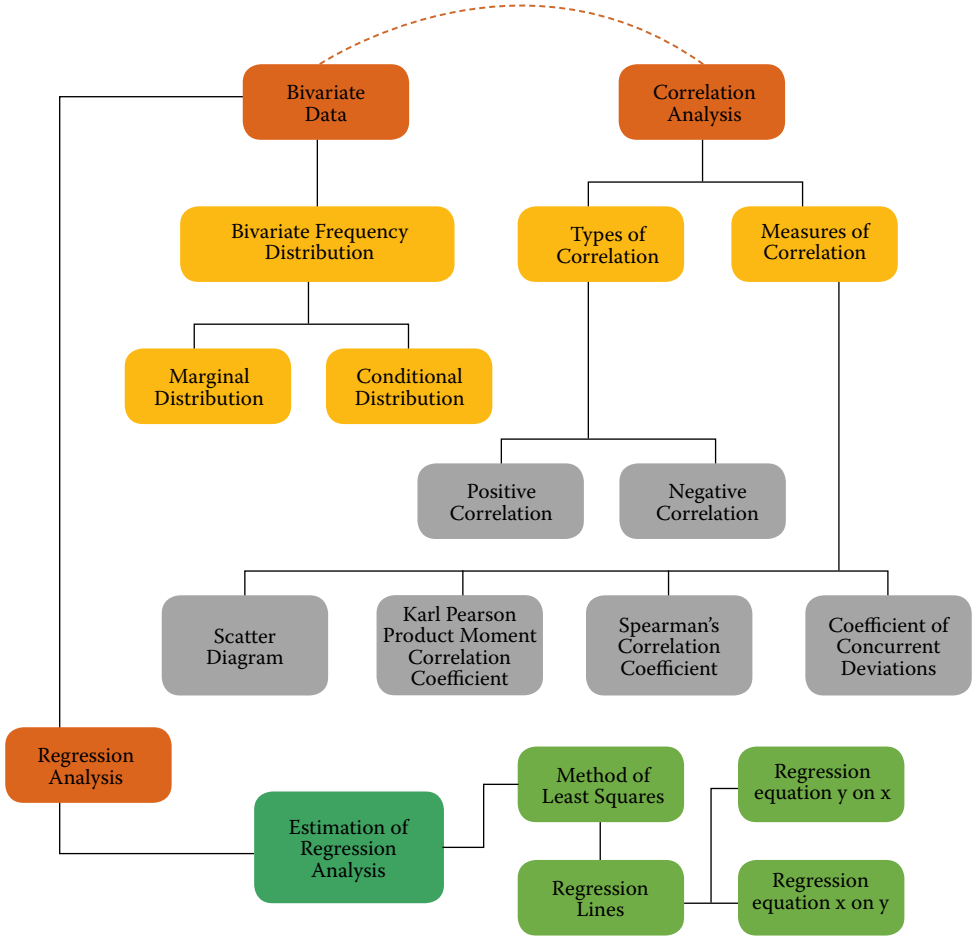
CA FOUNDATION - PAPER 3 - BUSINESS MATHEMATICS, LOGICAL REASONING AND STATISTICS

At the foundation level with regards to Paper 3: Business Mathematics, Logical Reasoning and Statistics, Chapter 17: Correlation and Regression is very important for students not only to acquire professional knowledge but also for examination point of view. Here in this capsule an attempt is made for solving and understanding the concepts of Correlation and Regression.

CHAPTER 17 OVERVIEW: CORRELATION AND REGRESSION



Univariate Distribution: Statistical measure relating to Univariate distribution i.e. distribution of one variable like height, weight, mark, profit, wage and so on. However, there are situations that demand study of more than one variable simultaneously. A businessman may be keen to know what amount of investment would yield a desired level of profit or a student may want to know whether performing better in the selection test would enhance his or her chance of doing well in the final examination. With a view to answering this series of questions, we need to study more than one variable at the same time.

Bivariate Data: When data are collected on two variables simultaneously, they are known as bivariate data and the corresponding frequency distribution, derived from it, is known as Bivariate Frequency Distribution. If x and y denote marks in Maths and Stats for a group of 30 students, then the corresponding bivariate data would be (x_i, y_i) for i= 1, 2, ..., 30 where (x_1, y_1) denotes the marks in Mathematics and Statistics for the student with serial number or Roll Number 1, (x_2, y_2), that for the student with Roll Number 2 and so on and lastly (x_{30}, y_{30}) denotes the pair of marks for the student bearing Roll Number 30.

Correlation Analysis and Regression Analysis are the two analyses that are made from a multivariate distribution i.e. a distribution of more than one variable. In particular, when there are two variables, say x and y, we study bivariate distribution. We restrict our discussion to bivariate distribution only.

Correlation analysis helps us to find an association or the lack of it between the two variables x and y. Thus, if x and y stand for profit and investment of a firm or the marks in Statistics and Mathematics for a group of students, then we may be interested to know whether x and y are associated or independent of each other. The extent or amount of correlation between x and y is provided by different measures of Correlation namely Product Moment Correlation Coefficient or Rank Correlation Coefficient or Coefficient of Concurrent Deviations. In Correlation analysis, we must be careful about a cause-and-effect relation between the variables under consideration because there may be situations where x and y are related due to the influence of a third variable although no causal relationship exists between the two variables.

Regression analysis, on the other hand, is concerned with predicting the value of the dependent variable corresponding to a known value of the independent variable on the assumption of a mathematical relationship between the two variables and also an average relationship between them.

As in the case of a Univariate Distribution, we need to construct the frequency distribution for bivariate data. Such a distribution takes into account the classification in respect of both the variables simultaneously. Usually, we make horizontal classification in respect of x and vertical classification in respect of the other variable y. Such a distribution is known as Bivariate Frequency Distribution or Joint Frequency Distribution or Two way classification of the two variables x and y. Frequency Distribution, we can obtain two types of univariate distributions which are known as:

(a) Marginal distribution: Marginal distributions always divide the column or row totals by the table total

(b) Conditional distribution: To calculate a conditional distribution, you must first establish a condition. For instance, we could ask what the distribution of gender is among students who watched the last football game. So, the condition here would be that the student watched the game. In particular, if there are m classifications for x and n classifications for y, then there would be altogether (m + n) conditional distribution.

