Measurement of Correlation: Karl Pearson’s Method, Spearman Rank Correlation

Karl Pearson’s Coefficient of Correlation

It is widely used mathematical method wherein the numerical expression is used to calculate the degree and direction of the relationship between linear related variables.

Pearson’s method, popularly known as a Pearsonian Coefficient of Correlation, is the most extensively used quantitative methods in practice. The coefficient of correlation is denoted by “r”.

If the relationship between two variables X and Y is to be ascertained, then the following formula is used:

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

Properties of Coefficient of Correlation

- The value of the coefficient of correlation (r) always lies between±1. Such as: r=+1, perfect positive correlation

  r=-1, perfect negative correlation

  r=0, no correlation

- The coefficient of correlation is independent of the origin and scale.By origin, it means subtracting any non-zero constant from the given value of X and Y the vale of “r” remains unchanged. By scale it means, there is no effect on the value of “r” if the value of X and Y is divided or multiplied by any constant.

- The coefficient of correlation is a geometric mean of two regression coefficient. Symbolically it is represented as:

$$r = \sqrt{b_{xy} + b_{yx}}$$

- The coefficient of correlation is “ zero” when the variables X and Y are independent. But, however, the converse is not true.

Merits:

1. This method indicates the presence or absence of correlation between two variables and gives the exact degree of their correlation.

2. In this method, we can also ascertain the direction of the correlation; positive, or negative.

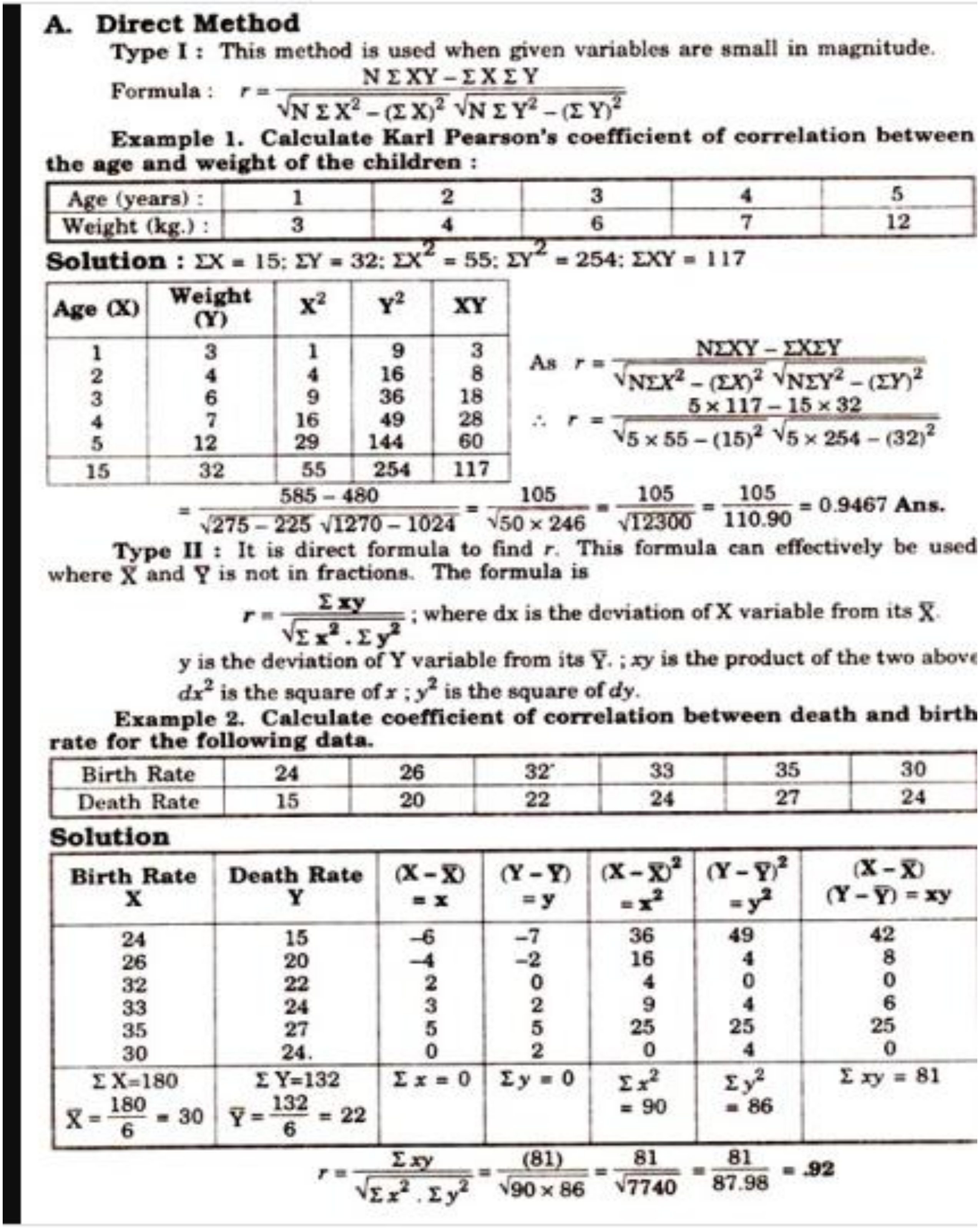
3. This method has many algebraic properties for which the calculation of co-efficient of correlation, and other related factors, are made easy.

