# UNIT V

**Correlation and Regression analysis:** Definitionof correlation, Scatter Diagram, Karl Pearson's Coefficient of correlation; Partial and Multiple correlation coefficients (for three variables); Definition of Regression, Simple Linear Regression (for 2 variables).

**Small Sample Tests:** Basic Definitions of testing of hypothesis; **t-Test:** t-test for single Mean, t-test for difference of Means, Paired t-test. **F-Test:** F-test for equality of two population variances. **CHI-SQUARE Test:** test for single variance(population variance)and test of independence of attributes.

## 5.1 Definition of Correlation

**Q1. Define the term correlation.**

**(OR)**

**What do you mean by correlation?**

*Ans :*

**Meaning**

Correlation is the study of the linear relationship between two variables. When there is a relationship of 'quantitative measure between two set of variables, the appropriate statistical tool for measuring the relationship and expressing each in a precise way is known as correlation.

**For example,** there is a relationship between the heights and weights of persons, demand and prices of commodities etc.

Correlation analysis is the statistical tool we can use to describe the degree to which one variable is linearly related to another.

**Definitions**

(i) **According to L.R. Connor** "If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated."

(ii) **According to A.M. Tuttle** "Correlation is an analysis of covariation between two or more variables".

(iii) **According to Croxton and Cowden** "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

(iv) **According to Ya Lun Chou** "Correlation analysis attempts to determine the 'degree of relationship' between variable".

**Q2. Explain the Significance of Measuring Correlation.**

**(OR)**

**How do you say that the correlation between the two variables is significant (or) not.**

*Ans :* **(Imp.)**

1. Correction is very useful to economists to study the relationship between variables, like price and quantity demanded. To business-men, it helps to estimate costs, sales, price and other related variables.
2. In economic theory we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply and quantity demanded; convenience, amenities, and service standards are related to customer retention; yield a crop related to quantity of fertilizer applied, type of soil, quality of seeds, rainfall and so on. Correlation analysis helps in measuring the degree of association and direction of such relationship.
3. The relation between variables can be verified and tested for significance, with the help of the correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.
4. The coefficient of correlation is a relative measure and we can compare the relationship between variables, which are expressed in different units.

5. Correlations are useful in the areas of health care such as determining the validity and reliability of clinical measures or in expressing how health problems are related to certain biological or environmental factors. For example, correlation coefficient can be used to determine the degree of inter-observer reliability for two doctors who are assessing a patient's disease.
6. Sampling error can also be calculated.
7. Correlation is the basis for the concept of regression and ratio of variation.
8. The decision making is heavily felicitated by reducing the range of uncertainty and hence empowering the predictions.

**Q3. Explain various Types of Correlation.**

**(OR)**

**What are the different Types of correlations.**

*Ans :*

Broadly speaking, there are four types of correlation, namely,

A) Positive correlation,

B) Negative correlation,

C) Linear correlation and

D) Non-Linear Correlation.

**A) Positive correlation**

If the values of two variables deviate in the same direction i.e., if increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, the corresponding correlation is said to be positive or direct.

**Examples**

i) Sales revenue of a product and expenditure on Advertising.

ii) Amount of rain fall and yield of a crop (up to a point)

iii) Price of a commodity and quantity of supply of a commodity

iv) Height of the Parent and the height of the Child.

v) Number of patients admitted into a Hospital and Revenue of the Hospital.

vi) Number of workers and output of a factory.

**i) Perfect Positive Correlation :** If the variables X and Y are perfectly positively related to each other then, we get a graph as shown in fig. below.

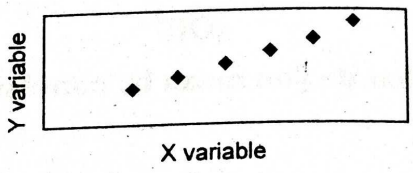



**Fig.: Perfect Positive Correlation (r = +1)**

**ii) Very High Positive Correlation :** If the variables X and Y are related to each other with a very high degree of positive relationship then we can notice a graph as in figure below.

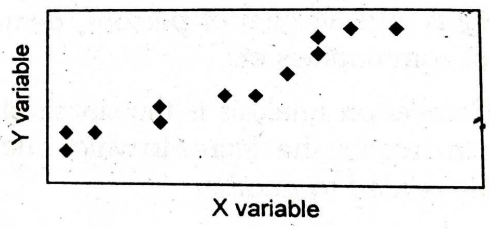



**Fig.: Very High Positive Correlation (r = nearly + 1)**

**iii) Very Low Positive Correlation :** If the variables X and Y are related to each other with a very low degree of positive relationship then we can notice a graph as in fig. below.

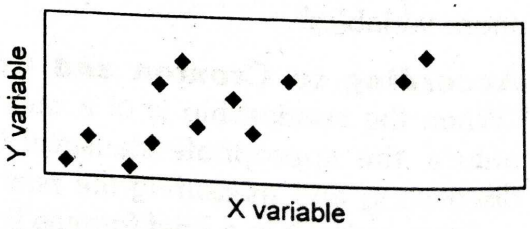



**Fig.: Very Low Positive Correlation (r = near to +0)**

### B) Negative Correlation

Correlation is said to be negative or inverse if the variables deviate in the opposite direction i.e., if the increase (decrease) in the values of one variable results, on the average, in a corresponding decrease (increase) in the values of the other variable.

**Examples**

1. Price and demand of a commodity.
2. Sales of Woolen garments and the day temperature.

**i) Perfect Negative Correlation :** If the variables X and Y are perfectly negatively related to each other then, we get a graph as shown in fig. below.

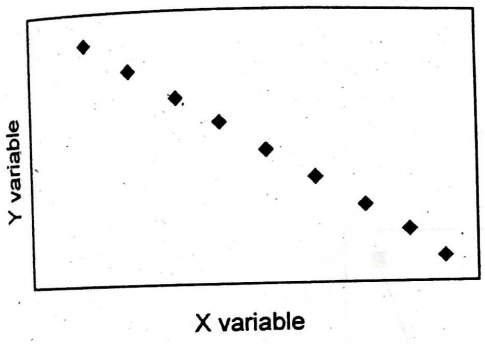



**Fig.: Perfect Negative Correlation (r = –1)**

**ii) Very High Negative Correlation :** If the variables X and Y are related to each other with a very high degree of negative relationship then we can notice a graph as in fig. below.

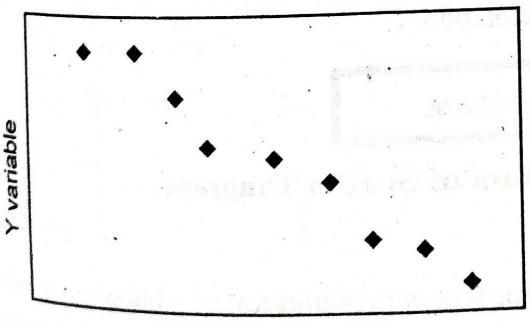



**Fig.: Very High Negative Correlation (r = near to –1)**

**iii) Very low Negative Correlation :** If the variables X and Y are related to each other with a very low degree of negative relationship then we can notice a graph as in fig. below.

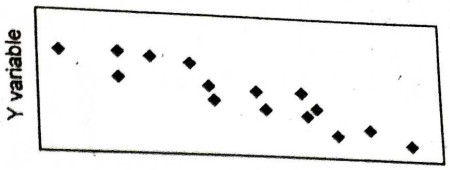



**Fig.: Very Low Negative Correlation**

**(r = near to 0 but negative)**

**iv) No Correlation :** If the scatter diagram show the points which are highly spread over and show no trend or patterns we can say that there is no correlation between the variables.

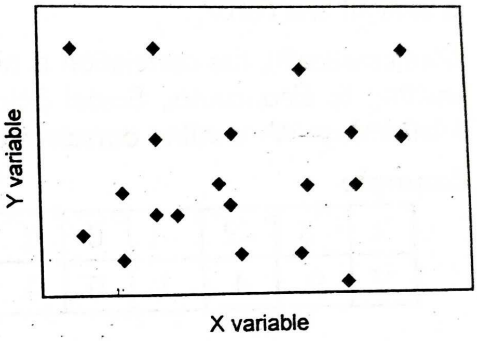



**Fig.: No Correlation (r = 0)**

### C) Linear Correlation

Two variables are said to be linearly related if corresponding to a unit change in one variable there is a constant change in the other variable over the entire range of the values.

If two variables are related linearly, then we can express the relationship as

$$Y = a + b X$$

where 'a' is called as the "intercept" (If X = 0, then Y = a) and 'b' is called as the "rate of change" or slope.

If we plot the values of X and the corresponding values of Y on a graph, then the graph would be a straight lines as shown in fig. below.

**Example**

| X | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Y | 8 | 11 | 14 | 17 | 20 |

For a unit change in the value of x, a constant 3 units changes in the value of y can be noticed. The above can be expressed as : Y = 5 + 3x.

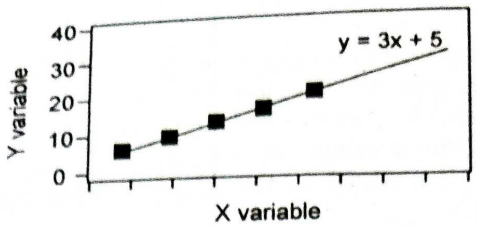



Fig.: Linear Correlation

**D) Non Linear (Curvilinear) Correlation**

If corresponding to a unit change in one variable, the other variable does not change in a constant rate, but change at varying rates, then the relationship between two variables is said to be nonlinear or curvilinear as shown in fig. below. In this case, if the data are plotted on the graph, we do not get a straight line curve.

Mathematically, the correlation is non-linear if the slope of the plotted curve is not constant. Data relating to Economics, Social Science and Business Management do exhibit often non-linear relationship. We confine ourselves to linear correlation only.

**Example**

| X | –3 | –2 | –1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| Y | 9 | 4 | 1 | 0 | 1 | 4 | 9 |

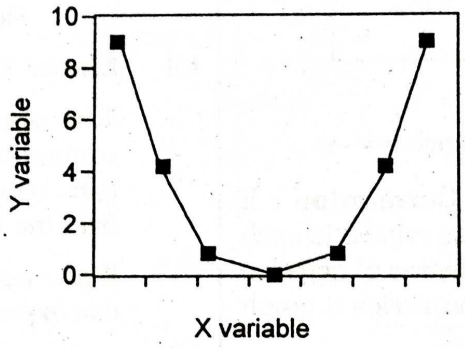



Fig.: Non Linear Correlation

## 5.2 Scatter Diagram

**Q4. What is Scatter Diagram? Explain the procedure of Scatter Diagram.**

*Ans :*

Scatter diagram method is the simplest way of diagrammatic representation of a bivariate distribution and helps in ascertaining the correlation between the two variables under study i.e., it portraits the relationship between these two variables graphically.

**Procedure of Scatter Diagram**

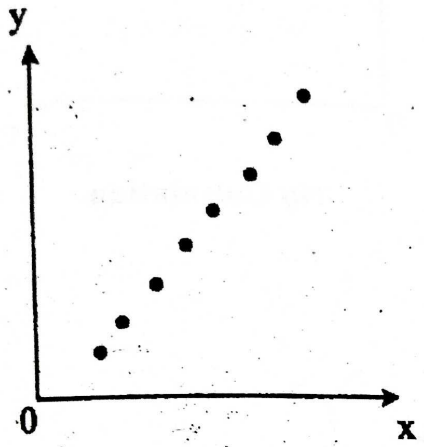
Given pair of values $(x_1, y_2)$ $(x_2, y_2)$... $(x_n, y_n)$ of two variables X and Y. Take the independent variables on the X axis and the dependent variable on the F-axis. The V points denoted by the pair of values are plotted on the graph. The diagram of dots thus obtained is the scatter diagram. Regarding the correlation between the two variables, the scatter diagram can be interpreted as follows,

(i) If the points reveal any upward or downward trend, the variables are said to be correlated, otherwise uncorrelated.