Correlation Analysis: While studying two variables at the same time, if it is found that the change in one variable is reciprocated by a corresponding change in the other variable either directly or inversely, then the two variables are known to be associated or correlated. Otherwise, the two variables are known to be dissociated or uncorrelated or independent. There are two types of correlation.

Positive correlation	Negative correlation	The two variables are known to be uncorrelated if the movement on the part of one variable does not produce any movement of the other variable in a particular direction. As for example, Shoe-size and intelligence are uncorrelated
If two variables move in the same direction i.e. an increase (or decrease) on the part of one variable introduces an increase (or decrease) on the part of the other variable, then the two variables are known to be positively correlated.	If the two variables move in the opposite directions i.e. an increase (or a decrease) on the part of one variable results a decrease (or an increase) on the part of the other variable, then the two variables are known to have a negative correlation.	
For example, height and weight yield and rainfall, profit and investment etc. are positively correlated.	The price and demand of an item, the profits of Insurance Company and the number of claims it has to meet etc. are examples of variables having a negative correlation.	

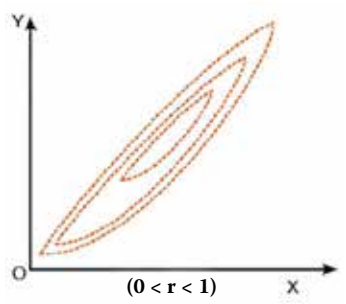
Measures of correlation:

- Scatter diagram
- Karl Pearson's Product moment correlation coefficient
- Spearman's rank correlation coefficient
- Coefficient of concurrent deviations

(a) SCATTER DIAGRAM: This is a simple diagrammatic method to establish correlation between a pair of variables. Unlike product moment correlation coefficient, which can measure correlation only when the variables are having a linear relationship, scatter diagram can be applied for any type of correlation – linear as well as non-linear i.e. curvilinear. Scatter diagram can distinguish between different types of correlation although it fails to measure the extent of relationship between the variables.

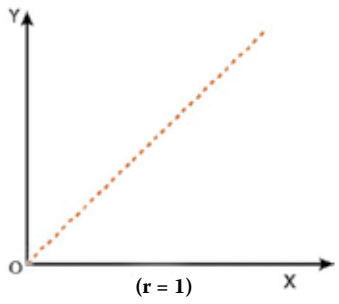
Each data point, which in this case a pair of values (x_i, y_i) is represented by a point in the rectangular axes of cordinates. The totality of all the plotted points forms the scatter diagram. The pattern of the plotted points reveals the nature of correlation. In case of a positive correlation, the plotted points lie from lower left corner to upper right corner, in case of a negative correlation the plotted points concentrate from upper left to lower right and in case of zero correlation, the plotted points would be equally distributed without depicting any particular pattern. The following figures show different types of correlation and the one-to-one correspondence between scatter diagram and product moment correlation coefficient.

Figure 1: Showing Positive Correlation



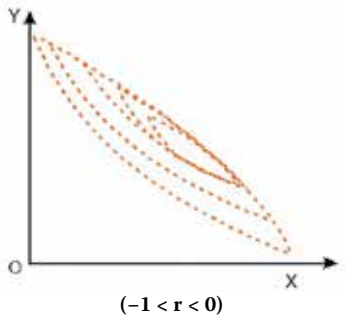
(0 < r < 1)

Figure 2: Showing Perfect Correlation



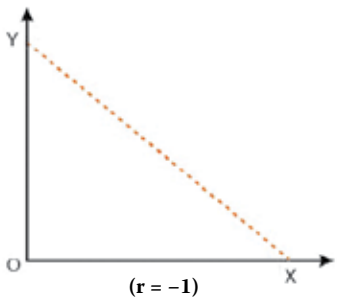
(r = 1)

Figure 3: Showing Negative Correlation



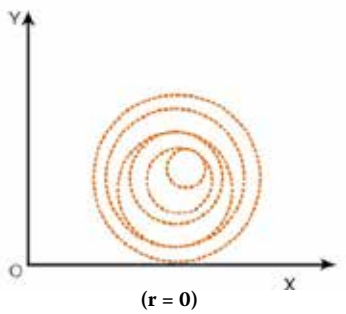
(–1 < r < 0)

Figure 4: Showing Perfect Negative Correlation



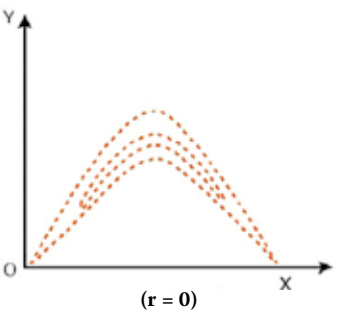
(r = –1)

Figure 5: Showing No Correlation



(r = 0)

Figure 6: Showing Curvilinear Correlation



(r = 0)

(b) KARL PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT

This is by far the best method for finding correlation between two variables provided the relationship between the two variables is linear. Pearson's correlation coefficient may be defined as the ratio of covariance between the two variables to the product of the standard deviations of the two variables. If the two variables are denoted by x and y and if the corresponding bivariate data are (x_i, y_i) for i = 1, 2, 3, ..., n, then the coefficient of correlation between x and y, due to Karl Pearson, is given by

$$r = r_{xy} = \frac{Cov(x,y)}{S_x \times S_y}$$

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

$$S_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2} \text{ and } S_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} = \sqrt{\frac{\sum y_i^2}{n} - \bar{y}^2}$$

A single formula for computing correlation coefficient is given by

$$r = \frac{n\sum x_i y_i - \sum x_i \times \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

In case of a bivariate frequency distribution, we have

$$\text{Cov}(x,y)= \frac{\sum_{i,j} x_i y_i f_{ij}}{N} - \bar{x} \times \bar{y}$$

$$S_x= \sqrt{\frac{\sum_i f_{io} x_i^2}{N} - \bar{x}^2} \text{ and } S_y= \sqrt{\frac{\sum_i f_{oj} y_j^2}{N} - \bar{y}^2}$$

where x_i = Mid-value of the i^{th} class interval of x. ,y_j =Mid-value of the j^{th} class interval of y
f_{io} = Marginal frequency of x f_{oj} =Marginal frequency of y
f_{ij} = frequency of the $(i, j)^{th}$ cell , N = Total frequency

PROPERTIES OF CORRELATION COEFFICIENT

(i) **The Coefficient of Correlation is a unit-free measure:** This means that if x denotes height of a group of students expressed in cm and y denotes their weight expressed in kg, then the correlation coefficient between height and weight would be free from any unit.