Demerits:

1. It is more difficult to calculate than other methods of calculations.

2. It is much affected by the values of the extreme items.

3. It is based on a many assumptions, such as: linear relationship, cause and effect relationship etc. which may not always hold well.



A. Direct Method

Type I : This method is used when given variables are small in magnitude.

Formula : $r = \dfrac{N\,\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{N\,\Sigma X^2 - (\Sigma X)^2}\,\sqrt{N\,\Sigma Y^2 - (\Sigma Y)^2}}$

Example 1. Calculate Karl Pearson's coefficient of correlation between the age and weight of the children :

Age (years) :	1	2	3	4	5
Weight (kg.) :	3	4	6	7	12

Solution : $\Sigma X = 15;\ \Sigma Y = 32;\ \Sigma X^2 = 55;\ \Sigma Y^2 = 254;\ \Sigma XY = 117$

Age (X)	Weight (Y)	X^2	Y^2	XY
1	3	1	9	3
2	4	4	16	8
3	6	9	36	18
4	7	16	49	28
5	12	29	144	60
15	32	55	254	117

As $r = \dfrac{N\Sigma XY - \Sigma X\Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\,\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$

$\therefore\ r = \dfrac{5 \times 117 - 15 \times 32}{\sqrt{5 \times 55 - (15)^2}\,\sqrt{5 \times 254 - (32)^2}}$

$$= \frac{585 - 480}{\sqrt{275 - 225}\,\sqrt{1270 - 1024}} = \frac{105}{\sqrt{50 \times 246}} = \frac{105}{\sqrt{12300}} = \frac{105}{110.90} = 0.9467 \textbf{ Ans.}$$

Type II : It is direct formula to find r. This formula can effectively be used where $\overline{X}$ and $\overline{Y}$ is not in fractions. The formula is

$r = \dfrac{\Sigma\,xy}{\sqrt{\Sigma\,x^2 \cdot \Sigma\,y^2}}$; where dx is the deviation of X variable from its $\overline{X}$.

y is the deviation of Y variable from its $\overline{Y}$. ; xy is the product of the two above

dx^2 is the square of x ; y^2 is the square of dy.