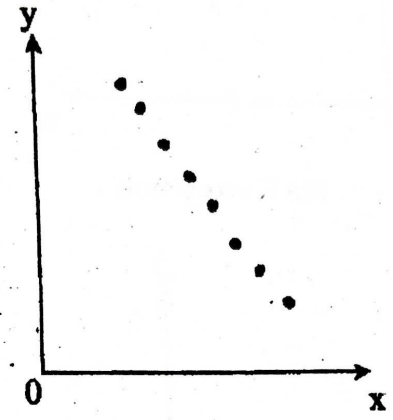
(ii) If the points are very close to each other, a good amount of correlation exists, else poor correlation exists.

(iii) Upward trend indicates positive correlation and downward trend indicates negative correlation.

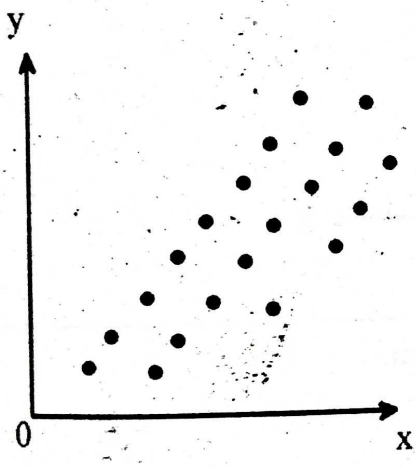
**Different Forms of Scatter Diagram**



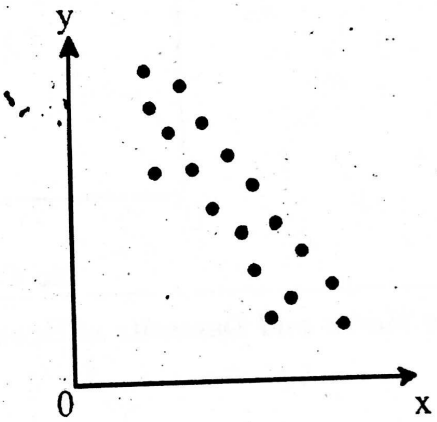

**Perfect Positive Correlation**



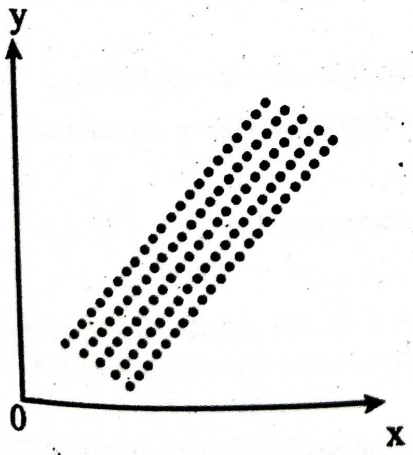

**Perfect Negative Correlation**



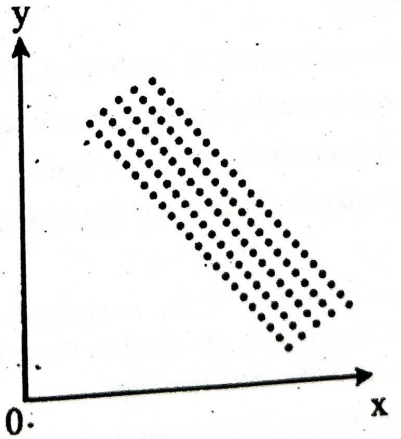

**Low Degree Positive Correlation**



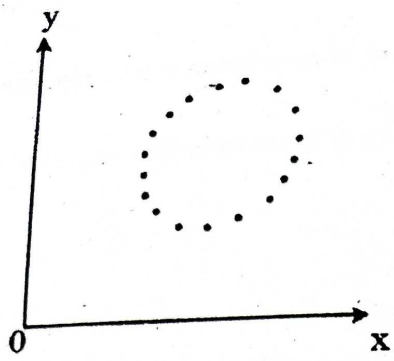

**Low Degree Negative Correlation**



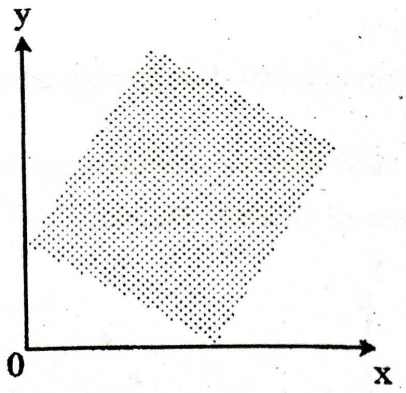

**High Degree Positive Correlation**



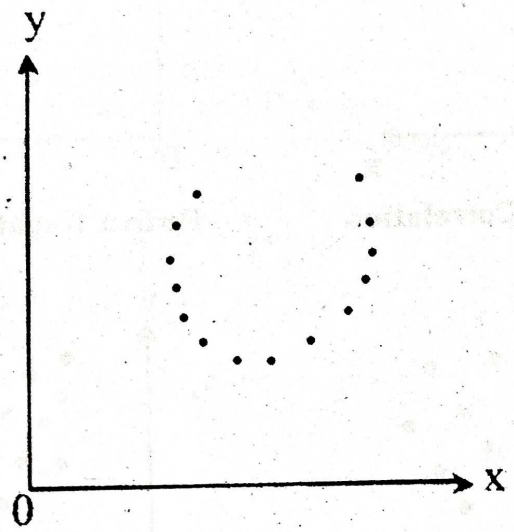

**High Degree negative Correlation**




No Correlation




No Correlation




No Correlation

**Q5. State the Merits and Demerits of Scatter Diagram.**

*Ans :*

**Merits**

- Scatter diagram is a simple and attractive method of finding out the nature of correlation between two variables.
- It is a non-mathematical method of studying correlation. It is easy to understand.
- We can get a rough idea at a glance whether it is a positive or negative correlation.
- It is not influenced by extreme items.
- It is a first step in finding out the relationship between two variables.

**Demerits**

- The major limitation of the method is that it only gives a visual picture of the relationship of two variables. It only tells us whether there is correlation between the variables, and if so, then in which direction, positive (or) negative.
- It does not give an idea about the precise degree of relationship as it is not amenable to mathematical treatment.

## 5.3 KARL PERSON'S COEFFICIENT OF CORRELATION

**Q6. What is Karl Pearson's Coefficient of Correlation? Explain properties of Coefficient of Correlation.**

*Ans :* **(Imp.)**

Karl Parson's Coefficient of Correlation is arrived at with the help of a statistical formula that takes into account the mean and standard deviation of the two variables, the number of such observations and the covariance between them. Since Karl Pearson's coefficient of correlation is a number, it can describe the strength of the correlation in greater detail and more objectively. A value of –1 signifies "absolute" negative correlation, a value between –1 and –0.5 signifies strong negative correlation, a value between –0.5 and –0.25 signifies moderate negative correlation and a value between –0.25 and 0 signifies weak negative correlation. Similarly, a value of +1 signifies "absolute" positive correlation, a value between +1 and +0.5 signifies strong positive correlation, a value between +0.5 and +0.25 signifies moderate positive correlation and a value between +0.25 and 0 signifies weak positive correlation.

**Properties of Karl Pearson's Coefficient of Correlation**

1. It is based on Arithmetic Mean and Standard Deviation.
2. It lies between –1
3. It measures both direction as well as degree of change. If r is less than 0, there is negative correlation, which means the direction of change of the two variables will be opposite. If r is more than 0, there is positive correlation, which means the direction of change of the two, variables will be same. Higher the value of r, greater is the degree of correlation. Hence, Karl Pearson's coefficient of correlation is said to be the ideal measure of correlation.
4. It is independent of change in scale. In other words, if a constant amount is added/ subtracted from all values of a variable, the value of r does not change.
5. It is independent of change in origin. Thus, if a constant amount is multiplied with or divides all values of a variable, Y does not change.
6. It is independent of direction. In other words, Correlation of X and Y is same as Correlation of Y and X.
7. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.
8. It takes into account all items of the variable(s).
9. It does not prove causation but is simply a measure of co-variation.
10. Correlation coefficient of two variables X and Y is the Geometric Mean of two regression coefficients, regression coefficient of X on Y and regression coefficient of Y on X. Symbolically,

    r = Square root of $(b_{xy} * b_{yx})$
11. Correlation coefficient can be calculated between two unrelated variables and such a number can be misleading. Such correlation is called accidental correlation, spurious correlation or non sense correlation.

**Q7. Explain Merits and Demerits of Coefficient of Correlation.**

*Ans :*

**Merits**

1. It takes into account all items of the variable(s).
2. It is a numerical measure and hence more objective.
3. It measures both direction as well as degree of change.
4. It facilitates comparisons between two series.
5. It is capable of further Algebraic treatment
6. It is more practical and hence popular and is more commonly used.

**Demerits**

1. It is not easy to calculate as complex formulae are involved.
2. It is more time consuming compared to methods such as rank correlation

3. It assumes a linear relationship between the two variables which may not be correct
4. It is impacted by extreme values as it is based o mean and standard deviation.
5. It is not easy to interpret.

**Q8. Explain the various methods of Coefficient of Correlation.**

*Ans :*

**i) Direct Method when deviations are taken from actual mean**

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

Where

$x = X - \bar{X}$ and $y = Y - \bar{Y}$

**Steps :**

1. Find the means of the two series ($\bar{X}$ , $\bar{Y}$)
2. Take the deviations of X series from the mean of X and denote these deviations as x.
3. Square these deviations and obtain the total. Denote it as $\Sigma x^2$.
4. Take the deviations of Y series from the Mean of Y and denote these deviations as y.
5. Square these deviations, obtain the total and denote it as $\Sigma y^2$.
6. Multiply the deviations of X and Y series, obtain the total and denote it $\Sigma xy$.
7. Substitute the above values in the formula.

**ii) Short-Cut Method**

**When deviations are taken from assumed mean.**

$$r = \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2}\sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

Where, $x = X - A_1$

$y = X - A_2$

$A_1$ = Assumed mean for X series

$A_2$ = Assumed mean for Y series

The values of coefficient of correlation as obtained by above formulae will always lie between $\pm 1$. When there is perfect positive correlation its value is +1 and when there is perfect negative correlation, its value is –1. When r = 0 means that there is no relationship between the two variables. We normally get values which lie between +1 and –1.