(ii) **The coefficient of correlation remains invariant under a change of origin and/or scale of the variables under consideration depending on the sign of scale factors.**

This property states that if the original pair of variables x and y is changed to a new pair of variables u and v by effecting a change of origin and scale for both x and y i.e.

$$u = \frac{x-a}{b} \text{ and } v = \frac{y-c}{d}$$

where a and c are the origins of x and y and b and d are the respective scales and then we have

$$rxy = \frac{bd}{|b||d|} r_{uv}.$$

r_{xy} and r_{uv} being the coefficient of correlation between x and y and u and v respectively, the two correlation coefficients remain equal and they would have opposite signs only when b and d, the two scales, differ in sign.

(iii) **The coefficient of correlation always lies between –1 and 1, including both the limiting values i.e.**
$-1 \le r \le 1$

(c) SPEARMAN'S RANK CORRELATION COEFFICIENT:

When we need finding correlation between two qualitative characteristics, say, beauty and intelligence, we take recourse to using rank correlation coefficient. Rank correlation can also be applied to find the level of agreement (or disagreement) between two judges so far as assessing a qualitative characteristic is concerned. As compared to product moment correlation coefficient, rank correlation coefficient is easier to compute, it can also be advocated to get a first hand impression about the correlation between a pair of variables.

Spearman's rank correlation coefficient is given by

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

where r_R denotes rank correlation coefficient and it lies between –1 and 1 inclusive of these two values.

$d_i = x_i - y_i$ represents the difference in ranks for the i-th individual and n denotes the number of individuals.

In case u individuals receive the same rank, we describe it as a tied rank of length u. In case of a tied rank, formula is changed to

$$r_R = 1 - \frac{6\left[\sum_i d_i + \sum_j \frac{(t_j^3 - t_j)}{12}\right]}{n(n^2-1)}$$

In this formula, t_j represents the j^{th} tie length and the summation extends over the lengths of all the ties for both the series.

(d) COEFFICIENT OF CONCURRENT DEVIATIONS

A very simple and casual method of finding correlation when we are not serious about the magnitude of the two variables is the application of concurrent deviations. This method involves in attaching a positive sign for a x-value (except the first) if this value is more than the previous value and assigning a negative value if this value is less than the previous value. This is done for the y-series as well. The deviation in the x-value and the corresponding y-value is known to be concurrent if both the deviations have the same sign.

Denoting the number of concurrent deviation by c and total number of deviations as m (which must be one less than the number of pairs of x and y values), the coefficient of concurrent deviation is given by

$$r_C = \pm\sqrt{\pm\frac{(2c-m)}{m}}.$$

If (2c–m) >0, then we take the positive sign both inside and outside the radical sign and if (2c–m) <0, we are to consider the negative sign both inside and outside the radical sign.

Like Pearson's correlation coefficient and Spearman's rank correlation coefficient, the coefficient of concurrent deviations also lies between –1 and 1, both inclusive.

Spurious Correlation: There are some cases when we may find a correlation between two variables although the two variables are not causally related. This is due to the existence of a third variable which is related to both the variables under consideration. Such a correlation is known as spurious correlation or non-sense correlation. As an example, there could be a positive correlation between production of rice and that of iron in India for the last twenty years due to the effect of a third variable time on both these variables. It is necessary to eliminate the influence of the third variable before computing correlation between the two original variables.

Correlation Coefficient: Correlation Coefficient measuring a linear relationship between the two variables indicates the amount of variation of one variable accounted for by the other variable. A better measure for this purpose is provided by the square of the correlation coefficient, known as 'coefficient of determination'. This can be interpreted as the ratio between the explained variance to total variance i.e.

$$r^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

Thus, a value of 0.6 for r indicates that $(0.6)^2 \times 100\%$ or 36 per cent of the variation has been accounted for by the factor under consideration and the remaining 64 per cent variation is due to other factors.

Coefficient of non-determination: The 'coefficient of non-determination' is given by $(1-r^2)$ and can be interpreted as the ratio of unexplained variance to the total variance.

Coefficient of non-determination = $(1-r^2)$

CA FOUNDATION - PAPER 3 - BUSINESS MATHEMATICS, LOGICAL REASONING AND STATISTICS

This capsule is in continuation to the previous edition featured in June 2022. Further here presented properties of Regression and their applications. Here an attempt is made to the enable the students to understand the Correlation and Regression with the help of examples.

CHAPTER 17: CORRELATION AND REGRESSION - PART 2

Regression Lines: (i) The two lines of regression coincide i.e. become identical when r = –1 or 1 or in other words, there is a perfect negative or positive correlation between the two variables. (ii) If r = 0, Regression lines are perpendicular to each other.

Regression Analysis: In regression analysis, we are concerned with the estimation of one variable for a given value of another variable (or for a given set of values of a number of variables) on the basis of an average mathematical relationship between the two variables (or a number of variables). Regression analysis plays a very important role in the field of every human activity. A businessman may be keen to know what would be his estimated profit for a given level of investment on the basis of the past records. Similarly, an outgoing student may like to know her chance of getting a first class in the final University Examination on the basis of her performance in the college selection test.

