Chapter 3: Correlation and Regression

The statistical tool with the help of which the relationship between two or more variables is studied is called correlation. The measure of correlation is called the Correlation Coefficient.

Uses of Correlation Coefficient

1. Helps us measure the relationship between the variables.
2. If the variables are closely related, we can estimate the value of one variable, given the value of another with the help of Regression Analysis
3. Helps in analyzing the economic behavior
4. Helps in the study of social science. For e.g. The relationship between smoking and lung cancer.

Correlation and Causation

1. The correlation may be due to pure chance, especially in a sample. For e.g., relationship between salary and weight.
2. Both the correlated variables may be influenced by one or more variables. For e.g., a high degree of correlation between the yield per acre of rice and wheat may be due to heavy rainfall or fertilizers used.
3. Both the variables may be mutually influencing each other, so that neither can be designated as cause and other effect. For e.g., demand and price.
4. **Nonsense / Illusory Correlation:** A correlation between two variables that is not due to any causal relationship but related to a third variable, or to random sampling fluctuations. E.g. Global warming and no. of pirates.

Types of Correlation

1. **Positive Correlation or Direct Correlation:** When the two variables are directly related, i.e., when one increases the other also increases, it is said to be positive correlation. For e.g., Supply and price.
2. **Negative or Indirect Correlation:** When the two variables are inversely related, i.e., when one increases the other decreases, it is said to be negative correlation. For e.g., Demand and supply
3. **Partial Correlation:** When one variable is independent and the other is dependent on the former, it is a case of partial correlation
4. **Simple Correlation:** When only two variable are studied, it is called simple correlation
5. **Multiple Correlation:** When three or more variables are studied, it is called multiple correlation
6. **Linear Correlation:** When the two variable change by a fixed proportion, thus forming a straight line, it is said to be linear correlation
7. **Non-linear or Curvilinear Correlation:** If the variables, when plotted on a graph do not form a straight line, it is said to be curvilinear correlation. In other words, the amount of change in one variable does not bear a constant change in the other variable.

Methods of Determining Correlation

1. Karl Pearson's Coefficient of Correlation 2. Spearman's Rank Coefficient of Correlation
3. Concurrent Deviation Method 4. Scatter Diagram method 5. Method of Least Squares

Karl Pearson's Coefficient of Correlation

This is the most widely used method of measuring correlation. It is denoted by the symbol 'r'.

Assumptions While Using Karl Pearson's Coefficient of Correlation

While using Karl Pearson's coefficient of correlation, it is assumed that,

1. The distribution is normal
2. There is cause and effect relationship between the variables.
3. There is a linear relationship between the variables.

Properties of Karl Pearson's Coefficient of Correlation

1. The value of r always lies between -1 and +1. Interpretation: ±1 – Perfect correlation; ±0.9 to ±0.1 – Very high degree; ±0.75 to ±0.9 – High degree; ±0.60 to ±0.75 – Moderate degree; ±0.30 to ±0.60 – Low degree; 0 to ±0.30 – Very low degree; 0 – No correlation.
2. It is independent of change of scale and origin of X and Y variables.
3. It is the geometric mean of two regression coefficients.$r = \sqrt{b_{xy} \text{ x } b_{yx}}$

Merits of Karl Pearson's Coefficient of Correlation

1. This is the most popular among the mathematical methods
2. It summarizes in one value the degree of correlation and its direction – direct or inverse.

Formulae

$$\mathrm{r} = \frac{\mathrm{N.\Sigma XY} - \mathrm{\Sigma X.\Sigma Y}}{\sqrt{\mathrm{N\Sigma X^2} - (\Sigma \mathrm{X})^2}\ \sqrt{\mathrm{N\Sigma Y^2} - (\Sigma \mathrm{Y})^2}}$$

Exercise 3.1

1. Compute the Karl Pearson's coefficient of correlation from the following data: (Ans.: +0.9243)

Internal Marks	25	30	22	12	19	24
External Marks	56	68	40	24	28	60

2. Compute the coefficient of correlation from the following data: (Ans.: +0.6051)

X	6	8	9	14	17	28	24	31	7
Y	10	12	15	15	18	25	22	26	28

3. Compute the coefficient of correlation from the following data: (Ans.: +0.8818)

X	45	55	56	58	60	65	68	70	75	80
Y	56	50	48	60	62	64	65	70	74	82

4. Compute the coefficient of correlation from the following data: (Ans.: – 0.7327)

X	43	44	46	40	44	42	45	42	38	40	42	57
Y	29	31	19	18	19	27	27	29	41	30	26	10

Spearman's Rank Correlation

Formulae

Unique Ranks: $\mathrm{r_s} = 1 - \frac{6\,\Sigma \mathrm{d}^2}{\mathrm{N}^3 - \mathrm{N}}$ $d = R_1 - R_2$

Tied Ranks: $\mathrm{r_s} = 1 - \frac{6\,[\Sigma \mathrm{d}^2 + \frac{1}{12}(\mathrm{m}_1^3 - \mathrm{m}_1) + \frac{1}{12}(\mathrm{m}_2^3 - \mathrm{m}_2) + \cdots + \frac{1}{12}(\mathrm{m}_n^3 - \mathrm{m}_n)]}{\mathrm{N}^3 - \mathrm{N}}$ where m = No. of tied ranks