## PROBLEMS

**1. Find the coefficient of correlation from the following data :**

| X | 46 | 54 | 56 | 56 | 58 | 60 | 62 |
|---|---|---|---|---|---|---|---|
| Y | 36 | 40 | 44 | 54 | 42 | 58 | 54 |

**Calculation of Correlation**

| X | $(X - \bar{X})$ (x) | Y | $(Y - \bar{Y})$ (y) | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|---|---|
| 46 | – 10 | 36 | –11 | 100 | 121 | 110 |
| 54 | – 2 | 40 | –7 | 4 | 49 | 14 |
| 56 | 0 | 44 | – 3 | 0 | 9 | 0 |
| 56 | 0 | 54 | 7 | 0 | 49 | 0 |
| 58 | + 2 | 42 | – 5 | 4 | 25 | –10 |
| 60 | + 4 | 58 | 11 | 16 | 121 | 44 |
| 62 | + 6 | 54 | 7 | 36 | 49 | 42 |
| | | | | 160 | 423 | 200 |

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$x = (x - \bar{x})$$

$$\bar{X} = \frac{392}{7} = 56,$$

Where,

$x^2 = 160$ , $y^2 = 1423$, $xy = 200$

$$= \frac{200}{\sqrt{160 \times 423}}$$

$$y = (y - \bar{y})$$

$$\bar{Y} = \frac{328}{7} = 46.85 = 47$$

$$= \frac{200}{260.15} = 0.768$$

**Calculate Karl Pearson's Coefficient of Correlation for the following data.**

| X | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|
| Y | 18 | 16 | 14 | 12 | 10 | 6 | 8 |

(Imp.)

*Sol :*

**Calculations of Karl pearson's coefficient of correlation**

| X | Y | $X - \bar{X}$ (x) | $Y - \bar{Y}$ (y) | xy | $x^2$ | $y^2$ |
|---|---|---|---|---|---|---|
| 7 | 18 | 3 | 6 | 18 | 9 | 36 |
| 6 | 16 | 2 | 4 | 8 | 4 | 16 |
| 5 | 14 | 1 | 2 | 2 | 1 | 4 |
| 4 | 12 | 0 | 0 | 0 | 0 | 0 |
| 3 | 10 | -1 | -2 | +2 | 1 | 4 |
| 2 | 6 | -2 | -6 | 12 | 4 | 36 |
| 1 | 8 | -3 | -4 | 12 | 9 | 16 |
| | | | | **54** | **28** | **112** |

$$\bar{X} = \frac{28}{7} = 4$$

$$\bar{Y} = \frac{84}{7} = 12$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$r = \frac{54}{\sqrt{28 \times 112}}$$

$$= \frac{54}{\sqrt{3136}} = \frac{54}{56} = 0.96$$

## 5.4 Partial and Multiple Correlation Coefficients (for Three Variables)

**Q9. Explain partial correlation coefficient with an example.**

(Imp.)

*Ans :*

Two variables, A and B, are closely related. The correlation between them is partialled out, or controlled for the influence of one or more variables is called as partial correlation. So when it is assumed that some other variable is influencing the correlation between A and B, then the influence of this variable(s) is partialled out for both A and B. Hence it can be considered as a correlation between two sets of residuals. Here we discuss a simple case of correlation between A and B is partialled out for C. This can be represented as $r_{AB.C}$ which is read as correlation between A and B partialled out for C. the correlation between A and B can be partialled out for more variables as well.

**Formula and Example**

For example, the researcher is interested in computing the correlation between anxiety and academic achievement controlled from intelligence. Then correlation between academic achievement (A) and anxiety (B) will be controlled for Intelligence (C). This can be represented as: $r_{\text{Academic Achievement(A) Anxiety (B) . Intelligence (C)}}$. To calculate the partial correlation ($r_p$) we will need a data on all three variables. The computational formula is as follows:

$$r_p = r_{AB.C} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1-r_{AC}^2)(1-r_{BC}^2)}}$$

Look at the data of academic achievement, anxiety and intelligence. Here, the academic achievement test, the anxiety scale and intelligence test is administered on ten students. The data for ten students is provided for the three variables in the table below.

| Subject | Academic Achievement | Anxiety | Intelligence |
|---|---|---|---|
| 1 | 15 | 6 | 25 |
| 2 | 18 | 3 | 29 |
| 3 | 13 | 8 | 27 |
| 4 | 14 | 6 | 24 |
| 5 | 19 | 2 | 30 |
| 6 | 11 | 3 | 21 |
| 7 | 17 | 4 | 26 |
| 8 | 20 | 4 | 31 |
| 9 | 10 | 5 | 20 |
| 10 | 16 | 7 | 25 |

**Table : Data of academic achievement, anxiety and intelligence for 10 subjects**

In order to compute the partial correlation between the academic achievement and anxiety partialled out for Intelligence, we first need to compute the Pearson's Product moment correlation coefficient between all three variables. We have already learned to compute it in the first Unit of this Block. So I do not again explain it here.

The correlation between anxiety (B) and academic achievement (A) is – 0.369.

The correlation between intelligence (C) and academic achievement (A) is 0.918.

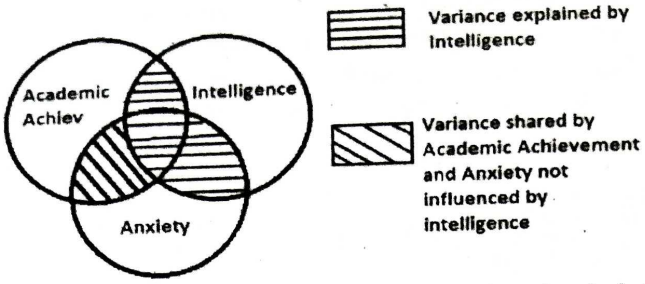
The correlation between anxiety (B) and intelligence (C) is – 0.245.

Give the correlations, we can now calculate the partial correlation.

$$r_{AB.C} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1-r_{AC}^2)(1-r_{BC}^2)}} = \frac{-0.369 - (0.918 \times -0.245)}{\sqrt{(1-0.918^2)(1-(0.245^2))}} = \frac{-.1441}{0.499} = -0.375$$

The partial correlation between the two variables, academic achievement and anxiety controlled for intelligence, is -0.375. You will realise that the correlation between academic achievement and anxiety is -0.369. Whereas, after partialling out for the effect of intelligence, the correlation between them has almost remained unchanged. While computing this correlation, the effect of intelligence on both the variables, academic achievement and anxiety, was removed.

The following figure explains the relationship between them.




**Graphical Explanation of partial correlation between Academic Intelligence and Anxiety controlled for intelligence**

**Fig. : Venn diagram explaining the partial correlation**

Significance testing of the partial correlation

We can test the significance of the partial correlation for the null hypothesis

$$H_0 : \tilde{n}_p = 0$$

and the alternative hypothesis

$$H_0 : \tilde{n}_p = 0$$

Where, the $\tilde{n}_p$ denote the population partial correlation coefficient. The t-distribution is used for this purpose. Following formula is used to calculate the t-value.

$$t = \frac{r_p\sqrt{n-v}}{\sqrt{1-r_p}}$$

Where,

$r_p$ = partial correlation computed on sample, $r_{AB.C}$

n = sample size,

v = total number of variables employed in the analysis.

The significance of the $r_p$ is tested at the df = n – v.

In the present example, we can employ significance testing as follows:

$$t = \frac{r_p\sqrt{n-v}}{\sqrt{1-r_p^2}} = \frac{-.375\sqrt{10-3}}{\sqrt{1-(-.375^2)}} = \frac{-0.992}{0.927} = 1.69$$

We test the significance of this value at the df = 7 in the table for t-distribution in the appendix. You will realise that at the df = 7, the table provides the critical value of 2.36 at 0.05 level of significance. The obtained value of 1.69 is smaller than this value. So we accept the null hypothesis stating that $H_0$: $\tilde{n}_p$ =0.

## Q10. Explain briefly about multiple correlation coefficient.

**(Imp.)**

***Ans :***

If information on two variables like height and weight, income and expenditure, demand and supply, etc. are available and we want to study the linear relationship between two variables, correlation

coefficient serves our purpose which provides the strength or degree of linear relationship with direction whether it is positive or negative. But in biological, physical and social sciences, often data are available on more than two variables and value of one variable seems to be influenced by two or more variables. For example, crimes in a city may be influenced by illiteracy, increased population and unemployment in the city, etc.

The production of a crop may depend upon amount of rainfall, quality of seeds, quantity of fertilizers used and method of irrigation, etc. Similarly, performance of students in university exam may depend upon his/her IQ, mother's qualification, father's qualification, parents income, number of hours of studies, etc. Whenever we are interested in studying the joint effect of two or more variables on a single variable, multiple correlation gives the solution of our problem.

In fact, multiple correlation is the study of combined influence of two or more variables on a single variable.

Suppose, $X_1$, $X_2$ and $X_3$ are three variables having observations on N individuals or units. Then multiple correlation coefficient of $X_1$ on $X_2$ and $X_3$ is the simple correlation coefficient between $X_1$ and the joint effect of $X_2$ and $X_3$. It can also be defined as the correlation between $X_1$ and its estimate based on $X_2$ and $X_3$.

Multiple correlation coefficient is the simple correlation coefficient between a variable and its estimate.

Let us define a regression equation of $X_1$ on $X_2$ and $X_3$ as

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$

Let us consider three variables $x_1$, $x_2$ and $x_3$ measured from their respective means. The regression equation of $x_1$ depends upon $x_2$ and $x_3$ is given by

$$X_1 = b_{12.3}X_2 + b_{13.2}X_3 \quad \text{... (1}$$

Where

$$X_1 - \bar{X}_1 = x_1,\ X_2 - \bar{X}_2 = x_2 \text{ and } X_3 = \bar{X}_3 = x_3$$

$$\therefore \quad \Sigma x_1 = \Sigma x_2 = \Sigma x_3 = 0$$

Right hand side of equation (1) can be considered as expected or estimated value of $x_1$ based on $x_2$ and $x_3$ which may be expressed as

$$x_{1.23} = b_{12.3}x_2 + b_{13.2}x_3 \quad \text{... (2)}$$

Residual $e_{1.23}$ is written as

$$e_{1.23} = x_1 - b_{12.3}x_2 - b_{13.2}x_3 = x_1 - x_{1.23}$$

$$\Rightarrow \quad e_{1.23} = x_1 - x_{1.23} \quad \text{... (3)}$$

$$\Rightarrow \quad x_{1.23} = x_1 - e_{1.23}$$

The multiple correlation coefficient can be defined as the simple correlation coefficient between $x_1$ and its estimate $e_{1.23}$. It is usually denoted by $R_{1.23}$ and defined as

$$R_{1.23} = \frac{\text{Cov}(x_1, x_{1.23})}{\sqrt{V(x_1)V(x_{1.23})}} \quad \text{... (4)}$$

Now,

$$\text{Cov}(x_1, x_{1.23}) = \frac{1}{N}\Sigma (x_1 - \bar{x}_1)(x_{1.23} - \bar{x}_{1.23})$$

(By the definition of covariance)

Since, $x_1$, $x_2$ and $x_3$ are measured from their respective means, so

$$\Sigma x_1 = \Sigma x_2 = \Sigma x_3 = 0$$

$$\Rightarrow \quad \bar{x}_1 = \bar{x}_2 = \bar{x}_3 = 0$$

and consequently

$$\bar{x}_{1.23} = b_{12.3}\bar{x}_2 + b_{13.2}\bar{x}_3 = 0 \qquad \text{(From equation (2))}$$