When there are two variables x and y and if y is influenced by x i.e., if y depends on x, then we get a simple linear regression or simple regression. y is known as dependent variable or regression or explained variable and x is known as independent variable or predictor or explanator. In the previous examples since profit depends on investment or performance in the University Examination is dependent on the performance in the college selection test, profit or performance in the University Examination is the dependent variable and investment or performance in the selection test is the independent variable.

In case of a simple regression model if y depends on x, then the regression line of y on x in given by

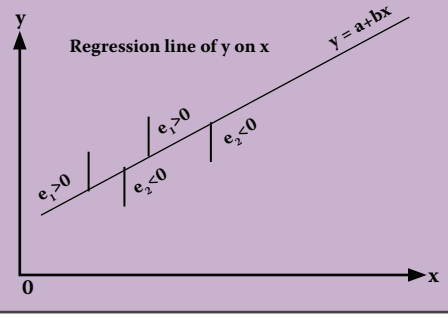
$$y = a + bx$$

Here, a and b are two constants and they are also known as regression parameters. Furthermore, b is also known as the regression coefficient of y on x and is also denoted by b_{yx}. We may define the regression line of y on x as the line of best fit obtained by the method of least squares and used for estimating the value of the dependent variable y for a known value of the independent variable x.

The method of least squares involves in minimising

$$\Sigma e_i^2 = \Sigma (y_i^2 - y\hat{}_i)^2 = \Sigma(y_i - a - bx_i)^2$$

where y_i denotes the actual or observed value and $y\hat{}_i = a + b_{xi}$, the estimated value of y_i for a given value of x_i, e_i is the difference between the observed value and the estimated value and ei is technically known as error or residue. This summation intends over n pairs of observations of (x_i, y_i). The line of regression of y or x and the errors of estimation are shown in the following figure.



SHOWING REGRESSION LINE OF y on x AND ERRORS OF ESTIMATION

Minimisation of the equation yields the following equations known as 'Normal Equations'

$$\Sigma y_i = na + b\Sigma x_i$$
$$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2$$

Solving there, two equations for b and a, we have the "least squares" estimates of b and a as

$$b = \frac{cov(x_i y_i)}{S_x^2} = \frac{r.S_x.S_y}{S_x^2}$$
$$= r.\frac{S_y}{S_x}$$

After estimating b, estimate of a is given by

$$a = \bar{y} - b\bar{x}$$

Substituting the estimates of b and a in equation, we get

$$\frac{(y-\bar{y})}{S_y} = \frac{r(x-\bar{x})}{S_x}$$

There may be cases when the variable x depends on y and we may take the regression line of x on y as

$$x = a\hat{} + b\hat{}y$$

Unlike the minimisation of vertical distances in the scatter diagram as shown in figure for obtaining the estimates of a and b, in this case we minimise the horizontal distances and get the following normal equation in a^ and b^, the two regression parameters:

$$\Sigma x_i = na\hat{} + b\hat{}\Sigma y_i$$
$$\Sigma x_i y_i = a\hat{}\Sigma y_i + b\hat{}\Sigma y_i^2$$

or solving these equations, we get

$$b\hat{} = b_{xy} = \frac{cov(x,y)}{S_y^2} = \frac{r.S_x}{S_y}$$
$$a\hat{} = \bar{x} - b\hat{}\,\bar{y}$$

A single formula for estimating b is given by

$$b\hat{} = b_{yx} = \frac{n\sum x_i y_i - \sum x_i.\sum y_i}{n\sum x^2 - \left(\sum x\right)^2}.$$

Similarly, $b\hat{} = b_{xy} = \dfrac{n\sum x_i y_i - \sum x_i.\sum y_i}{n\sum y_i^2 - \left(\sum y_i\right)^2}$

The standardised form of the regression equation of x on y is given by

$$\frac{x-\bar{x}}{S_x} = r\frac{\left(y-\bar{y}\right)}{S_y}$$

PROPERTIES of Regression lines: We consider the following important properties of regression lines:

(i) **The regression coefficients remain unchanged due to a shift of origin but change due to a shift of scale.**

This property states that if the original pair of variables is (x, y) and if they are changed to the pair (u, v) where

$$u = \frac{x-a}{p} \text{ and } v = \frac{y-c}{q}$$

$$b_{xy} = \frac{p}{q} \times b_{vu} \text{ and } b_{yx} = \frac{q}{p} \times b_{uv}$$

(ii) **The two lines of regression intersect at the point, where x and y are the variables under consideration.**

According to this property, the point of intersection of the regression line of y on x and the regression line of x on y is the solution of the simultaneous equations in x and y.

(iii) **The coefficient of correlation between two variables x and y in the simple geometric mean of the two regression coefficients. The sign of the correlation coefficient would be the common sign of the two regression coefficients.**

This property says that if the two regression coefficients are denoted by b_{yx} (=b) and b_{xy} (=b') then the coefficient of correlation is given by

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

If both the regression coefficients are negative, r would be negative and if both are positive, r would assume a positive value.

1. If for two variables x and y, the covariance, variance of x and variance of y are 40, 16 and 256 respectively, what is the value of the correlation coefficient?

Solution:

Cov(x,y)=30, V(x)=25, V(y)=144

As we know formula of Corelation coefficient is:

Let r be Corelation coefficient of x,y

$$r = \frac{Cov\,(x,y)}{\sqrt{V\,(x)\ x\ V(y)}} = \frac{30}{\sqrt{25\ x\ 144}} = \frac{30}{5\ x\ 12} = 0.5$$

$\Rightarrow$r=0.5

2. If the covariance between two variables is 20 and the variance of one of the variables is 16, what would be the variance of the other variable?