Exercise 3.2

1. Two ladies ranked seven brands of lipsticks as follows. Find the degree of agreement (Ans.: 0.786):

Lady 1	1	3	2	7	6	4	5
Lady 2	2	1	4	6	7	3	5

2. In a beauty competition, two judges ranked 12 participants as follows. What is the degree of agreement between them? (Ans.: – 0.4546)

X	3	4	1	5	2	10	6	9	8	7	12	11
Y	6	10	12	3	9	2	5	8	7	4	1	11

3. Compute the rank correlation from the following data (Ans.: 0.8322):

X	60	34	40	50	45	41	22	43	42	66	64	46
Y	75	32	35	40	45	33	12	30	36	72	41	57

4. From the marks scored in accountancy and statistics by 12 students, compute rank correlation (Ans.: 0):

Accountancy	60	15	20	28	12	40	80	20
Statistics	10	40	30	50	30	20	60	30

5. Compute the coefficient of rank correlation (Ans.: 0.733):

X	48	33	40	9	16	16	65	24	16	57
Y	13	13	24	6	15	4	20	9	6	19

6. Ten competitors in a voice contest are ranked by three judges in the following order. Find which pair of judges have the nearest approach to common liking in voice (Ans.: -0.212, -0,297, 0.6364; Judges 1 & 3):

Judge 1	1	6	5	10	3	2	4	9	7	8
Judge 2	3	5	8	4	7	10	2	1	6	9
Judge 3	6	4	9	8	1	2	3	10	5	7

Linear Regression

The statistical tool with the help of which we are in a position to estimate or predict the unknown values of one variable from known values of another variable is called regression.

Correlation vs. Regression

1. Correlation coefficient is a measure of degree of co-variability between two variables, but regression analysis helps to predict the value of one variable given the value of the other.
2. The cause and effect relation is clearly indicated more through regression analysis than by correlation, which is more a tool of ascertaining the degree of relationship between the variables.

Formulae

Equation X on Y: $(X - \overline{X}) = b_{xy} (Y - \overline{Y})$

Equation Y on X: $(Y - \overline{Y}) = b_{yx} (X - \overline{X})$

Formulae to Find the Regression Coefficients:

Using Assumed Mean: $b_{xy} = \frac{N\,\Sigma XY - \Sigma X.\Sigma Y}{N\,\Sigma Y^2 - (\Sigma Y)^2}$; $b_{yx} = \frac{N\,\Sigma XY - \Sigma X.\Sigma Y}{N\,\Sigma dX^2 - (\Sigma X)^2}$

Using Standard Deviation: $b_{xy} = r.\frac{\sigma_x}{\sigma_y}$; $b_{yx} = r.\frac{\sigma_y}{\sigma_x}$

Coefficient of Correlation: $r = \sqrt{b_{xy} \text{ x } b_{yx}}$

Exercise 3.3

1. Find the Regression Equations (Answer: X = 1.3Y – 4.4 & Y = 0.65X + 4.1):

X	2	4	6	8	10
Y	5	7	9	8	11

2. A panel of judges P & Q graded seven dramatic performances by awarding marks as follows. Obtain the two Regression Equations: (Answer: X = 0.75Y + 14.5 & Y = 0.75X + 5.75)

Performance	1	2	3	4	5	6	7
Marks by P	46	42	44	40	43	41	45
Marks by Q	40	38	36	35	39	37	41

3. Following Table shows the exports of raw cotton and the imports of manufactured goods into India for seven years.

Exports	42	44	58	55	89	98	60
Imports	56	49	53	58	67	76	58

Obtain the two Regression Equations and estimate the imports when export in a particular year was ₹ 70 crore. (Answer: 62.03; X = 2.198Y – 67.244 & Y = 0.391X + 34.651)

4. The advertisement expenses and sales data of ABC company are as follows:

Advertisement Expenses (₹ Lakh)	60	62	65	70	73	75	71
Sales (₹ Crore)	10	11	13	15	16	19	14

Find:

a. Sales for advertisement expenses of ₹ 90 lakhs. (Answer: ₹ 25.224 Crore)

b. Advertisement expenses for a sales target of ₹ 25 Crore. (Answer: ₹ 87.643 Lakh)

c. Coefficient of Correlation (Answer: 0.9545)

(The Regression Equations are: X = 1.786Y + 43 and Y = 0.51X – 20.694)

5. Following data are available on sales and advertisement:

	Sales (₹)	Advertisement Expenses (₹)
Mean	70,000	15,000
Standard Deviation	15,000	3,000

Coefficient of correlation is +0.8

Find:

a. The two Regression Equations (Answer: X = 4Y + 10,000 & Y = 0.16X + 3,800)

b. The advertisement budget if the company desires to achieve the target sales of ₹ 1,00,000 (Answer: ₹ 19,800)

6. Coefficient of correlation between the ages of brothers and sisters in a community was found to be 0.8. Average age of the brothers was 25 and that of sisters 22 years. Their standard deviations were 4 and 5 respectively.

Find:

a. The expected age of the brother when sister's age is 12 years. (Answer: 18.6 years)

b. The expected age of the sister when brother's age is 33 years. (Answer: 30 years)

(The Regression Equations are: X = 0.64Y + 10.92 and Y = X – 3)