Thus

$$\text{Cov}(x_1, x_{1.23}) = \frac{1}{N}\Sigma x_1 x_{1.23}$$

$$= \frac{1}{N}\Sigma x_1 (x_1 - e_{1.23}) \qquad \text{(From equation (3))}$$

$$= \frac{1}{N}\Sigma x_1^2 - \frac{1}{N}\Sigma x_1 e_{1.23} \qquad \text{(By third property of residuals)}$$

$$= \frac{1}{N}\Sigma x_1^2 - \frac{1}{N} e_{1.23}^2$$

$$= \sigma_1^2 - \sigma_{1.23}^2$$

Now

$$V(x_{1.23}) = \frac{1}{N}\Sigma (x_{1.23} - \bar{x}_{1.23})^2$$

$$= \frac{1}{N}\Sigma (x_{1.23})^2 \qquad \text{(Since } \bar{x}_{1.23} = 0\text{)}$$

$$= \frac{1}{N}\Sigma (x_1 - e_{1.23})^2 \qquad \text{(From equation (3))}$$

$$= \frac{1}{N}\Sigma (x_1^2 + e_{1.23}^2 - 2x_1 e_{1.23})$$

$$= \frac{1}{N}\Sigma x_1^2 + \frac{1}{N} e_{1.23}^2 - 2\frac{1}{N}\Sigma x_1 e_{1.23}$$

$$= \frac{1}{N}\Sigma x_1^2 + \frac{1}{N} e_{1.23}^2 - 2\frac{1}{N}\Sigma e_{1.23}^2$$

(By third property of residuals)

$$= \frac{1}{N}\Sigma x_1^2 - \frac{1}{N} e_{1.23}^2$$

$$V(x_{1.23}) = \sigma_1^2 - \sigma_{1.23}^2$$

Substituting the value of Cov $(x_1, x_{1.23})$ and $V(x_{1.23})$ in equation (4),

We have

$$R_{1.23} = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma_{1.23}^2)}}$$

$$R_{1.23}^2 = \frac{(\sigma_1^2 - \sigma_{1.23}^2)^2}{\sigma_1^2(\sigma_1^2 - \sigma_{1.23}^2)}$$

$$R_{1.23}^2 = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sigma_1^2} = 1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}$$

Here, $\sigma_{1.23}^2$ is the variance of residual, which is

$$\sigma_{1.23}^2 = \frac{\sigma_1^2}{1 - r_{23}^2}\ (1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{23}r_{13})$$

Then,

$$R_{1.23}^2 = 1 - \frac{\sigma_1^2(1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23})}{\sigma_1^2(1 - r_{23}^2)}$$

$$R_{1.23}^2 = \frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$R_{1.23}^2 = \frac{1 - r_{23}^2 - 1 + r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \qquad \ldots (5)$$

which is required formula for multiple correlation coefficient.

Where,

$r_{12}$ is the total correlation coefficient between variable $X_1$ and $X_2$.

$r_{23}$ is the total correlation coefficient between variable $X_2$ anx $X_3$.

$r_{13}$ is the total correlation coefficient between variable $X_1$ and $X_3$.

## PROBLEMS

3. **From the following data, obtain $R_{1.23}$ and $R_{2.13}$.**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 65 | 72 | 54 | 68 | 55 | 59 | 78 | 58 | 57 | 51 |
| $X_2$ | 56 | 58 | 48 | 61 | 50 | 51 | 55 | 48 | 52 | 42 |
| $X_3$ | 9 | 11 | 8 | 13 | 10 | 8 | 11 | 10 | 11 | 7 |

**Sol :**

To obtain multiple correlation coefficients $R_{1.23}$ and $R_{2.13}$, we use following formulae

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \text{ and}$$

$$R^2_{2.23} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$$

We need, $r_{12}$,$r_{13}$ and $r_{23}$ which are obtained from the following tanle :

| S.No | $X_1$ | $X_2$ | $X_3$ | $(X_1)^2$ | $(X_2)^2$ | $(X_3)^2$ | $X_1X_2$ | $X_1X_3$ | $X_2X_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 65 | 56 | 9 | 4225 | 3136 | 81 | 3640 | 585 | 504 |
| 2 | 72 | 58 | 11 | 5184 | 3364 | 121 | 4176 | 792 | 638 |
| 3 | 54 | 48 | 8 | 2916 | 2304 | 64 | 2592 | 432 | 384 |
| 4 | 68 | 61 | 13 | 4624 | 3721 | 169 | 4148 | 884 | 793 |
| 5 | 55 | 50 | 10 | 3025 | 2500 | 100 | 2750 | 550 | 500 |
| 6 | 59 | 51 | 8 | 3481 | 2601 | 64 | 3009 | 472 | 408 |
| 7 | 78 | 55 | 11 | 6084 | 3025 | 121 | 4290 | 858 | 605 |
| 8 | 58 | 48 | 10 | 3364 | 2304 | 100 | 2784 | 580 | 480 |
| 9 | 57 | 52 | 11 | 3249 | 2704 | 121 | 2964 | 627 | 572 |
| 10 | 51 | 42 | 7 | 2601 | 1764 | 49 | 2142 | 357 | 294 |
| Total | 617 | 521 | 98 | 38753 | 27423 | 990 | 32495 | 6137 | 5178 |

Now we get the total correlation coefficient $r_{12}$, $r_{13}$ and $r_{23}$.

$$r_{12} = \frac{N(\sum X_1X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_2^2) - (\sum X_2)^2\}}}$$

$$r_{12} = \frac{(10 \times 32495) - (617)\text{x}(521)}{\sqrt{\{(10 \times 38753) - (617)\text{x}(617)\}\{(10 \times 27423) - (521)\text{x}(521)\}}}$$

$$r_{12} = \frac{3493}{\sqrt{\{6841\} \text{x} \{2789\}}} = \frac{3493}{4368.01} = 0.80$$

$$r_{13} = \frac{N(\sum X_1X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{13} = \frac{(10\text{x}6137) - (617)\text{x}(98)}{\sqrt{\{(10\text{x}38753) - (617\text{x}617)\}\{(10\text{x}990) - (98\text{x}98)\}}}$$

$$r_{13} = \frac{904}{\sqrt{\{6841\}\{296\}}} = \frac{904}{1423.00} = 0.64$$

and

$$r_{23} = \frac{N(\sum X_2X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\{N(\sum X_2^2) - (\sum X_2)^2\}\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{23} = \frac{(10x5178) - (521)x(98)}{\sqrt{\{(10x27423) - (521x521)\}\{(10x990) - (98x98)\}}}$$

$$r_{23} = \frac{722}{\sqrt{\{2789\}\{296\}}} = \frac{722}{908.59} = 0.79$$

Now, we calculate $R_{1.23}$

We have, $r_{12} = 0.80$, $r_{13} = 0.64$ and $r_{23} = 0.79$, then

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$= \frac{0.80^2 + 0.64^2 - 2x0.80x0.64x0.79}{1 - 0.79^2}$$

$$= \frac{0.64 + 0.41 - 0.81}{1 - 0.62}$$

$$R_{1.23}^2 = \frac{0.24}{0.38} = 0.63$$

Then

$$R_{1.23} = 0.79.$$

$$R_{1.23} = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}$$

$$= \frac{0.80^2 + 0.79^2 - 2x0.80x0.64x0.79}{1 - 0.64^2}$$

$$= \frac{0.64 + 0.62 - 0.81}{1 - 0.49}$$

$$= \frac{0.45}{0.51} = 0.88$$

Thus,

$$R_{2.13} = 0.94$$

4. **From the following data, obtain $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$**

| | | | | |
|---|---|---|---|---|
| $X_1$ | 2 | 5 | 7 | 11 |
| $X_2$ | 3 | 6 | 10 | 12 |
| $X_3$ | 1 | 3 | 6 | 10 |

*Sol :*

To obtain multiple correlation coefficients $R_{1.23}$ $R_{2.13}$ and $R_{3.12}$, we use following formulae

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$$

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{13}} \text{ and}$$

$$R^2_{3.12} = \frac{r^2_{13} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{12}}$$

We need $r_{12}$, $r_{13}$ and $r_{23}$ which are obtained from the following table :

| S.No. | $X_1$ | $X_2$ | $X_3$ | $(X_1)^2$ | $(X_2)^2$ | $(X_3)^2$ | $X_1X_2$ | $X_1X_3$ | $X_2X_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 1 | 4 | 9 | 1 | 6 | 2 | 3 |
| 2 | 5 | 6 | 3 | 25 | 36 | 9 | 30 | 15 | 18 |
| 3 | 7 | 10 | 6 | 49 | 100 | 36 | 70 | 42 | 60 |
| 4 | 11 | 12 | 10 | 121 | 144 | 100 | 132 | 110 | 120 |
| Total | 25 | 31 | 20 | 199 | 289 | 146 | 238 | 169 | 201 |

Now we get the total correlation coefficient $r_{12}$, $r_{13}$ and $r_{23}$.

$$r_{12} = \frac{N(\sum X_1X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_2^2) - (\sum X_2)^2\}}}$$

$$r_{12} = \frac{(4x238) - (25)x(31)}{\sqrt{\{(4x199) - (25)x(25)\}\{(4x289) - (31)x(31)\}}}$$

$$r_{12} = \frac{177}{\sqrt{\{171\}\{195\}}} = \frac{177}{182.61} = 0.0.97$$

$$r_{13} = \frac{N(\sum X_1X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{13} = \frac{(4x169) - (25)x(20)}{\sqrt{\{(4x199) - (25x25)\}\{(4x146) - (20x20)\}}}$$

$$r_{13} = \frac{176}{\sqrt{\{171\}\{184\}}} = \frac{176}{177.38} = 0.99$$

and

$$r_{23} = \frac{N(\sum X_2X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\{N(\sum X_2^2) - (\sum X_2)^2\}\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{23} = \frac{(4x201)-(31)x(20)}{\sqrt{\{(4x289)-(31x31)\}\{(4x146)-(20x20)\}}}$$

$$r_{13} = \frac{184}{\sqrt{\{195\}\{184\}}} = \frac{184}{189.42} = 0.97$$

Now, we calculate $R_{1.23}$.

We have, $r_{12} = 0.97$, $r_{13} = 0.99$ and $r_{23} = 0.97$, then

$$R^2_{1.23} = \frac{r^2_{12}+r^2_{13}-2r_{12}r_{13}r_{23}}{1-r^2_{23}}$$

$$= \frac{0.97^2+0.99^2-2x0.97x0.99x0.97}{1-0.97^2}$$

$$= \frac{0.058}{0.059} = 0.98$$

Then

$$R_{1.23} = 0.99$$

$$R^2_{2.13} = \frac{r^2_{12}+r^2_{13}-2r_{12}r_{13}r_{23}}{1-r^2_{13}}$$

$$= \frac{0.97^2+0.97^2-2x0.97x0.99x0.97}{1-0.99^2}$$

$$= \frac{0.19}{0.20} = 0.95$$

Thus,

$$R_{2.13} = 0.97$$

$$R^2_{3.12} = \frac{r^2_{13}+r^2_{23}-2r_{12}r_{13}r_{23}}{1-r^2_{12}}$$

$$= \frac{0.99^2+0.97^2-2x0.97x0.99x0.97}{1-0.97^2}$$

$$= \frac{0.58}{0.591}$$

$$= 0.981$$

Thus,

$$R_{3.12} = 0.99$$

## 5.5 Definition of Regression

**Q11. What do you understand by Regression?**

**(OR)**

**What is meant by Regression?**

**(OR)**

**Define Regression?**

*Ans :*

**Meaning**

Regression analysis which confines itself to a study of only two variables is called simple regression. The regression analysis which studies more than two variables at a time is called multiple regression. In the simple regression analysis there are two variables-one of which is known as 'independent variable' or 'regressor' or 'predictor'. On the basis of the values of this variable the values of the other variable are predicted. The other variable whose values are predicted is called the 'dependent' or 'regressed' variable.