Solution: Given, Cov (x, y) =20 and variance of one of the variables is 16.
so the standard deviation SD is 4.
we know the formula,
r= cov (x, y) /(SD of x * SD of Y)
r= 20 /4 * SD of the other variable
r= 5/ SD of the other variable
we also know that coefficient of correlation, r, lies between –1 and +1 including them.
so, SD of the other variable has to be atleast 5 or more.
so the variance will be $5^2 = 25$ atleast or more. $S^2 \geq 25$

3. If r = 0.6, then the coefficient of non-determination is
Solution: Given r = 0.6
The coefficient of non-determination = $1 - r^2 = 1 - 0.36 = 0.64$

4. If u+5x=6 and 3y+7v=20 and correlation coefficient between x and y is 0.58, then what is correlation coefficient between u and v ?

Solution: Correlation coefficient between x and y is 0.58
u + 5x = 6
=> u = 6 – 5x
–5 is the factor (constant does not have any impact)
3y + 7v = 20
=> 7v = –3y + 20
=> v = (–3/7) y + 20/7
(–3/7) is the factor (constant does not have any impact)
correlation coefficient between u and v = 0.58 x (–5)(–3/7) / $\sqrt{(-5)^2}\sqrt{(-3/7)^2}$
= 0.58
correlation coefficient between u and v = 0.58

5. If the relation between x and u is 3x + 4u + 7 = 0 and the correlation coefficient between x and y is –0.6, then what is the correlation coefficient between u and y?

Solution: Given x and u is 3x + 4u + 7 = 0

$$\therefore\ u = \frac{-3x-7}{4}$$

We can write this as

$$u = \left(-\frac{3}{4}\right)x - \left(\frac{7}{4}\right)$$

Therefore, perfect negative correlation between x and y and that is –0.6

$$= \frac{-(0.6) \times \left(-\frac{3}{4}\right)}{\left(\frac{3}{4}\right)} = 0.6$$

So, the correlation between u and y is 0.6

6. If the sum of squares of difference of ranks, given by two judges A and B of 8 students is 21, what is the value of rank correlation coefficient?

Solution: Here, n = 8.
and '$\Sigma\ (d^2)$' = 21.
Now, n x $\{(n^2) - 1\}$ = 8 x 63 = 504. So, rank correlation coefficient
= 1 – (6 x 21 / 504)
= 1 – 0.25 = 0.75

7. If the rank correlation coefficient between marks in management and mathematics for a group of student is 0.6 and the sum of squares of the differences in ranks is 66, what is the number of students in the group?

Solution: Rank correlation coefficient = $1 - 6\ \Sigma\ (di)^2 / (n(n^2 - 1))$
n = number of students of group
$\Sigma\ (di)^2$ = sum of squares of the differences in ranks = 66
rank correlation coefficient = 0.6
$=> 0.6 = 1 - 6\ \Sigma\ (di)^2 / (n(n^2 - 1))$
$=> -0.4 = -6\ x\ 66 / (n(n^2 - 1))$
$=> (n(n^2 - 1)) = 990$
$=> (n(n^2 - 1)) = 10\ x\ 99$
$=> (n(n^2 - 1)) = 10\ x\ (100 - 1)$
$=> (n(n^2 - 1)) = 10\ x\ (10^2 - 1)$
=> n = 10
Therefore, the number of students of group = 10.

8. While computing rank correlation coefficient between profit and investment for the last 6 years of a company the difference in rank for a year was taken 3 instead of 4. What is the rectified rank correlation coefficient if it is known that the original value of rank correlation coefficient was 0.4?

Solution: rank correlation coefficient $= 1 - 6\ \Sigma (di)^2 / (n(n^2 - 1))$

n = number of years = 6

$0.4 = 1 - 6\ \Sigma (di)^2 / (6(6^2 - 1))$

$\Rightarrow 0.6 = \Sigma (di)^2 / 35$

$\Rightarrow \Sigma (di)^2 = 21$

rank for a year was taken 3 instead of 4.

$\Rightarrow$ Actual $\Sigma (di)^2 = 21 - (3)^2 + 4^2 = 28$

Actual rank correlation coefficient $= 1 - 6 \times 28 / (n(n^2 - 1))$

$= 1 - 28/35 = 7/35 = 1/5 = 0.2$

9. For 10 pairs of observations, No. of concurrent deviations was found to be 4. What is the value of the coefficient of concurrent deviation?

Solution: Step-by-step explanation:

The Formula for Coefficient of Concurrent deviation is:

$$R = \pm \sqrt{\left|\frac{2c-m}{m}\right|}$$

here, m = n − 1 (and n is the total number of observation) = 10 − 1 = 9

c = Number of pair of concurrent deviation = 4

Substituting all values in formula,

$$R = \pm \sqrt{\left|\frac{2x4-9}{9}\right|}$$

Also, 2×4−9 < 0 so take negative value

$\Rightarrow R = \frac{-1}{3}$

10. For 10 pairs of observations, No. of concurrent deviations was found to be 4. What is the value of the coefficient of concurrent deviation?

Solution: Given coefficient of concurrent deviation is given by $R = \pm\sqrt{2}\ c - m / m$

Now m = p − 1 and c = 6 and R = $1/\sqrt{3}$

So $1/\sqrt{3} = \sqrt{2} \times 6 - (p - 1) / p - 1$

$1/\sqrt{3} = \sqrt{12} - p + 1 / p - 1$

Squaring both sides, we get

$1/3 = 13 - p / p - 1$

Or $p - 1 = 39 - 3p$

$4p = 40$ Or $p = 40 / 4$ Or $p = 10$ pairs

11. Following are the two normal equations obtained for deriving the regression line of y and x:
5a + 10b = 40, 10a + 25b = 95. The regression line of y on x is given by

Solution:

The normal equations obtained for deriving the regression line of y and x: 5a + 10b = 40 and 10a + 25b = 95.

In order to find the regression line of y on x, we need to solve the 5a + 10b = 40 and 10a + 25b = 95 to find out the values of a and b. So, we have,

Given,

5a + 10b = 40 .....(1)

10a + 25b = 95 .....(2)

2 × (1) gives,

10a + 20b = 80 .....(3)

(2) − (3) gives,

10a + 25b = 95

10a + 20b = 80

5b = 15, **b = 3** from (1)

5a + 10b = 40

5a + 10 (3) = 40

5a + 30 = 40

5a = 40 − 30

5a = 10

a = 2

As, "a" represents the y-intercept and "b" represents the slope, so we have,

The regression line of y on x is given by **y = 2x + 3**

12. Given the regression equations as 2x+3y = 6 and 5x+7y =12, then which one is regression equation x on y?

Solution: For regression equations, both coefficients both b_{xy} and b_{yx} need to be of the same sign.

