, pp. 187–202.
18. D.E. Goldberg, Real coded genetic algorithms, virtual alphabets and blocking. Complex Systems, 5 (1991) 139–167.
19. W.M. Spears, Cross-over or mutation? in: Foundations of Genetic Algorithms 2, L.D. Whitley (Ed.). Morgan Kaufmann, San Mateo, 1993, pp. 221–238.
20. A.H.C. van Kampen, L.M.C Buydens, C.B. Lucasius and M.J.J. Blommers, Optimisation of metric matrix embedding by genetic algorithms. J. Biomolec. NMR, 7 (3) (1996) 214–224.
21. J.D. Schaffer, R.A. Caruani, L.J. Eschelman and R. Das, A study of control parameters affecting online performance of genetic algorithms for function optimization, in: Proceedings of the Third International Conference on Genetic Algorithms, J.D. Schaffer (Ed.). Morgan Kaufmann, San Mateo, 1989, pp. 51–60.
22. M.L.M. Beckers, L.M.C. Buydens, J.A. Pikkemaat and C. Altona, Application of a genetic algorithm in the conformational analysis of methylene-acetal-linked thymine dimers in DNA: comparison with distance biometry calculations. J. Biomolec. NMR, 7 (1997) 25–34.
23. C.B. Lucasius and G. Kateman, Understanding and using genetic algorithms. Part 1. Concepts, properties and context. Chemom. Intell. Lab. Syst., 19 (1993) 1–33.
24. A.S. Fraser, Simulation of genetic systems. J. Theoretical Biol., 2 (1962) 329–346.
25. J.T. Alander, Indexed bibliography of genetic algorithms in chemistry and physics, Technical report of the University of Vaasa, Department of Information Technology and Production Economics, 1996.
26. D.B. Hibbert, Genetic algorithms for the estimation of kinetic parameters. Chemom. Intell. Lab. Syst., 19 (1993) 319–329.
27. P. Vankeerberghen, J. Smeijers-Verbeke, R. Leardi, C.L. Carr and D.L. Massart, Robust regression and outlier detection for non-linear models using genetic algorithms. Chemom. Intell. Lab. Syst., 28 (1995) 73–88.
28. A.T. Hatjimihail, Genetic algorithms-based design and optimization of statistical quality control procedures. Clin. Chem., 38 (1993) 1972–1978.
29. D. Wienke, C.B. Lucasius and G. Kateman, Multicriteria target optimization of analytical procedures using genetic algorithms, 1. Theory, numerical simulations and applications to atomic emission spectroscopy. Anal. Chim. Acta, 265 (1992) 211–225.
30. W.J. Dunn and D. Rogers, Genetic Partial Least Squares in QSAR in Genetic Algorithms in Molecular Modeling. Academic Press, Harcourt Brace, London, 1996, pp. 109–130.
31. D. Wienke, C.B. Lucasius, M. Ehrlich and G. Kateman, Multicriteria target optimization of analytical procedures using genetic algorithms, 2. Polyoptimization of the photometric calibration graph of dry glucose sensors for quantitative clinical analysis. Anal. Chim. Acta, 271 (1993) 253–268.
32. M. Hahn and D. Rogers, Receptor surface models, 2. Application to quantitative relationship studies. J. Med. Chem., 38 (1995) 2091–2102.
33. J. Devillers (Ed.), Genetic algorithms in molecular modeling. Academic Press, Harcourt Brace, London, 1996.