**Definitions**

1. "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data."
2. **According to Morris Hamburg** The term 'regression analysis' refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process."
3. **According to Taro Yamane** "One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related causally, is regression analysis."
4. **According to YaLum Chou** "Regression analysis attempts to establish the 'nature of the relationship between variables that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting."

**Q12. What is the importance of regression analysis ?**

*Ans :*

1. Regression analysis helps in establishing a functional relationship between two or more variables. Once this is established it can be used for various advanced analytical purposes.
2. Since most of the problems of economic analysis are based on cause and effect relationship, the regression analysis is a highly valuable tool in economics and business research.
3. This can be used for prediction or estimation of future production, prices, sales, investments, income, profits and population which are indispensable for efficient planning of an economy and are of paramount importance to a businessman or an economist.
4. Regression analysis is widely used in statistical estimation of demand curves, supply curves, production functions, cost functions, consumption functions, etc. Economists have discovered many types of production functions by fitting regression lines to input and output data.

### 5.5.1 Simple Linear Regression (for 2 variables)

**Q13. Define Simple Linear Regression.**

*Ans :*

Linear regression is a form of regression which is used for modeling the relationship between scalar variables like X and F under linear regression, linear functions are used to model the data and the unknown parameters, of models are estimated from the data. Hence, these models are known as linear models.

Linear models more commonly refers to those models, where the conditional mean of variable 'F for a given value of variable X will be an affine function of X. A linear regression may also refer to a model, where median or other quantile of the conditional distribution of 'F for a given value of 'X is termed as linear function of X. Similar, to all

types of regression analysis, linear regression also aims on the conditional probability distribution of ' F for a given 'X, instead of joint probability distribution of 'F and X.

**Q14. Differentiate between Correlation and Regression.**

**(OR)**

**What are the differences between Correlation and Regression.**

*Ans :*

**(Imp.)**

| S.No. | Basis for Comparison | Correlation | Regression |
|---|---|---|---|
| 1. | **Meaning** | Correlation is a statistical measure which determines co-relationship or association of two variables. | Regression describe how an independent variable is numerically related to the dependent variable. |
| 2. | **Usage** | To represent linear relationship between two variables. | To fit a best line and estimate one variable on the basis of another variable. |
| 3. | **Dependent and Independent variables** | No difference | Both variables are different |
| 4. | **Indicates** | Correlation coefficient indicates the extent to which two variables move together. | Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (Y). |
| 5. | **Objective** | To find a numerical value expressing the relationship between variables. | To estimate values of random variable on the basis of the values of fixed variable. |

**Q15. What do you mean by line of regression? Derive the equations of lines of regression.**

*Ans :*

**(Imp.)**

In a bi-variate distribution, if the variables are related then the points when plotted in the scatter diagram will lie near a straight line which is called the line of regression and the regression is said to be linear regression. If points lie on some non-linear curve then the regression is said to be curvilinear regression.

**(I) Regression of Y on X.**

The regression equation of Y on X is expressed as follows :

$Y = a + bX$

It may be noted that in this equation 'Y' is a dependent variable, i.e., its value depends on X. 'X is independent variable, i.e., we can take a given value of X and compute the value of Y.

'a' is "Y-intercept" because its value is the point at which the regression line crosses the Y-axis, that is, the vertical axis, 'b' is the "slope" of line. It represents change in Y variable for a unit change in X variable.

'a' and 'b' in the equation are called numerical constants because for any given straight line, their value does not change.

If the values of the constants 'a' and 'b' are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of Least Squares which states that the line should be drawn through the plotted points in such a manner that the sum

of the squares of the deviations of the actual Y values from the computed Y values is the least, or in other words, in order to obtain a line which fits the points best $\Sigma(Y-Y_C)^2$, should be minimum. Such a line is known as the line of 'best fit'.

A straight line fitted by least squares has the following characteristics :

(i) It gives the best fit to the data in the sense that it makes the sum of the squared deviations from the line, $\Sigma(Y-Y_C)^2$, smaller than they would be from any other straight line. This property accounts for the name 'Least Squares'.

(ii) The deviations above the line equal those below the line, on the average. This means that the total of the positive and negative deviations is zero, or $\Sigma(Y-Y_C) = 0$.

(iii) The straight line goes through the overall mean of the data (X, Y).

(iv) When the data represent a sample from a large population the least squares line is a 'best' estimate of the population regression line.

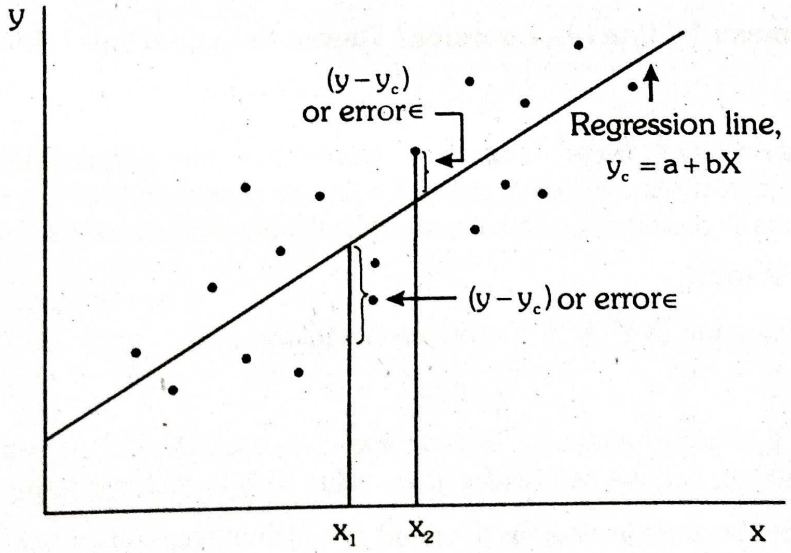
With a little algebra and differential calculus it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters *a* and *b* such that the least squares requirement is fulfilled :

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

These equations are usually called the *normal equations.* In the equations $\Sigma X$, $\Sigma XY$, $\Sigma X^2$ indicate totals which are computed from the observed pairs of values of two variables X and Y to which the least squares estimating line is to be fitted and *N* is the number of observed pairs of values.

This will be shown in figure below.




$\Sigma(y-y_c)$ or $\Sigma \epsilon = 0$

$\Sigma(y-y_c)^2$ or $\Sigma \epsilon^2$ = a minimum

**Figure : Regression of *Y* on X: Least Squares**

If the parameter estimates be *a* for $\alpha$ and *b* for $\beta$, then the line would be

$$Y_c = a + bX$$

Since we seek to minimize $\Sigma(Y-Y_c)^2$, which works out to be $\Sigma(Y-a-bX)^2$, we can find the values of $a$ and $b$ by applying calculus. This results in a pair of what are called normal equations. The normal equations are :

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

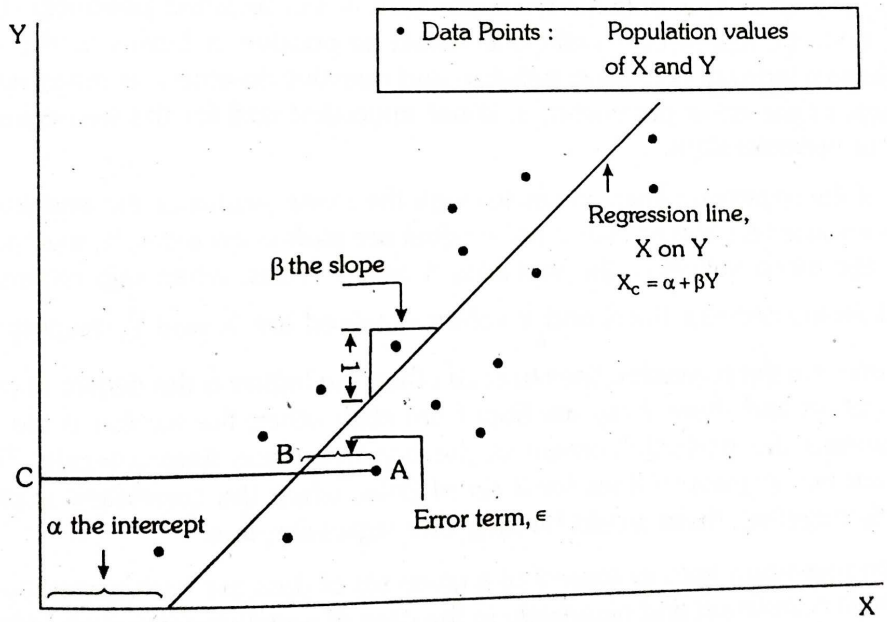
### (II) Regression of X on Y

In general, usually X-variable is taken to be independent and the Y-variable as dependent one. However, if the X-variable is treated as the dependent variable and Y as an independent variable, we can also have regression of X on Y. In the regression of X on Y, the population regression model is :

$$X = \alpha + \beta Y + \varepsilon \text{ or } X_C = a + bY$$

in which X is the dependent variable (the variable to be predicted); Y is the independent variable (the predictor variable); $\alpha$ is the population X-intercept: $\beta$ is the population slope (measured as change in the X variable corresponding to a unit change in Y); and $\varepsilon$ is the error term. Here the Y-variable is fixed and the randomness in the X variable comes from the error term, $\varepsilon$.

The population regression model is shown in Figure below.




For a given pair of X and Y values, represented by a point, say A, the actual value of X equal to AC is composed of the non-random part BC (given by the regression line) and the random component AB.

**Assumptions**

(i) There exists a linear relationship between X and Y variables.

(ii) The values of independent variable Y are fixed while those of dependent variable X are random - with randomness arising from the error term.

(iii) The errors, e, are normally distributed with mean equal to zero, and constant variance a . Further, they are independent in different observations.

Observe here that if X and Y are plotted on the graph on X-axis and Y-axis respectively, then we consider horizontal deviations in the case of regression of X on Y and vertical deviations in the case of regression of Y on X. The estimates of the parameters are given by $a$ and $b$, which are obtained from a pair of normal equations given below :

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

We calculate the input values and substitute them into the above equations, solve them simultaneously to get $a$ and $b$. This yields the regression equation $X = a + bY$. This equation is then used to get the expected values of X for given values of Y.

**Some important points may be noted as follows:**

(i) For a given set of paired data, there are two regression lines - one showing regression of Y on X and the second one showing regression of X on Y. One of these is obtained by minimizing the squared vertical deviations and the other by horizontal deviations. As such, they are different lines with separate parameter values.

(ii) The slope parameters of the regression lines are of particular significance. To distinguish, they are designated as $b_{yx}$ and $b_{xy}$, called regression co-efficient of Y on X and, regression co-efficient of X on Y, respectively.

(iii) For a given set of data, both the regression lines would be either positively sloped or negatively. Thus, both the regression co-efficients would be positive or both would be negative. Positive co-efficients indicate positive correlation and negative co-efficients mean negative correlation. The sign of the other parameter, a, is not important and for the two equations, it may bear same or opposite signs.

(iv) Each of the regression lines passes through the mean values of the variables. When both the regression lines for a given set of paired data are plotted on a graph, their point of intersection yields the mean values of the variables X and Y. Thus, when two regression equations are solved simultaneously, the X and Y values obtained are $\bar{X}$ and $\bar{Y}$, respectively.

(v) The closer are the regression lines to each other, the higher is the degree of correlation between the variables and more away are they from each other, the weaker is the correlation. When the variables are perfectly correlated, the two regression lines coincide. Thus, while usually there are two regression lines for a set of data, when the correlation is perfectly positive or perfectly negative, there would be only one regression line.