If 2x+3y = 6 is y on x and 5x+7y−12 = 0 is x on y

then y=6/3−2/3x then $x = \frac{12 - 7y}{5}$

$b_{yx}1 = -2/3$ and $b_{xy}1 = -7/5$

If 2x+3y = 6 is x on y and 5x+7y = 12 is y on x

Then $x = \frac{6 - 3y}{2}$ and $y = \frac{12 - 5x}{7}$

$b_{xy}2 = -3/2$ and $b_{yx}2 = -5/7$

$b_{xy}1 \times b_{yx}1 = 14/15 < 1$

$b_{yx}2 \times b_{xy}2 = 15/14 > 1$

We know $b_{xy} \times b_{yx} = r^2$ which is always <1

So first assumption is correct.

Regression equation of y on x is 2x+3y = 6

Regression equation of x on y is 5x+7y−12 = 0

13. If y = 3x + 4 is the regression line of y on x and the arithmetic mean of x is −1, what is the arithmetic mean of y?

Solution: Given regression line y on x is y = 3x + 4

Given $\bar{y} = 3\bar{x} + 4 = 3(-1) + 4 = 1$

Therefore, arithmetic mean of y =1

14. Compute the correlation coefficient between x and y from the following data n = 10, $\Sigma xy = 220$, $\Sigma x^2 = 200$, $\Sigma y^2 = 262$
$\Sigma x = 40$ and $\Sigma y = 50$

Solution: From the given data, we have by applying

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

$$= \frac{10\times 220 - 40\times 50}{\sqrt{10\times 200 - (40)^2}\times\sqrt{10\times 262 - (50)^2}}$$

$$= \frac{2200 - 2000}{\sqrt{2000 - 1600}\times\sqrt{2620 - 2500}}$$

$$= \frac{200}{20\times 10.9545} = 0.91$$

Thus, there is a good amount of positive correlation between the two variables x and y.

Alternately

As given, $\bar{x} = \frac{\sum x}{n} = \frac{40}{10} = 4$, $\bar{y} = \frac{\sum y}{n} = \frac{50}{10} = 5$

Cov (x, y) $= \frac{\sum xy}{n} - \bar{x}\bar{y} = \frac{220}{10} - 4\times 5 = 22 - 20 = 2$

$S_x = \sqrt{\frac{\sum x^2}{n} - (\bar{x})^2} = \sqrt{\frac{200}{10} - 4^2} = 2$

$S_y = \sqrt{\frac{\sum y_i^2}{n} - \bar{y}^2} = \sqrt{\frac{262}{10} - 5^2} = \sqrt{26.20 - 25} = 1.0954$

Thus, applying formula, we get

$$r = \frac{\text{cov}(x,y)}{S_x . S_y} = \frac{2}{2\times 1.0954} = 0.91$$

As before, we draw the same conclusion.

15. For a group of 10 students, the sum of squares of differences in ranks for Mathematics and Statistics marks was found to be 50. What is the value of rank correlation coefficient?

Solution: As given n = 10 and $\sum d_i^2 = 50$. Hence the rank correlation coefficient between marks in Mathematics and Statistics is given by

$$r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6\times 50}{10(10^2-1)} = 1 - 0.30 = 0.70$$

16. For a number of towns, the coefficient of rank correlation between the people living below the poverty line and increase of population is 0.50. If the sum of squares of the differences in ranks awarded to these factors is 82.50, find the number of towns.

Solution: As given $r_R = 0.50$, $\sum d_i^2 = 82.50$.

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

$0.50 = 1 - \frac{6\times 82.50}{n(n^2-1)}$

$= n(n^2 - 1) = 990$

$= n(n^2 - 1) = 10(10^2 - 1)$

Therefore n = 10 as n must be a positive integer.

17. While computing rank correlation coefficient between profits and investment for 10 years of a firm, the difference in rank for a year was taken as 7 instead of 5 by mistake and the value of rank correlation coefficient was computed as 0.80. What would be the correct value of rank correlation coefficient after rectifying the mistake?

Solution: We are given that n = 10,

r_R = 0.80 and the wrong $d_i = 7$ should be replaced by 5.

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

$$0.80 = 1 - \frac{6\sum d_i^2}{10(10^2-1)}$$

$$\sum d_i^2 = 33$$

Corrected $\sum d_i^2 = 33 - 7^2 + 5^2 = 9$

Hence rectified value of rank correlation coefficient $= 1 - \frac{6\times 9}{10(10^2-1)}$ = 0.95

18. Find product moment correlation coefficient from the following information:

x	:	2	3	5	5	6	8
y	:	9	8	8	6	5	3

Solution: In order to find the covariance and the two-standard deviation, we prepare the following table:

Table: Computation of Correlation Coefficient

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
(1)	(2)	(3)= (1) x (2)	(4)= $(1)^2$	(5)= $(2)^2$
2	9	18	4	81
3	8	24	9	64
5	8	40	25	64
5	6	30	25	36
6	5	30	36	25
8	3	24	64	9
29	39	166	163	279

We have

$\bar{x} = \frac{29}{6} = 4.8333$ $\bar{y} = \frac{39}{6} = 6.50$

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

$= 166/6 - 4.8333 \times 6.50 = -3.7498$

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

$= \sqrt{\frac{163}{6} - (4.8333)^2}$

$= \sqrt{27.1667 - 23.3608} = 1.95$

$S_y = \sqrt{\frac{\sum y_i^2}{n} - (\bar{y})^2}$

$= \sqrt{\frac{279}{6} - (6.50)^2}$

$= \sqrt{46.50 - 42.25} = 2.0616$

Thus the correlation coefficient between x and y is given by

$$r = \frac{\text{cov}(x,y)}{S_x . S_y} = \frac{-3.7498}{1.9509\times 2.0616} = -0.93$$

We find a high degree of negative correlation between x and y.