34. D.E. Clark and D.R. Westhead, Evolutionary algorithms in computer-aided molecular design. J. Computer-Aided Molec. Des., 10 (1996) 337–358.
35. T. Brodmeier and E. Pretsch, Application of genetic algorithms in molecular modeling. J. Computat. Chem., 15 (6) (1994) 588–595.
36. G. Jones, R.D. Brown, D.E. Clark, P. Willett and R.C. Glen, Searching databases of two-dimensional and three-dimensional chemical structures using genetic algorithms, in: Proceedings of the Fifth International Conference on Genetic Algorithms. Morgan Kaufmann, San Mateo, 1993, pp. 567–602.
37. E. Fontain, Application of genetic algorithms in the field of constitutional similarity. J. Chem. Inf. Computer. Sci., 32 (1992) 748–752.
38. R. Unger and J. Moult, Genetic algorithms for protein folding simulations. J. Molec. Biol., 231, 75–81.
39. P. Tuffrey, C. Etchebest, S. Hazout and R. Lavery, A critical comparison of search algorithms applied to the protein side-chain conformation. J. Computat. Chem., 14, 790–798.
40. R. Wehrens, C.B. Lucasius, L.M.C. Buydens and G. Kateman, Sequential assignment of 2D-NMR spectra of proteins using genetic algorithms. J. Chem. Inf. Computer Sci., 33 (1993) 245–251.
41. C.B. Lucasius, M.J.J. Blommers, L.M.C. Buydens and G. Kateman, A genetic algorithm for conformational analysis of DNA, in: The Handbook of Genetic Algorithms, L. Davis (Ed.). Van Nostrand Reinhold, New York, 1991, pp. 251–281.
42. S.P. van Helden, H. Hamersma and V.J. van Geerestein, Prediction of progesterone receptor binding of steroids using a combination of genetic algorithms and neural networks, in: Genetic Algorithms in Molecular Modeling. Academic Press, Harcourt Brace, London, 1996, pp. 159–192.
43. P. Willet, Genetic algorithms in molecular recognition and design. Trends Biochem., 13 (1995) 516–521.
44. M. Bos and H.T. Weber, Comparison of the training of neural networks for quantitative X-ray fluorescence spectrometry by a genetic algorithm and backward error propagation. Anal. Chim. Acta, 247 (1991) 97–105.
45. D.B. Hibbert, Generation and display of chemical structures by genetic algorithms. Chemom. Intell. Lab. Syst., 20 (1993) 35–43.
46. R. Leardi, R. Boggia and M. Terrile, Genetic algorithms as a strategy for feature selection. J. Chemom., 6 (1992) 267–281.
47. R. Leardi, Application of genetic algorithms to feature selection under full validation conditions and to outlier detection. J. Chemom., 8 (1994) 65–79.
48. C.B. Lucasius, M.L.M. Beckers and G. Kateman, Genetic algorithms in wavelength selection: a comparative study. Anal. Chim. Acta, 286 (1994) 135–153.
49. W.H.A.M. van den Broek, D. Wienke, W.J. Melssen and L.M.C. Buydens, Optimal wavelength range selection by a genetic algorithm for discrimination purposes in spectroscopic infrared imaging. Appl. Spectros., accepted for publication and scheduled for August 1997.
50. C.B. Lucasius and G. Kateman, Genetic algorithms for large-scale optimization problems in chemometrics: an application. Trends Anal. Chem., 10 (1991) 254–261.
51. D. Jouan-Rimbaud, D.L. Massart and O.E. de Noord, Random correlation in variable selection for multivariate calibration with a genetic algorithm. Chemom. Intell. Lab. Syst., 35 (1996) 213–220.
52. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, Equation of state calculation by fast computing machines. J. Chem. Phys., 21 (1953) 1087–1091.
53. S. Kirckpatrick, C.D. Gellat and M.P. Vecchi, Optimization by simulated annealing. Science, 220 (1983) 671–680.

54. C.R. Reeves (Ed.), Modern heuristic techniques for combinatorial problems. Blackwell Scientific, London, 1993.
55. E. Aarts and J. Korst, Simulated annealing and Boltzmann machines. John Wiley, Chichester, 1990.
56. J.H. Kalivas, Generalized simulated annealing for calibration sample selection from an existing set and orthogonalization of undesigned experiments. J. Chemom., 5 (1991) 37–48.
57. U. Hoerchner and J.H. Kalivas, Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection. Anal. Chim. Acta, 311 (1995) 1–14.
58. F. Glover, Heuristics for integer progamming using surrogate constraints. Decimal Sci., 8, (1977) 156–166.
59. F. Glover, Future path for integer programming and links to artificial intelligence. Computers Ops. Res., 5 (1986) 533–549.
60. F. Glover, Tabu search — Part I. ORSA J. Computing, 1 (1989) 190–206.
61. F. Glover, Tabu search — Part II. ORSA J. Computing, 2 (1990) 4–32.
62. D. Werra and A. Hertz, Tabu search techniques, a tutorial and an application to neural networks. OR Spektrum, 11 (1989) 131–141.

Recommended reading

Books

Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, D.W. Loveland (Ed.). Springer-Verlag, 1992.

Davis L. (Ed.), Genetic Algorithms and Simulated Annealing, Morgan Kaufmann, Los Altos, CA, 1987.

P.J.M. Laarhove and E.H.L. Aarts, Simulated Annealing: Theory and Applications. D. Reidel, Dordrecht, 1987.

Reviews

D.B. Hibbert, Genetic algorithms in chemistry. Chemom. Intell. Lab. Syst., 19 (1993) 277–293.

Articles

J.J. Grefenstette and J.E. Baker, How genetic algorithms work: A critical look at implicit parallelism, in: Proceedings of the Third International Conference on Genetic Algorithms, J.D. Schaffer (Ed.). Morgan Kaufmann, San Mateo, 1989, pp. 20–27.

K. Park, A comparative study of genetic search, in: Proceedings of the Sixth International Conference on Genetic Algorithms, L.J. Eschelman (Ed.). Morgan Kaufmann, San Francisco, CA, 1995, pp. 512–519.

Index