The two regression lines in respect of a given set of data are both sloped positively in the case of positive correlation and negatively in the case of negative correlation between the variables.

They intersect at $\bar{X}$ and $\bar{Y}$, and their closeness to each other is indicative of the degree of correlation between the variables.

## PROBLEMS

**5. From the following data obtain the two regression equations and calculate the correlation co-efficient.**

| x | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| y | 18 | 16 | 20 | 24 | 22 | 26 | 28 | 32 | 30 |

**Calculate the value of y when x = 6.2**

(Imp.)

X on Y

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

Y on X

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\bar{X} = \frac{\Sigma x}{n} = \frac{90}{9} = 10$$

$$\bar{Y} = \frac{\Sigma y}{n} = \frac{216}{9} = 24$$

**Calculation of Regression Equation and Correlation Coefficient**

| X | Y | $x - \bar{x}$ x | $y - \bar{y}$ y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|---|---|
| 2 | 18 | –8 | –6 | 64 | 36 | 48 |
| 4 | 16 | –6 | –8 | 36 | 64 | 48 |
| 6 | 20 | –4 | –4 | 16 | 16 | 16 |
| 8 | 24 | –2 | 0 | 4 | 0 | 0 |
| 10 | 22 | 0 | –2 | 0 | 4 | 0 |
| 12 | 26 | 2 | 2 | 4 | 4 | 4 |
| 14 | 28 | 4 | 4 | 16 | 16 | 16 |
| 16 | 32 | 6 | 8 | 36 | 64 | 48 |
| 18 | 30 | 8 | 6 | 64 | 36 | 48 |
| 90 | 216 | 0 | 0 | 240 | 240 | 228 |

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{228}{240} = 0.95$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{228}{240} = 0.95$$

**Equation X on Y**

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 10 = 0.95(X - 24)$$

$$X = 0.95Y - 22.8 + 10$$

$$X = 0.95Y - 12.8$$

**Equation Y on X**

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 24 = 0.95\,(X - 10)$$

$$Y - 24 = 0.95,\ X - 95 + 24$$

$$Y = 0.95\,x - 14.5$$

Regression equation y on x = 6.2

$$y = 0.95x - 14.5$$

$$y = 095(6.2) - 14.5$$

$$y = -8.61$$

**6. Given :**

**$\Sigma x=56$, $\Sigma y=40$, $\Sigma x^2=524$, $\Sigma y^2=256$, $\Sigma xy=364$, $N = 8$**

**(i) Find the two Regression equations and**

**(ii) The Correlation Coefficient.**

***Sol :***

We have

$$\bar{X} = \frac{\Sigma x}{N} = \frac{56}{8} = 7; \quad \bar{y} = \frac{\Sigma y}{N} = \frac{40}{8} = 5$$

byx = co-efficient of regression of y on x $= \dfrac{N\Sigma xy - (\Sigma x)(\Sigma y)}{N(\Sigma y^2) - (\Sigma x)^2}$

$$\frac{8(364) - (56)(40)}{8(524) - (56)^2} = \frac{2912 - 2240}{4192 - 3136} = \frac{672}{1056}$$

$$\boxed{byx = 0.6363}$$

bxy = co-efficient of regression of x on y $= \dfrac{N\Sigma xy - (\Sigma x)(\Sigma y)}{N(\Sigma y^2) - (\Sigma y)^2}$

$$\frac{8(364) - (56)(40)}{8(256) - (40)^2} = \frac{2912 - 2240}{2048 - 1600} = \frac{674}{448} = \boxed{bxy = 1.504}$$

(i) **Two Regression equations**

Regression equation x on y

$$(x - \bar{x}) = b_{xy}(y - 7)$$

$$(x - 7) = 1.504\,(y - 5)$$

$$(x - 7) = 1.504\,(y) - 1.504\,(5)$$

$$x = 1.504\,(y) - 7.522 + 7$$

$$x = 1.504\,(y) - 0.522 \ldots\ldots (1)$$

Regression equation y on x

$(y-\bar{y}) = b_{yx}(x-\bar{x})$

$(y-5) = 0.6363\,(x-7)$

$y-5 = 0.6363\,(x) - 0.6363(7)$

$y = 0.6363\,(x) - 4.4541+5$

$y = 0.6363\,(x) + 0.5459$ .....(2)

(ii) **The correlation co-efficient $\gamma_{xy}$ between x and y is given by**

$\gamma_{xy}^2 = b_{yx}.b_{xy} = (0.6363)(1.504)$

$r_{xy}^2 = 0.9569$

$\boxed{\gamma_{xy} = 0.9782}$

**7. Following are the marks in Statistics and English in an Annual Examination.**

| Particular | Statistics (X) | English (Y) |
|---|---|---|
| Mean | 40 | 50 |
| Standard Derivation | 10 | 16 |
| Co-efficient Correlation | 0.5 | |

**(i) Estimate the score of English, when the score in Statistics is 50.**

**(ii) Estimate the score of statistics, when the score in English is 30.**

*Sol :*

**(Imp.)**

Given mean of X denoted as $\bar{X} = 40$.

Given mean of Y denoted as $\bar{Y} = 50$.

SD of X denoted as $\sigma_x = 10$.

SD of Y denoted as $\sigma_y = 16$.

Coefficient of correlation denoted as $r = 0.5$

**Regression Equation X on Y**

$$[X-\bar{X}] = [r]\left[\frac{\sigma x}{\sigma y}\right][y-\bar{y}]$$

$$X-40 = [0.5]\left[\frac{10}{16}\right][Y-50]$$

$X-40 = [0.5]\,[0.625]\,[Y-50]$

$X-40 = [0.3125]\,[Y-50]$

$X-40 = 0.3125y - 15.625$

$$\bar{X} = 0.3125y - 15.625 + 40$$
$$X = 0.3125y + 24.375$$

**Regression Equation Y on X**

$$[Y - \bar{Y}] = [r]\left[\frac{\sigma y}{\sigma x}\right][X - \bar{X}]$$

$$[Y - 50] = [0.5]\left[\frac{16}{10}\right][X - 40]$$

$$Y - 50 = [0.5]\,[1.6]\,[X - 40]$$
$$Y - 50 = (0.8)\,(X - 40)$$
$$Y - 50 = 0.8X - 32$$
$$Y = 0.8\,X - 32 + 50$$
$$Y = 0.8\,X + 18$$

Estimation of English (Y) when Statistics (X) is 50

$$Y = 0.78\,X + 18$$
$$= 0.8\,(50) + 18$$
$$= 40 + 18$$
$$\therefore Y = 18 \text{ marks.}$$

Estimation of statistics (X) when English (Y) is 30

$$X = 0.3125\,Y + 24.375$$
$$= 0.3125(30) + 24.375$$
$$= 9.375 + 24.375$$
$$X = 33.75 \text{ marks.}$$

## 5.6 Basic Definitions of Testing of Hypothesis

**Q16. Define Hypothesis. What are the characteristics of Hypothesis ?**

*Ans :* **(Imp.)**

**Introduction**

The term 'hypothesis' is derived from the ancient Greek word, 'hypothesis' that means 'to put under' (or) 'to suppose'. Hypothesis is also a combination of two words 'Hypo, Thesis where 'Hypo' means tentative or subject to verification and 'Thesis' a statement based on concepts, theories and past experiences about the solution of the problem.

The term hypothesis literally means an assumption or a supposition about the state of affairs of a certain thing or phenomena or facts or variable or situation. Thus, "hypothesis is perceived as a proposition or set of propositions set forth as an explanation for occurrence of some specified group of phenomenon either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts. Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variables"

**Definitions**

The view point of various thinkers has been presented as under :

1. **According to Good, Barr and Scates** "Hypothesis is a statement temporarily accepted as true in the light of what is at the time, known about a phenomenon, and it is employed as a basis for action in the search for new truth. When hypothesis is fully established, it may take the form of facts, principles and theories".
2. **According to Lundberg** "Hypotheses is a tentative generalization, the validity of which is remains to be tested. In the most elementary stage the hypothesis may be any hunch, guess, imaginative idea which become base for further investigation".
3. **According to Best** "Hypothesis id a shrewd guess or inference that is formulated and provisionally adopted to explain observed facts or conditions and to guide in further investigation".
4. **According to Mouly** "Hypothesis is an assumption whose testability is to be tested on the basis of the compatibility of its implications with empirical evidences and previous knowledge".
5. **According to Gopal** "Hypothesis is a tentative solution posed on cursory observation of known and available data and adopted provisionally to explain certain events and to guide in the investigation of others. It is in fact, a possible solution to the problem".

6. **According to Theodorson and Theodorson** "A hypotheses is a tentative statement asserting a relationship between certain facts".

7. **According to Goode and Hutt** "A proposition which can be put to a test to determine its validity".

8. **According to Kerlinger** "A hypothesis is a conjectural statement of the relationship between two or more variables".

9. **According to Black and Champion** "A hypotheses is a tentative statement about something, the validity of which is usually unknown".

10. **According to While Bailey** "Hypothesis is a proposition that is stated in a testable form and that predicts a particular relationship between two or more variables".

11. **According to Grinnell** "Hypothesis is written in such a way that it can be proven or disproved by valid or reliable data it is in order to obtain these data that we perform our study".

12. **According to Palmar O Johnson:** "A hypothesis in statistics is simply a quantitative statement about a population".

13. **According to Webster :** "Hypothesis is a tentative assumption made in order to draw out and test its logical or empirical consequences".

**Characteristics**

The following the basic characteristics of a hypothesis :

1. **Valid**

Hypothesis must be valid and related to the phenomena or situation which it is trying to explain.

2. **Pivot of Research**

Hypothesis is backbone of all kinds of researches because the all research activities are designed to verify the hypothesis from 360 degree angle.

3. **Conceptual Clarity**

Hypothesis must be clearly and precisely stated. There should be no ambiguity in the formulation of hypothesis. It means hypothesis should be defined lucidly, should be operationalsed, should be commonly accepted and should be communicable.

4. **Testability**

A hypothesis should be testable and not moral judgement. It should be possible to collect empirical evidences to test the hypothesis. In the words of C. William Emory, "A hypothesis is testable if other deductions can be confirmed or disapproved by observation".

5. **Specificity**

A hypothesis should be specific and explain the expected the relations between variables and the situations under which these relation will hold.

6. **Consistency**

A hypothesis should be logically consistency. Two or more hypothesis logically derived from the same population must not be mutually contradictory.

7. **Objectivity**

Hypothesis should be free from value judgement. In the scientific research the researcher's value system has no place in scientific enquiry.

8. **Simplicity**

Hypothesis should be a simple one requiring fewer assumptions. But the simplicity does not mean vague idea.

9. **Theoretical Relevance**

Hypothesis should be based upon some theoretical foundations. When a research is systematically based upon a body of existence knowledge, only then a genuine contribution is more likely to result in.

10. **Availability of Technique**

Hypothesis should be related to available techniques, otherwise it will not be

researchable. Hence, the researcher must ensure that statistical or mathematical techniques are available for testing the proposed hypothesis.

11. **Future Oriented**

Hypothesis is forward looking concept as it is related to future verification not the past facts, information or situations.

**Q17. Explain the procedure for testing a hypothesis.**

**(OR)**

**Explain the procedure generally followed in testing of hypothesis.**

*Ans :* **(Imp.)**

Test of Hypothesis involves the following steps :

**Step 1: Statement (or assumption) of hypothesis**

There are two types of hypothesis :

(i) Null Hypothesis

(ii) Alternative Hypothesis.