19. If y is independent variable and x is independent variable and SD of x and y are 5 and 8 respectively and coefficient of correlation between x and y is 0.8. Find the regression coefficient y on x

Solution: SD of x (σ_x) = 5, SD of y (σ_y) = 8,

Coefficient of correlation (r) = 0.8

Regression coefficient y on x = $b_{yx} = r \times \frac{\sigma_y}{\sigma_x} = \frac{0.8\times 8}{5} = \frac{6.4}{5} = 1.28$

CA FOUNDATION - PAPER 3 - BUSINESS MATHEMATICS, LOGICAL REASONING AND STATISTICS

This capsule is in continuation to the previous edition featured in November 2022. Here, an attempt is made to enable the students to understand the Correlation and Regression with the help of examples.

CHAPTER 17: CORRELATION AND REGRESSION - PART 3

1. Given that the correlation coefficient between x and y is 0.8, write down the correlation coefficient between u and v where

(i) $2u + 3x + 4 = 0$ and $4v + 16x + 11 = 0$

(ii) $2u - 3x + 4 = 0$ and $4v + 16x + 11 = 0$

(iii) $2u - 3x + 4 = 0$ and $4v - 16x + 11 = 0$

(iv) $2u + 3x + 4 = 0$ and $4v - 16x + 11 = 0$

Solution: Using formula, we find that

$$r_{xy} = \frac{bd}{|b||d|} r_{uv}$$

i.e., $r_{xy} = r_{uv}$ if b and d are of same sign and $r_{uv} = -r_{xy}$ when b and d are of opposite signs, b and d being the scales of x and y respectively. In (i), $u = (-2) + (-3/2)\,x$ and $v = (-11/4) + (-4)y$.

Since $b = -3/2$ and $d = -4$ are of same sign, the correlation coefficient between u and v would be the same as that between x and y, i.e., $r_{xy} = 0.8 = r_{uv}$

In (ii), $u = (-2) + (3/2)x$ and $v = (-11/4) + (-4)y$. Hence $b = 3/2$ and $d = -4$ are of opposite signs and we have $r_{uv} = -r_{xy} = -0.8$

Proceeding in a similar manner, we have $r_{uv} = 0.8$ and -0.8 in (iii) and (iv).

2. For the variables x and y, the regression equations are given as $7x - 3y - 18 = 0$ and $4x - y - 11 = 0$

(i) Find the arithmetic means of x and y.

(ii) Identify the regression equation of y on x.

(iii) Compute the correlation coefficient between x and y.

(iv) Given the variance of x is 9, find the SD of y.

Solution:

(i) Since the two lines of regression intersect at the point $(\bar{x}, \bar{y})$ replacing x and y by $\bar{x}$ and $\bar{y}$ respectively in the given regression equations, we get

$$7\bar{x} - 3\bar{y} - 18 = 0$$

and $$4\bar{x} - \bar{y} - 11 = 0$$

Solving these two equations, we get $\bar{x} = 3$ and $\bar{y} = 1$

Thus, the arithmetic means of x and y are given by 3 and 1 respectively.

(ii) Let us assume that $7x - 3y - 18 = 0$ represents the regression line of y on x and $4x - y - 11 = 0$ represents the regression line of x on y.

Now $7x - 3y - 18 = 0$

$$\Rightarrow \quad y = (-6) + \frac{(7)}{3}x$$

$$\Rightarrow \quad b_{yx} = \frac{7}{3}$$

Again $4x - y - 11 = 0$

$$\Rightarrow \quad x = \frac{(11)}{4} + \frac{(11)}{4}y \qquad \therefore\ b_{xy} = \frac{1}{4}$$

Thus $r^2 = b_{yx} \times b_{xy}$

$$= \frac{7}{3} \times \frac{1}{4}$$

$$= \frac{7}{12} < 1$$

Since $|r| \le 1 \Rightarrow r^2 \le 1$, our assumptions are correct. Thus, $7x - 3y - 18 = 0$ truly represents the regression line of y on x.

(iii) Since $r^2 = \frac{7}{12}$

$$\therefore r = \sqrt{\frac{7}{12}}$$ (We take the sign of r as positive since both the regression coefficients are positive)

$$= 0.7638$$

(iv) $$b_{yx} = r \times \frac{S_y}{S_x}$$

$$\Rightarrow \frac{7}{3} = 0.7638 \times \frac{S_y}{3} \quad (\because S_x^2 = 9 \text{ as given})$$

$$\Rightarrow S_y = \frac{7}{0.7638} = 9.1647$$

3. The following data relate to the test scores obtained by eight salesmen in an aptitude test and their daily sales in thousands of rupees:

Salesman :	1	2	3	4	5	6	7	8
Scores :	60	55	62	56	62	64	70	54
Sales :	31	28	26	24	30	35	28	24

Solution:

Let the scores and sales be denoted by x and y respectively. We take a, origin of x as the average of the two extreme values, i.e., 54 and 70. Hence, a = 62 similarly, the origin of y is taken as $b = \frac{24 + 35}{2} \cong 30$

Table

Computation of Correlation Coefficient Between Test Scores and Sales

Scores	Sales in	u_i	v_i	u_iv_i	u_i^2	v_i^2
	Rs. 1000	$= x_i - 62$	$= y_i - 30$			
(x_i)	(y_i)					
(1)	(2)	(3)	(4)	(5)=(3) x(4)	$(6)=(3)^2$	$(7)=(4)^2$
60	31	−2	1	−2	4	1
55	28	−7	−2	14	49	4
62	26	0	−4	0	0	16
56	24	−6	−6	36	36	36
62	30	0	0	0	0	0
64	35	2	5	10	4	25
70	28	8	−2	−16	64	4
54	24	−8	−6	48	64	36
Total	—	−13	−14	90	221	122

Since correlation coefficient remains unchanged due to change of origin, we have

$$r = r_{xy} = r_{uv} = \frac{n\sum u_iv_i - \sum u_i \times \sum v_i}{\sqrt{n\sum u_i^2 - (\sum u_i)^2} \times \sqrt{n\sum v_i^2 - (\sum v_i)^2}}$$

$$= \frac{8 \times 90 - (-13) \times (-14)}{\sqrt{8 \times 221 - (-13)^2} \times \sqrt{8 \times 122 - (-14)^2}} = \frac{538}{\sqrt{1768 - 169} \times \sqrt{976 - 196}}$$

$$= 0.48$$

In some cases, there may be some confusion about selecting the pair of variables for which correlation is wanted. This is explained in the following problem.

