Correlation

8

8.1 Causation and Correlation

Suppose we find direct correlation between two variables. But it does not mean that the change in variable "*Y*" is a direct cause of a change in variable "*X*." If at all the change in "*Y*" is directly associated with a change in the variable "*X*," then it would be certain that *X* and *Y* are correlated. The existence of correlation may be due to any one of the following:

8.2 One Variable Being a Cause of Another

The cause variable is taken as an independent variable (X), and the effect variable is considered as a dependent one (Y). Suppose "age" and "height" are correlated. Age is an independent variable which is a cause for change in height, the dependent variable.

8.2.1 Both Variables Being the Result of a Common Cause

Women in a group were followed, after a given operation. The duration of survival and number of children born to a woman were recorded. These factors were related, and it was found that there was a high degree of "positive correlation." It would be interesting to interpret this data in either of the following two ways:

1. Prolonged life of a woman tends to bear more children.
2. Bearing of children tends to prolong the life of a woman.

Note Both these interpretations are absurd. Neither prolonged life has any effect on bearing of children nor bearing of children increases the life span of a woman. One

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_8

can therefore think of some other factors such as age and state of health at the time of operation, which could tend to affect both the survival time and bearing of children.

8.2.2 Chance

Rainfall of some place in north may find high degree of correlation with per acre yield of rice in the south. It would be meaningless to think that the rainfall recorded in the north has any effect on the yield of rice in the south. Such correlations are called spurious or chance correlations. Hence, one must reasonably think of any likelihood relationship existing between the two variables under study.

So, one should be very careful in interpreting the relationship when correlation between the two variables exists.

8.3 Methods of Studying Correlation

1. The scatter diagram
2. Pearson's coefficient of correlation
3. The regression line

8.3.1 The Scatter Diagram

Usually the scale on Y-axis starts from zero though the scale on X-axis need not start from zero. But in cases of "scatter diagram," this restriction on the side of Y-axis is also removed. Both X- and Y-axes may be stared at the minimum values of the respective variables.

8.3.2 Pearson Coefficient of Correlation for Ungrouped Data

Pearson's coefficient of correlation is a measure of the degree of relationship between the two variables. It is denoted by "r" in the case of the sample estimate and by "ρ" in case of the correlation obtained from the whole population. This is also known as the product moment component of correlation. The computation formulae for the both have been illustrated in Table 8.1.

The formulae may also be written in different forms for the sake of convenience in calculations. These are as below:

$$\mathbf{r} = \frac{\sum_1^n \left(Xi - \bar{X}\right) \cdot \left(Yi - \bar{Y}\right)}{\sqrt{\sum_1^n \left(Xi - \bar{X}\right)^2 \cdot \sum_1^n \left(Yi - \bar{Y}\right)^2}}$$

Table 8.1 Pearson's coefficient of correlation formulae

Obtained from the sample	Obtained from the whole population
$\mathbf{r} = \frac{\Sigma \frac{x}{sx} \cdot \Sigma \frac{y}{sy}}{n-1}$	$\boldsymbol{\rho} = \frac{\Sigma \frac{x}{\sigma x} \cdot \Sigma \frac{y}{\sigma y}}{N}$
where	where
$x = X - \bar{X}$	$x = X - \mu X$
$y = Y - \bar{Y}$	$y = Y - \mu Y$
n = no. of pairs of items	N = no. of pairs of items
sx = SD of X-variables	σx = SD of X-variables
sy = SD of Y-variables	σy = SD of Y-variables

$$\mathbf{r} = \frac{\sum_1^n XiYi - n\bar{X}\bar{Y}}{\sqrt{\left[\sum_1^n Xi^2 - n\bar{X}^2\right] \cdot \left[\sum_1^n Yi^2 - n\bar{Y}^2\right]}}.$$

$$\mathbf{r} = \frac{\sum_1^n XiYi - \frac{\sum_i^n Xi \cdot \sum_i^n Yi}{\mathrm{n}}}{\sqrt{\left[\frac{\sum_i^n Xi^2}{1} - \frac{\left(\sum_i^n Xi\right)^2}{n}\right] \cdot \left[\frac{\sum_i^n Yi^2}{1} - \frac{\left(\sum_i^n Yi\right)^2}{n}\right]}}$$

$$\mathbf{r} = \frac{\sum_1^n uv - \frac{\sum_i^n u \cdot \sum_i^n v}{\mathrm{n}}}{\sqrt{\left[\frac{\sum_i^n u^2}{1} - \frac{\left(\sum_i^n u\right)^2}{n}\right] \cdot \left[\frac{\sum_i^n v^2}{1} - \frac{\left(\sum_i^n v\right)^2}{n}\right]}}$$