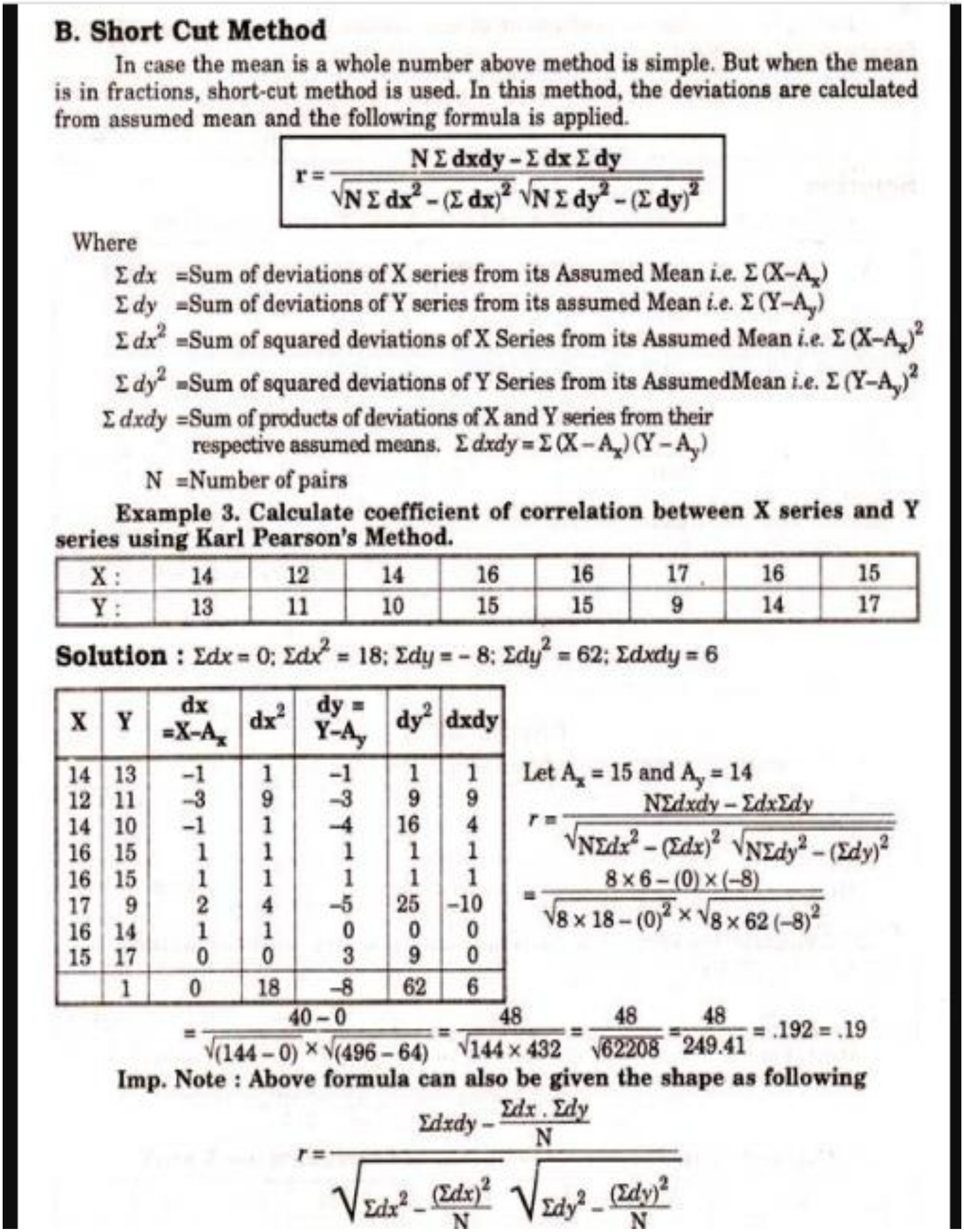
Example 2. Calculate coefficient of correlation between death and birth rate for the following data.

Birth Rate	24	26	32	33	35	30
Death Rate	15	20	22	24	27	24

Solution

Birth Rate X	Death Rate Y	$(X - \overline{X}) = x$	$(Y - \overline{Y}) = y$	$(X - \overline{X})^2 = x^2$	$(Y - \overline{Y})^2 = y^2$	$(X - \overline{X})(Y - \overline{Y}) = xy$
24	15	–6	–7	36	49	42
26	20	–4	–2	16	4	8
32	22	2	0	4	0	0
33	24	3	2	9	4	6
35	27	5	5	25	25	25
30	24.	0	2	0	4	0
$\Sigma X = 180$ $\overline{X} = \frac{180}{6} = 30$	$\Sigma Y = 132$ $\overline{Y} = \frac{132}{6} = 22$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 90$	$\Sigma y^2 = 86$	$\Sigma xy = 81$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{(81)}{\sqrt{90 \times 86}} = \frac{81}{\sqrt{7740}} = \frac{81}{87.98} = \mathbf{.92}$$