4. Examine whether there is any correlation between age and blindness on the basis of the following data:

Age in years :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Persons (in thousands):	90	120	140	100	80	60	40	20
No. of blind Persons:	10	15	18	20	15	12	10	06

Solution:
Let us denote the mid-value of age in years as x and the number of blind persons per lakh as y. Then as before, we compute correlation coefficient between x and y.

Table : Computation of correlation between age and blindness

Age in	Mid-value	No. of	No. of	No. of	xy	x^2	y^2
years	x	Persons	blind	blind per	(2)×(5)	$(2)^2$	$(5)^2$
(1)	(2)	('000)	B	lakh	(6)	(7)	(8)
		P	(4)	y=B/P × 1 lakh			
		(3)		(5)			
0-10	5	90	10	11	55	25	121
10-20	15	120	15	12	180	225	144
20-30	25	140	18	13	325	625	169
30-40	35	100	20	20	700	1225	400
40-50	45	80	15	19	855	2025	361
50-60	55	60	12	20	1100	3025	400
60-70	65	40	10	25	1625	4225	625
70-80	75	20	6	30	2250	5625	900
Total	320	—	—	150	7090	17000	3120

The correlation coefficient between age and blindness is given by

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

$$= \frac{8\times 7090 - 320\times 150}{\sqrt{8\times 17000 - (320)^2} \times \sqrt{8\times 3120 - (150)^2}}$$

$$= \frac{8720}{183.3030 \times 49.5984} = 0.96$$

which exhibits a very high degree of positive correlation between age and blindness.

5. Coefficient of correlation between x and y for 20 items is 0.4. The AM's and SD's of x and y are known to be 12 and 15 and 3 and 4 respectively. Later on, it was found that the pair (20, 15) was wrongly taken as (15, 20). Find the correct value of the correlation coefficient.

Solution:
We are given that n = 20 and the original r = 0.4, $\bar{x}$ = 12, $\bar{y}$ = 15, S_x = 3 and S_y = 4

$$r = \frac{cov(x,y)}{S_x.S_y} \qquad = 0.4 = \frac{cov(x,y)}{3\times 4}$$

$$= Cov\,(x, y) = 4.8$$

$$= \frac{\sum xy}{n} - \bar{x}\bar{y} = 4.8$$

$$= \frac{\sum xy}{20} - 12\times 15 = 4.8$$

$$= \Sigma xy = 3696$$

Hence, corrected $\Sigma xy = 3696 - 20 \times 15 + 15 \times 20 = 3696$

Also, $S_x^2 = 9$

$= (\Sigma x^2/20) - 12^2 = 9$

$\Sigma x^2 = 3060$

Similarly, $S_y^2 = 16$

$$S_y^2 = \frac{\sum y^2}{20} - 15^2 = 16$$

$\Sigma y^2 = 4820$

Thus, corrected $\Sigma x = n\bar{x}$ – wrong value + correct value.

$= 20 \times 12 - 15 + 20$

$= 245$

Similarly, corrected $\Sigma y = 20 \times 15 - 20 + 15 = 295$

Corrected $\Sigma x^2 = 3060 - 15^2 + 20^2 = 3235$

Corrected $\Sigma y^2 = 4820 - 20^2 + 15^2 = 4645$

Thus, corrected value of the correlation coefficient by applying formula

$$= \frac{20\times 3696 - 245\times 295}{\sqrt{20\times 3235 - (245)^2} \times \sqrt{20\times 4645 - (295)^2}}$$

$$= \frac{73920 - 72275}{68.3740 \times 76.6480}$$

$= 0.31$

6. If the regression line of y on x is given by y = x + 2 and Karl Pearson's coefficient of correlation is 0.5,

then $\frac{\sigma_y^2}{\sigma_x^2} =$

Solution: The regression line of y on x is given by y = x + 2

$x - y + 2 = 0$

$$b_{yx} = -\frac{\text{coefficent of x}}{\text{coefficent of y}} = \frac{-1}{-1} = 1$$

$b_{yx} = 1$

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$ [*Coefficent of correlation* $(r) = 0.5$]

$1 = 0.5\frac{\sigma_y}{\sigma_x}$

$\frac{\sigma_y}{\sigma_x} = \frac{1}{0.5} = 2$

$\left(\frac{\sigma_y}{\sigma_x}\right)^2 = 2^2 \rightarrow \frac{\sigma_y^2}{\sigma_x^2} = 4$

7. Two variables x and y are related according to 4x + 3y = 7, then x and y are:

Solution: Given regression equation

$4x + 3y = 7$ and $4x + 3y = 7$

$3y = 7 - 4x$ and $4x = 7 - 3y$

$y = \frac{7}{3} - \frac{4x}{3}$ and $x = \frac{7}{4} - \frac{3}{4}y$

$y = a + bx$ and $x = a + by$

We get

$b_{yx} = -4/3$ and $b_{xy} = -3/4$

$r = \pm\sqrt{b_{yx} \times b_{xy}} = \pm\sqrt{\left(\frac{-4}{3}\right)\left(-\frac{3}{4}\right)} = -\sqrt{1}$ (both b_{xy},b_{yx} &r are same sign)

r = –1(Perfectly Negative correlation)