In the above formula, "u" and "v" are the new variables used to simplify the computation: $u = X - X_o$ and $v = Y - Y_o$, where X_o and Y_o are the assumed means.

Pearson's coefficient of correlation (**r**) can also be computed by the "difference formula" as given below:

$$\mathbf{r} = \frac{\sum_1^n x^2 + \sum_1^n y^2 - \sum_1^n d^2}{2\sqrt{\sum_1^n x^2 \cdot \sum_1^n y^2}}$$

In which $\sum_1^n d^2 = \sum_1^n (x - y)^2$ and $x = X - \bar{X}$; $y = Y - \overline{Y}$.

The above equation can also be modified as below:

$$\mathbf{r} = \frac{\sum_1^n x^2 + \sum_1^n y^2 - \sum_1^n (x - y)^2 - 2\left(\sum_1^n x\right)\left(\sum_1^n y\right)}{2\sqrt{\left[n\sum_1^n x^2 - \left(\sum_1^n x\right)^2\right]\left[n\sum_1^n y^2 - \left(\sum_1^n y\right)^2\right]}}$$

8.3.2.1 Examples Illustrating the Computations of *r*-Test

Example 1 Computations of "*r*" when deviations are taken from their means. Data of height and weight of five students have been tabulated in Table 8.2.

$$\mathbf{r} = \frac{\sum_1^n (Xi - \bar{X}) \cdot (Yi - \bar{Y})}{\sqrt{\sum_1^n (Xi - \bar{X})^2 \cdot \sum_1^n (Yi - \bar{Y})^2}} = \frac{\sum_1^n xy}{\sqrt{\sum_1^n (x)^2 \cdot \sum_1^n (y)^2}} = \frac{55}{\sqrt{20 \times 750}} = \mathbf{0.449}$$

$\mathrm{df} = n-2 = 5-2 = 3$
$\mathbf{r}_{0.05} = \mathbf{0.878}$

Decision
The computed value of $\mathbf{r} = 0.449$ is less than the "table value" of $\mathbf{r}_{0.05} = 0.878$. So, "null hypothesis" (H_o) is accepted. Hence, there is no significant correlation between the height and weight of students.

Example 2 Computations of "*r*" when deviations are taken from the assumed means. Data of height and weight of five students has been tabulated in Table 8.3.

Table 8.2 Data of height and weight for *r*-test

Student	Height "inches" X	Weight "kg" Y	$x = X–\bar{X}$	$y = Y–\bar{Y}$	x^2	y^2	xy
1	72	70	+3	0	9	0	0
2	69	65	0	−5	0	25	0
3	66	50	−3	−20	9	400	60
4	70	80	+1	+10	1	100	10
5	68	85	−1	+15	1	225	−15
Sum	**345**	**350**			**20**	**750**	**55**
Mean	**69**	**70**					

Table 8.3 Data of height and weight for *r*-test

Student	Height "inches" X	Weight "kg" Y	$u = X–Xo$ $X_o = 70$	$v = Y–Yo$ $Y_o = 70$	u^2	v^2	uv
1	72	70	+2	0	4	0	0
2	69	65	−1	−5	1	25	5
3	66	50	−4	−20	16	400	80
4	70	80	0	+10	0	100	0
5	68	85	−2	+15	4	225	−30
Sum			**−5**	**0**	**25**	**750**	**+55**

$$\mathbf{r} = \frac{\sum_1^n uv - \frac{\sum_i^n u \cdot \sum_i^n v}{\mathrm{n}}}{\sqrt{\left[\frac{\sum_i^n u^2}{1} - \frac{\left(\sum_i^n u\right)^2}{n}\right] \cdot \left[\frac{\sum_i^n v^2}{1} - \frac{\left(\sum_i^n v\right)^2}{n}\right]}}$$

$$= \frac{55 - \frac{(-5) \cdot (0)}{5}}{\sqrt{\left[25 - -\frac{(5)^2}{5}\right] \cdot \left[750 - -\frac{(0)^2}{5}\right]}} = \frac{55}{\sqrt{20 \times 750}} = \mathbf{0.449}$$