B. Short Cut Method

In case the mean is a whole number above method is simple. But when the mean is in fractions, short-cut method is used. In this method, the deviations are calculated from assumed mean and the following formula is applied.

$$\boxed{r = \frac{N\,\Sigma\,dxdy - \Sigma\,dx\,\Sigma\,dy}{\sqrt{N\,\Sigma\,dx^2 - (\Sigma\,dx)^2}\,\sqrt{N\,\Sigma\,dy^2 - (\Sigma\,dy)^2}}}$$

Where

$\Sigma\,dx$ =Sum of deviations of X series from its Assumed Mean *i.e.* $\Sigma\,(X-A_x)$

$\Sigma\,dy$ =Sum of deviations of Y series from its assumed Mean *i.e.* $\Sigma\,(Y-A_y)$

$\Sigma\,dx^2$ =Sum of squared deviations of X Series from its Assumed Mean *i.e.* $\Sigma\,(X-A_x)^2$

$\Sigma\,dy^2$ =Sum of squared deviations of Y Series from its AssumedMean *i.e.* $\Sigma\,(Y-A_y)^2$

$\Sigma\,dxdy$ =Sum of products of deviations of X and Y series from their respective assumed means. $\Sigma\,dxdy = \Sigma\,(X-A_x)\,(Y-A_y)$

N =Number of pairs

Example 3. Calculate coefficient of correlation between X series and Y series using Karl Pearson's Method.

X :	14	12	14	16	16	17	16	15
Y :	13	11	10	15	15	9	14	17

Solution : $\Sigma dx = 0;\ \Sigma dx^2 = 18;\ \Sigma dy = -8;\ \Sigma dy^2 = 62;\ \Sigma dxdy = 6$

X	Y	dx $=X-A_x$	dx^2	dy = $Y-A_y$	dy^2	dxdy
14	13	–1	1	–1	1	1
12	11	–3	9	–3	9	9
14	10	–1	1	–4	16	4
16	15	1	1	1	1	1
16	15	1	1	1	1	1
17	9	2	4	–5	25	–10
16	14	1	1	0	0	0
15	17	0	0	3	9	0
	1	0	18	–8	62	6

Let $A_x = 15$ and $A_y = 14$

$$r = \frac{N\Sigma dxdy - \Sigma dx\Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{8 \times 6 - (0) \times (-8)}{\sqrt{8 \times 18 - (0)^2} \times \sqrt{8 \times 62\,(-8)^2}}$$

$$= \frac{40 - 0}{\sqrt{(144-0)} \times \sqrt{(496-64)}} = \frac{48}{\sqrt{144 \times 432}} = \frac{48}{\sqrt{62208}} = \frac{48}{249.41} = .192 = .19$$

Imp. Note : Above formula can also be given the shape as following

$$r = \frac{\Sigma dxdy - \dfrac{\Sigma dx\,.\,\Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \dfrac{(\Sigma dx)^2}{N}}\ \sqrt{\Sigma dy^2 - \dfrac{(\Sigma dy)^2}{N}}}$$

Spearman Rank Correlation

This implies a negative correlation between the considered variables i.e. The higher the number of cigarettes smoked per week in last 5 years, the lesser the number of years lived. Note that it DOES NOT mean that smoking cigarettes decreases the life span. Because, many other factors might be responsible for one’s death. Still, it is an important conclusion nevertheless.

Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. The Spearman rank correlation test does not carry any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson’s correlation assesses linear relationships, Spearman’s correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.

The following formula is used to calculate the Spearman rank correlation:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

ρ= Spearman rank correlation

di= the difference between the ranks of corresponding variables

n= number of observations

Merits

- It is easy to calculate.
- It is simple to understand.
- It can be applied to any type of data. Qualitative or Quantitative. Hence correlation with qualitative data such as honesty, beauty can be found.
- This is most suitable in case there are two attributes.

Demerits

- It is only an approximately calculate measure as actual values are not used for calculations.
- For large samples, it is not a convenient method.
- Combined r of different series cannot be obtained as in case of mean and S.D.
- It cannot be treated further algebraically.

2. Pearson correlation coefficient

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by ‘r’. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data