Correlation and Regression

د. سلمى حسين

Correlation: refers to the relationship of two variables or more. (such that, relation between the height & weight).

Correlation Analysis attempts to determine the degree of relationship between variables.

Then, correlation expresses the inter_dependence of two sets of variables upon each other. one variable may be called as (subject) independent and the other relative variable (dependent). Relative variable is measured in terms of subject.

Types of correlation: The most important ones are

① positive and negative: it depends upon the direction of change of the variables, if two variables tend to move together in the same direction.
(i.e) the increase in the value of one variable is accompanied by an increase in the value of the other, or (decrease in one by decrease in the other) Then the correlation called positive.
Correlation called negative, if two variables tend to move in opposite directions (increase by decrease or decrease by increase).

② linear and non_linear correlation: If the ratio between The two variables is a constant, then there will be linear correlation while, if the amount of change in one variable doesn't bear a constant ratio of the amount of change in the other, then the relation is called Curvilinear or non-linear correlation, the graph will be a curve.

③ Simple and multiple correlation: when we study only two variables, the relationship is simple, But in a multiple we study more than two variables.

Correlation Coefficient: when there exists some relationship between two variables, we have to measure the degree of relationship. this measure is called the measure of correlation and denoted by (r)

بين

Co-variation: the Co-variation between the variables x and y is defined as

$$\operatorname{cov}(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{n}$$

Karl Pearson's Coefficient of Correlation: suggested a mathematical method for measuring the magnitude of linear relationship between the two variables. The formula for calculating (r) is:

① $$r = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}$$, σ_x, σ_y are S.D of x & y

② $$r = \frac{\sum xy}{n\sigma_x \sigma_y}$$

③ $$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$, $X = x - \bar{x}$, $Y = y - \bar{y}$

* The third formula is easy to calculate,

يرى بيرسون ان افضل مقياس للارتباط بين متغيرين تختلفان في وحدات
القياس عن طريق حساب الانحرافات لكل من المتغيرين عن وسطهما
الحسابي وقسمة الانحرافات على الانحراف المعياري لكل منهما فنحصل على
مايسمى الوحدات المعيارية لكل متغير.

د. علي حسين

Correlation coefficient (r)

r = +1	r = 0	r = −1
positive	no correlation	negative

درجات العلاقات (r)

−1		−0.9		−0.7		−0.5		−0.3		0		0.3		0.5		0.7		0.9		1
تام	قوي جدا		قوي		متوسط		ضعيف		ضعيف جدا	منعدمة	ضعيف جدا		ضعيف		متوسط		قوي		قوي جدا	تام

ارتباط عكسي [(−1) — 0]

[0 — 1] ارتباط طردي

Example: Find Pearson's coefficient correlation from the following data between height of father (X) and son (Y) [comment on the result]

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Solution:-

X	Y	$x = X - \bar{X}$	x^2	$y = Y - \bar{Y}$	y^2	xy
64	66	−3	9	−2	4	6
65	67	−2	4	−1	1	2
66	65	−1	1	−3	9	3
67	68	0	0	0	0	0
68	70	1	1	2	4	2
69	68	2	4	0	0	0
70	72	3	9	4	16	12
Σ 469	476	0	28	0	34	25

$\bar{X} = \frac{469}{7} = 67$, $\bar{Y} = \frac{476}{7} = 68$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{25}{\sqrt{28 \times 34}} = 0.81$$

Since $r = +0.81$, then the variables are highly positively correlated (i.e) Tall fathers have tall sons.

Another rule: we can also find (r) with the following formula.

د. علي بن

we have $r = \frac{cov(x,y)}{\delta_x \delta_y}$, $cov(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$

$$cov(x,y) = \frac{\sum(xy - \bar{x}y - \bar{y}x + \bar{x}\bar{y})}{n}$$

$$= \frac{\sum xy}{n} - \frac{\bar{x}\sum y}{n} - \frac{\bar{y}\sum x}{n} + \frac{\sum \bar{x}\bar{y}}{n}$$

$$cov(x,y) = \frac{\sum xy}{n} - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y} = \frac{\sum xy}{n} - \bar{x}\bar{y}$$

Now, $\delta_x^2 = \frac{\sum x^2}{n} - \bar{x}^2$, $\delta_y^2 = \frac{\sum y^2}{n} - \bar{y}^2$

$$r = \frac{cov(x,y)}{\delta_x \delta_y} = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\sqrt{\frac{\sum x^2}{n} - \bar{x}^2}\ \sqrt{\frac{\sum y^2}{n} - \bar{y}^2}}$$

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

* here we need not find the mean & standard deviation

Example: Calculate coefficient of correlation from the following data. د. تحسين

X	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

Sol

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
= 45	108	285	1356	597

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

$$= \frac{9 \times 597 - 45 * 108}{\sqrt{\left[9 * 285 - (45)^2\right]\left[9 * 1356 - (108)^2\right]}} = \frac{513}{540} = 0.95$$