$\mathrm{df} = \mathrm{n}-2 = 5-2 = 3$
$\mathbf{r}_{0.05} = \mathbf{0.878}$

Decision
The computed value of $\mathbf{r} = 0.449$ is less than the "table value" of $\mathbf{r}_{0.05} = 0.878$. So, "null hypothesis" (H_o) is accepted. Hence, there is no significant correlation between the height and weight of students.

Example 3 Computations of "*r*" from observed data without taking deviations. Data of height and weight of five students has been tabulated in Table 8.4.

$$\mathbf{r} = \frac{\sum_1^n XiYi - \frac{\sum_i^n Xi \cdot \sum_i^n Yi}{n}}{\sqrt{\left[\frac{\sum_i^n Xi^2}{1} - \frac{\left(\sum_i^n Xi\right)^2}{n}\right] \cdot \left[\frac{\sum_i^n Yi^2}{1} - \frac{\left(\sum_i^n Yi\right)^2}{n}\right]}}$$

$$= \frac{24205 - \frac{345 \times 350}{5}}{\sqrt{\left[23825 - \frac{(345)^2}{5}\right] \cdot \left[25250 - \frac{(350)^2}{5}\right]}}$$

$$= \frac{24205 - 24150}{\sqrt{[23825 - 23805] \cdot [25250 - 24500]}} = \frac{55}{\sqrt{20 \times 750}} = \frac{55}{122.47} = \mathbf{0.449}$$

$\mathrm{df} = n-2 = 5-2 = 3$
$\mathbf{r}_{0.05} = \mathbf{0.878}$

Decision
The computed value of $\mathbf{r} = 0.449$ is less than the "table value" of $\mathbf{r}_{0.05} = 0.878$. So, "null hypothesis" (H_o) is accepted. Hence, there is no significant correlation between the height and weight of students.

Table 8.4 Data of height and weight for *r*-test

Student	Height "inches" X	Weight "kg" Y	X^2	Y^2	XY
1	72	70	5184	4900	5040
2	69	65	4761	4225	4485
3	66	50	4356	2500	3300
4	70	80	4900	6400	5600
5	68	85	4624	7225	5780
Sum	**345**	**350**	**23,825**	**25,250**	**24,205**

Table 8.5 Data of height and weight for *r*-test

Student	Height "inches" X	Weight "kg" Y	$x = X–\bar{X}$ $(\bar{X} = 69)$	$y = Y–\bar{Y}$ $(\bar{Y} = 70)$	x^2	y^2	$(x–y)^2$
1	72	70	+3	0	9	0	9
2	69	65	0	−5	0	25	25
3	66	50	−3	−20	9	400	289
4	70	80	+1	+10	1	100	81
5	68	85	−1	+15	1	225	256
Sum	**345**	**350**			**20**	**750**	**660**
Mean	**69**	**70**					

Example 4 Computations of "*r*" by the "difference formula." Data of height and weight of five students has been tabulated in Table 8.5.

$$\mathbf{r} = \frac{\sum_1^n x^2 + \sum_1^n y^2 - \sum_1^n d^2}{2\sqrt{\sum_1^n x^2 \cdot \sum_1^n y^2}} = \frac{20 + 750 - 660}{2\sqrt{20 \times 750}}$$
$$= \frac{110}{2\sqrt{20 \times 750}} = \frac{55}{\sqrt{20 \times 750}} = 0.449$$

$\mathrm{df} = n-2 = 5-2 = 3$
$\mathbf{r}_{0.05} = 0.878$

Decision
The computed value of $\mathbf{r} = 0.449$ is less than the "table value" of $\mathbf{r}_{0.05} = 0.878$. So, "null hypothesis" (H_o) is accepted. Hence, there is no significant correlation between the height and weight of students.