**(i) Null Hypothesis :**

For applying the tests of significance, we first set up a hypothesis a definite statement about the population parameter. Such a hypothesis is usually a hypothesis of no-difference, is called Null Hypothesis.

It is in the form $H_0 : \mu = \mu_0$

$\mu_0$ is the value which is assumed or claimed for the population characteristic. It is the reference point against which the Alternative Hypothesis is set up, as explained in the next step

**Definition**

A null hypothesis is the hypothesis which asserts that there is no significant: difference between the statistic and the population parameter and whatever observed difference is there, is merely due to fluctuations in sampling from the same population. It is always denoted by $H_0$ . To test whether one procedure is better than another, we assume that there is no difference between the procedures. Similarly to test whether there is a relationship between two variates , we take $H_0$ that there is no relationship.

**For example,** in case of a single statistic, $H_0$ will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics ($H_0$) will be that the sample statistics do not differ significantly.

**(ii) Alternative Hypothesis**

Any hypothesis which contradicts the Null Hypothesis is called an Alternative Hypothesis, usually denoted by $H_1$. The two hypothesis $H_0$ and $H_1$ are such that if one is true, the other is false and vice versa.

**For example,** if we want to test the null hypothesis that the population has a specified mean $\mu_0$ (say) i.e., $H_0 : \mu = \mu_0$, then the Alternative Hypothesis would be

(i) $H_1 : \mu \neq \mu_0$(i e., either $\mu > \mu_0$ (or) $\mu < \mu_0$)

(or)

(ii) $H_1 : \mu > \mu_0$

(or)

(iii) $H_t : \mu < \mu_0$

The Alternative Hypothesis

(i) is known as a two-tailed alternative and the Alternative Hypothesis in

(ii) is known as right-tailed and in

(iii) is known as left-tailed.

The setting of alternative hypothesis is very important to decide whether we have to use a single-tailed (right or left) or two-tailed test.

Alternate Hypothesis is in one of the following forms :

$H_1 : \mu \neq \mu_0$

or $H_1 : \mu > \mu_0$

or $H_1 : \mu < \mu_0$

**Step 2 : Specification of the Level of Significance**

The level of significance denoted by a is the confidence with which we rejects or accepts the Null hypothesis $H_0$ i.e., it is the maximum possible probability with which we are willing to risk an error in rejecting $H_0$ when it is true. The level of significance is generally specified before a test

procedure so that the results obtained may not influence our decision. In practice, we take either 5% (i.e., .05) or 1% (i.e., 0.01) level of significance, although other levels such as 2%, 1/2% etc. may also be used. 5% Level of significance in a test procedure indicates that there are about 5 cases in 100 that we would reject the null hypothesis $H_0$ when it is true i.e., we are about 95% confident that we have made the right decision. Similarly, in 1% Level of significance, there is only 1 case in 100 that the null hypothesis $H_0$ is rejected when it is true i.e., we are about 99% confident that we have made the right decision. Level of significance is also known as the size of the test.

**Step 3 : Identification of the Test Statistic**

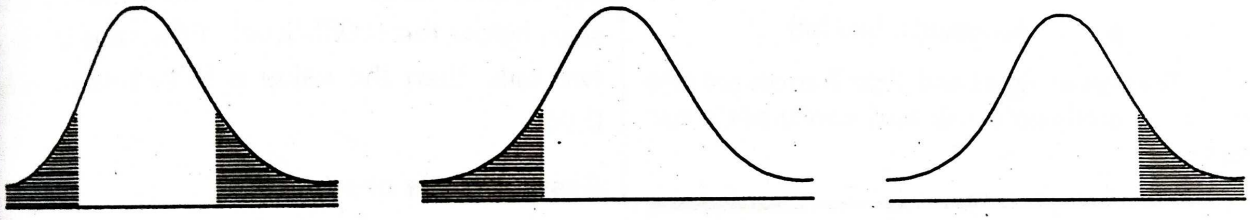
There are several tests of significance, viz., z,t,F etc. First we have to select the right test depending on the nature of the information given in the problem. Then we construct the test criterion and select the appropriate probability distribution.

**Step 4: Critical Region**

The critical region is formed based on following factors.

(a) Distribution of the Statistic i.e., whether the statistic follows the normal, 't' $\chi^2$ (or) 'F' distribution (will be discussed later).

(b) Form of Alternative Hypothesis: If the form has $\neq$ sign, the critical region is divided equally in the left and right tailes, sides of the distribution.



Two sided Left sided Right sided

If the form of alternative hypothesis has < sign, the entire critical region is taken in the left tail of the distribution.

If the form of alternative hypothesis has > sign, the entire critical region is taken on the right side of the distribution.

**Step 5 : Making Decision**

We compute the value of the appropriate statistic and ascertain whether the computed value falls in acceptance or rejection region depending on the specified Level of significance.

In finding the acceptance or rejection region we have to use critical values given in Statistical Tables. By comparing the computed value and the critical value decision is taken for accepting or rejecting $H_0$. If the computed value < critical value, we accept $H_0$, otherwise we reject $H_0$.

**Q18. What are the errors of sampling ?**

*Ans :*

The main objective in sampling theory is to draw valid inferences about the on parameters on the basis of the sample results. In practice we decide to accept or to reject the lot after examining a sample from it. As such we have two types of errors.

(i) **Type I error**

Reject $H_0$ when it is true.

If the Null Hypothesis $H_0$ is true but it rejected by test procedure, then the error made is called Type I error or a error.

**(ii) Type II error**

Accept $H_0$ when it is wrong i.e., accept $H_0$ when $H_1$ is true. If the Null Hypothesis is false but it is accepted by test, then error committed is called Type II error or $\alpha$ error.

If we write

P (Reject $H_0$ when it is true) = P (Type I error) = $\alpha$

and P (Accept $H_0$ when it is wrong) = P (Type II error) = $\beta$

then a and $\beta$ are called sizes of Type I and Type II errors respectively

i.e., $\alpha$ = P (Rejecting a good lot)

$\beta$ = P (Accepting a bad lot)

The sizes of Type I and Type II errors are also known as producer's risk and consumer's risk respectively.

## 5.7 Small Sample Tests

### 5.7.1 t-Test

**Q19. What is Small sample test ?**

*Ans :*

**Meaning**

Small sample size referred to size of sample which is less than 30. In case of small sample size the z-test is not appropriate test statistic as the assumptions on which it is based do not hold good in case of small sample. The theoretical work on t-distribution was done by W.S. Gosset (1876-1937) under the pen name "student' as he was the employee of the company Guinness & Sons, a Dublin bravery, Ireland, which did not allowed it employees to publish research findings under their own names. The t-distribution is used when sample size is less than 30 and the population standard deviation is not known.

**Q20. Explain briefly about T-test. State the assumptions of T-test.**

*Ans :*

When the size of the sample is small i.e., less than 30, the Z-tests using normal distribution are not applicable because the assumptions on which they are based generally do not hold good in case of small samples. The sampling distribution of small samples follow student's t-distribution. The student's t-distribution has a greater dispersion than the standard normal distribution. As 'n' gets larger, the t-distribution approaches the normal form.

**Degree of Freedom:** Degree of freedom is used to see the table value for testing the hypothesis as V = n –1. If hypothesis is to be tested 5% level of significance under one tail, then the value is to be seen below the 0.025 level. If the value of table is two tails, then the value is to be seen below the 0.05.

**Assumptions of t-test**

The following are the pre-requisites for the application of t-test :

1. The population from which a sample is drawn is normal.
2. The samples have been drawn at random.
3. The population standard deviation is not known.
4. Sample size should small i.e., less than 30.

**Q21. What are the properties of t-distribution?**

*Ans :*

1. The shape of t-distribution is bell-shaped, which is similar to that of a normal distribution and is symmetrical about the mean.
2. The t-distribution curve is also asymptotic to the t-axis, i.e., the two tails of the curve on both sides of t = 0 extends to infinity.

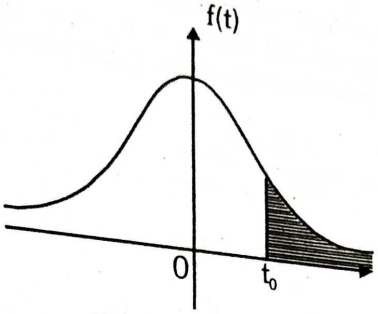



**Fig.: t-distribution**

3. It is symmetrical about the line $t = 0$.
4. The form of the probability curve varies with degrees of freedom i.e., with sample size.
5. It is unimodel with Mean = Median = Mode.
6. The mean of standard normal distribution and as well as t-distribution zero but the variance of t-distribution depends upon the parameter v which is called the degrees of freedom.

### 5.7.2 t-test for single Mean, t-test for difference of Means

**Q22. Explain the test concerning the significance of single and two mean**

***Ans :***

**(Imp.)**

**T-test for Single Mean**

(i) If a random sample x. of size n has been drawn from a normal population with a specified mean p.

(ii) If the sample mean differs significantly from the hypothetical value p the population mean.

In this case the statistic is given by $t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$ where $\bar{x}$, $\mu$, S, n have usual meanings.

Let a random sample of size n ($n < 30$) has a sample mean $\bar{x}$. To test the hypothesis that the population mean p has a specified value $p_0$ when population S. D. $\sigma$ is not known.

Let the Null Hypothesis be $H_0$: $\mu = \mu_0$

Then the Alternative Hypothesis is $H_1$ : $\mu \neq \mu_0$

Assuming that $H_0$ is true, the test statistic given by $t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$, where s is the sample S. D. follows t-distribution with $v = (n - 1)$ d.f.

We calculate the value of 11 \ and compare this value with the table value of t $\alpha$ level of significance. If the calculated value of t >the table value of t, we reject $H_0$ at $\alpha$ level. Otherwise we accept $H_0$.

In this case, 95% confidence limits for the population mean $\mu$ are $\bar{x} \pm t_\alpha \cdot \frac{s}{\sqrt{n-1}}$ where $\alpha = 0.025$ for two-tailed test and 5 = sample S. D. and 99% confidence limits $\mu$ are $\bar{x} \pm t_\alpha \cdot \frac{s}{\sqrt{n-1}}$ where $\alpha = 0.05$.

For a two-tailed test at $\alpha$ level of significance, value of $\alpha / 2$ is taken for $\alpha$.

The more common situation involving tests on two means are those in which variances are unknown. If we assume that distributions are normal and that $\sigma_1 = \sigma_2 = \sigma$. The pooled t-test (often called the two-sample t-test) may be used. The test statistic is given by the following test procedure,

$$t = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ (or) } t = \frac{\bar{X} - \bar{Y}}{S^2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

With $n_1 + n_2 - 2$ degrees of freedom.

Where,

$$S^2 = \frac{\Sigma(X_i - \bar{X}) + \Sigma(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

Where $\bar{x}_1$, $\bar{x}_2$ are the means of two samples of size $n_1$ and $n_2$.

The critical region with this t-distribution can be obtained in a similar war.

**For example,** when A.H is $\mu_1 - \mu_2 \neq \delta$, the null hypothesis $(H_0)$ is not rejected when, $-t_{\frac{\alpha}{2},n_1+n_2-2} < t < t_{\frac{\alpha}{2},n_1+n_2-2}$ and the critical region is $t < -t_{\frac{\alpha}{2},n_1+n_2-2}$ (or) $t_{\frac{\alpha}{2},n_1+n_2-2}$

Critical region for testing $H_0 : \mu_1 - \mu_2 = \delta$

Alternate hypothesis; Reject null hypothesis if ( ),

(i) $\mu_1 - \mu_2 \neq \delta$ $\quad t < -t_{\alpha/2}$ (or) $t > t_{\alpha/2}$

(ii) $\mu_1 - \mu_2 > \delta$ $\quad t > t_\alpha$

(iii) $\mu_1 - \mu_2 < \delta$ $\quad t < -t_\alpha$

**Note :**

1. The two-sample t-test can not be used if $\sigma_1 \neq \sigma_2$.
2. The two-sample t-test can not be used for 'before and after' kind of data, where the is naturally paired.