Example 5 The body weights of five chicks were 180, 170,170, 190, and 190 grams, respectively, and their comb weights were found to be 50, 40, 20, 60, and 60 grams, respectively. Find out if there is any correlation between the body weight and comb weight of chicks.

Table 8.6 Body weights and comb weights of chicks

Chick no	Body weight (x)	Comb weight (y)	$u-170$	$v = y-40$
1	180	50	10	10
2	170	40	0	0
3	170	20	0	−20
4	190	60	20	20
5	190	60	20	20
Total			**50**	**30**

Solution

Data has been transformed by subtracting 170 from the body weights and 40 from the comb weights as shown in Table 8.6.

$$\Sigma u^2 = 100 + 400 + 400 = 900$$
$$\Sigma v^2 = 100 + 400 + 400 + 400 = 1300$$
$$\Sigma uv = 100 + 400 + 400 = 900$$

$$\mathbf{r} = \frac{\Sigma uv - \frac{\Sigma u \cdot \Sigma v}{n}}{\sqrt{\left(\Sigma u^2 - \frac{(\Sigma u)^2}{n}\right)\left(\Sigma v^2 - \frac{(\Sigma v)^2}{n}\right)}}$$
$$= \frac{900 - \frac{50 \times 30}{5}}{\sqrt{\left(900 - \frac{50 \times 50}{5}\right)\left(1300 - \frac{30 \times 30}{5}\right)}}$$
$$= \frac{900 - 300}{\sqrt{(900 - 500)(1300 - 180)}}$$
$$= \frac{600}{\sqrt{400 \times 1120}} = \frac{600}{20\sqrt{1120}} = \frac{30}{\sqrt{1120}} = \frac{30}{33.5} = 0.895 = +\mathbf{0.895}$$

$\mathbf{df} = \text{n} - 2 = 5 - 2 = 3$; $\mathbf{r_{0.05}} = 0.878$

Decision

The calculated $\mathbf{r} = +0.895$ is greater than $\mathbf{r_{0.05}} = 0.878$. So, "null hypothesis" (H_o) is rejected ($p < 0.05$). Hence, there is direct correlation between the "body weights" and "comb weights" of chicks.

8.3.3 Regression Line

To determine the amount of change that normally takes place in the Y-variable for a unit change in the X-variable, a line is fitted to the points plotted on the scatter

diagram. This line is described as $Y = a + bx$ and is said to be line of regression of Y on X. Here, "a" and "b" are the two constants: $a = Y$-intercept and $b =$ slope of the regression line. The "b" may also be written as "by" = regression coefficient of Y on X. It is also possible to find $bXY =$ "regression coefficient" of X on Y. There is a definite relationship between "**r**" and these two "regression coefficients" bYX and bXY. The "**r**" is the geometric mean of bXY and bYX.

Therefore:

$$\mathbf{r} = \sqrt{bYX \cdot bXY}; \text{But } bYX = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(X - \bar{X})^2}$$

whereas:

$$\mathbf{r} = \frac{\sum_1^n (Xi - \bar{X}) \cdot (Yi - \bar{Y})}{\sqrt{\sum_1^n (Xi - \bar{X})^2 \cdot \sum_1^n (Yi - \bar{Y})^2}}$$

Hence it can be proved that $bYX = r\sqrt{\frac{(Y - \bar{Y})^2}{(X - \bar{X})^2}} = \frac{sy}{sx}$

8.4 Proportions of "*r*"

1. The "r" values range from -1.00 through 0.00 to +1.00.
2. It is a pure number, independent of the units of measurement of the variables X and Y.
3. If $r = -1$, a perfect inverse linear relationship exists between the variables (e.g., volume $\infty \frac{1}{\text{Power}}$).
4. If $r = -0$, linear relationship between the two variables X and Y does not exist (e.g., number of births registered vs number of cars registered).
5. If $r = +1$, there is a perfect direct linear relationship (e.g. diameter vs circumference).
6. If $r = -0.7$ or + 0.7 in a large set, the degree of relationship between the two variables seems to be high.
7. If $r = +0.6$, it does not mean that 60% of the values are related.
8. The computation of r is valid only if the variables are approximately normally distributed.
9. The r^2 is known as coefficient of determination. If $r^2 = 0.756$, it means that approximately 75.6% of the variation in Y is only due to the linear regression of Y on X.