   In other words the samples must be 'independent' for two sample t-test.

## PROBLEMS

**8. A mechanist making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a S.D. of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specification at level of significance.**

***Sol :***

Here the sample size n – 10 < 30

Hence the sample is small sample.

Also sample mean $\bar{x} = 0.742$ inches , the population mean $\mu = 0.700$ inches and S.D.= 0.040 inches are given.

$\therefore$ We use student's t-Test

(i) **Null Hypothesis $H_0$ :** The product is confirming to specification.

(ii) **Alternative Hypothesis $H_1$ :** $\mu \neq 0.700$

(iii) **Level of significance is,** $\alpha = 0.05$

(iv) **The test statistic is,** $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}}$

Here $\bar{x} = 0.742$ inches, $\mu = 0.700$ inches, S.D. $= 0.040$ inches and $n = 10$. Degrees of freedom (d.f) = w- l = 10 – l = 9

$$\therefore \quad t = \frac{0.742 - 0.700}{\frac{0.040}{\sqrt{10-1}}} = 3.15$$

$\therefore$ The calculated value of t = 3.15

The tabulated value of t at 5% level with 9 degrees of freedom is $t_{005} = 2.26$

Since calculated value of t > tabulated value of t, therefore, $H_0$ is rejected.

$\therefore$ The product is not meeting the specification.

---

**9. A sample 26 bulbs gives a mean life of 990 hours with a S.D. of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not upto the standard.**

*Sol :*

Here sample size, $n = 26 < 30$

$\therefore$ The sample is small sample.

Also given, sample mean, $\bar{x} = 990$

Population mean, $\mu = 1000$ and S.D., $s = 20$

Degrees of freedom $= n - 1 = 26 - 1 = 25$

Here we know x, p, S.D. and n.

$\therefore$ We use students 't' test.

(i) **Null Hypothesis $H_0$ :** The sample is upto the standard.

(ii) **Alternative Hypothesis $H_1$ :** $\mu < 1000$

(The sample is below standard) (left-tail test)

(iii) **Level of significance :** $\alpha = 0.05$

(iv) **The test statistic is** $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}} = \dfrac{990 - 1000}{20/\sqrt{25}} = -2.5$

$\therefore \quad |t| = 2.5$

i.e., Calculated value of t = 2.5

Tabulated value of 't' at 5% level with 25 degrees of freedom for left-tailed test is 1.708.

Since calculated t > tabulated t, we reject the null hypothesis $H_0$ and conclude: that the sample is not upto the standard.

**10. A machine is designed to produce insulating washers for electrical device of average thickness of 0.025 cm. A random sample of 10 washers was found to have a thickness of 0.024 cm with a S.D of 0.002 cm. Test the significance of the deviation. Value of t for 9 degrees of freedom at 5% level is 2.262.**

***Sol :***

Here the sample size is $10 < 30$

$\therefore$ The sample is small

Also given Sample mean, $\bar{x} = 0.024$ cm

Population mean, $\mu = 0.025$ cm

S.D. = 0.002 cm

Degrees of freedom (d.f) $= n - 1 = 10 - 1 = 9$

**(i) Null Hypothesis $H_0$ :** The difference between $\bar{x}$ and $\mu$ is not significant.

**(ii) Alternative Hypothesis $H_1$ :** $\mu_1 \neq 0.025$

**(iii) Level of significance :** $\alpha = 0.05$

**(iv) The test statistic is** 't' $= \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}} = \dfrac{0.024 - 0.025}{\dfrac{0.002}{\sqrt{10-1}}} = -1.5$

$\Rightarrow |t| = 1.5$

$\therefore$ Calculated value of t = 1.5 for two tailed test.

Tabulated value of t for 9 degrees of freedom at 5% level = 2.262

Since calculated t < tabulated t, we accept the null hypothesis and conclude that the difference between x and p is not significant.

**11. Two different types of drugs A and B were tried on certain patients for increasing weight, 5 persons were given drug A and 7 persons were given drug B. The increase in weight (in pounds) is given below**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Drug A** | **8** | **12** | **13** | **9** | **3** | | |
| **Drug B** | **10** | **8** | **12** | **15** | **6** | **8** | **11** |

**Do the drugs differ significantly with regard to their effect in increasing weight ?**

***Sol :***

**(Imp.)**

Let the weights (in kgs) of the patients treated with drugs A and B be denoted by suitable variances X and Y respectively.

We set up the null hypothesis, $H_0 : \mu_x = \mu_y$ i.e., there is no significant difference between the drugs A and B with regard to their effect on increase in patients weight.

Alternative hypothesis, $H_1 : \mu_x \neq \mu_y$

Under $H_0$, the appropriate test statistic is,

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Degree of freedom (d.f) = $t_{n_1+n_2-2}$

**Computation of Sample Means and Standard Deviations**

| X | $(X - \bar{X})$ $(X - 9)$ | $(X - \bar{X})^2$ | Y | $(Y - \bar{Y})$ $(Y - 10)$ | $(Y - \bar{Y})^2$ |
|---|---|---|---|---|---|
| 8 | –1 | 1 | 10 | 0 | 0 |
| 12 | 3 | 9 | 8 | – 2 | 4 |
| 13 | 4 | 16 | 12 | 2 | 4 |
| 9 | 0 | 0 | 15 | 5 | 25 |
| 3 | – 6 | 36 | 6 | – 4 | 16 |
| | | | 8 | – 2 | 4 |
| | | | 11 | 1 | 1 |
| **$\Sigma X = 45$** | **$\Sigma(X - \bar{X}) = 0$** | **$\Sigma(X - \bar{X})^2 = 62$** | **$\Sigma Y = 70$** | **0** | **54** |

Here, $n_1 = 5, \Sigma X = 45, \Sigma(X - \bar{X})^2 = 62$

$$\bar{X} = \frac{\Sigma X}{n_1} = \frac{45}{5} = 9$$

$n_2 = 7, \Sigma Y = 70, \Sigma(Y - \bar{Y})^2 = 54$

$$\bar{Y} = \frac{\Sigma Y}{n_2} = \frac{70}{7} = 10$$

and $$S^2 = \frac{1}{n_1 + n_2 - 2}[\Sigma (X - \bar{X})^2 + \Sigma(Y - \bar{Y})^2$$

$$= \frac{1}{5+7-2}[62 + 54]$$

$$= \frac{1}{10}[116]$$

$$S^2 = \frac{116}{10} = 11.6$$

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{9 - 10}{\sqrt{11.6 \times \frac{12}{35}}}$$

$$= \frac{-1}{\sqrt{3.98}}$$

$$= \frac{-1}{1.99}$$

Degree of freedom (df) $= t_{n_1+n_2-2}$

$= t_{5+7-2}$

$= t_{10}$

Hence, tabulate value of t for 10 df at 5% level of significance for the two tailed test is 2.228. Thus, calculated t = – 0.50, is less than tabulated value of t (i.e., 2.228).

Therefore, null hypothesis $H_0$ is accepted at 5% level of significance and we may conclude that the drugs A and B do not differ significantly with regard to their effect on increase in patiets weights.

### 5.7.3 Paired t-test

**Q23. Discuss in detail about Paired t-test.**

**(Imp.)**

*Ans :*

Paired observations arise in many practical situations where each homogeneous experimental unit receives both population conditions. As a result, each experimental unit has a pair of observations, one for each population.

For instance, to test the effectiveness of "drug" some 11 persons blood pressure is measured "before" and "after" the intake of certain drug. Here the individual person is the experimental unit and the two populations are blood pressure "before" and "after" the drug is given. Thus for each observation in one sample, there is a corresponding observation in the other sample pertaining to the same character. Hence the two samples are not independent.

Consider another example. Suppose a business concern is interested to know whether a particular media of promoting sales of a product is really effective or not. In this case we have to test whether the average sales before and after the sales promotion are equal.

If $(x_1, y_1)$, $(x_2, y_2)$,.....,$(x_n, y_n)$, be the pairs of sales data before and after the sales production in a business concern, we apply paired t - test to examine the significance of the difference of the two situations.

Let $d_i = x_i - y_i$ (or) $y_i - x_i$ for i = 1,2,3,...., n

Let the Null Hypothesis be $H_0 : \mu_1 = \mu_2$ (i.e., $\mu = 0$), there is no significant difference between the means in two situations.

Then the Alternative Hypothesis is $H_1 : \mu_1 = \mu_2$

Assuming that $H_0$ is true, the test statistic for n paired observations (which are dependent) by taking the differences $d_1, d_2, \ldots, d_n$ of the paired data.

$$t = \frac{\bar{d} - \mu}{S/\sqrt{n}} = \frac{\bar{d}}{S/\sqrt{n}} \ (\because \mu = 0)$$

where $t = \bar{d} = \frac{1}{n}\Sigma d_1$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$

or $S^2 = \frac{\sum d^2 - n(\bar{d})^2}{n-1}$ or $\frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$

are the mean and variance of the differences $d_u d_2, \ldots, d_n$ respectively and p is the are of the population of differences.

The above statistic follows student's t-distribution with (n – 1) degrees of freedom.

## PROBLEMS

**12. Ten workers were given a training programme with a view to study then assembly time for a certain mechanism. The results of the time and motion studies before and after the training programme are given below.**

| Workers | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 15 | 18 | 20 | 17 | 16 | 14 | 21 | 19 | 13 | 22 |
| $Y_1$ | 14 | 16 | 21 | 10 | 15 | 18 | 17 | 16 | 14 | 20 |

**$X_1$ = Time taken for assembling before training,**

**f, = Time taken for assembling after training.**

**Test whether there is significant difference in assembly times before and after training**

***Sol :***

From the given paired data, we see that we are to use paired t-test. Let p be the mean of population of differences.

(i) **Null Hypothesis $H_0$ :** $\mu_1 = \mu_2$ or $\mu = 0$ i.e., training is not useful.

(ii) **Alternative Hypothesis $H_1$:** $\mu_1 \neq \mu_2$ i.e., training is useful in assembly time.

(iii) **Level of significance** $\alpha = 0.05$

(iv) **Computation :** Differences $d_i$'s (before and after training) are

1, 2, –1, 7, 1, –4, 4, 3, –1, 2

$\bar{d}$ = mean of differences of sample data = $\frac{\sum d}{n} = \frac{14}{10} = 1.4$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_1 - \bar{d})^2$$

$$= \frac{1}{9}[(1-1\ 4)^2 + (2-1.4)^2 + (-1-1.4)^2 + (7-1.4)^2 + (1-1.4)^2 + (-4-1.4)^2$$
$$+ (4-1\ 4)^2 + (3-1\ 4)^2 + (-1-1\ 4)^2 + (2-1\ 4)^2]$$

$$= \frac{1}{9}[0.16 + 0.36 + 5.76 + 31.36 + 0.16 + 29{,}16 + 6.76 + 2.56 + 5.76 + 0.36]$$

$$= \frac{82.4}{9} = 9.1555$$

$$\therefore \quad S = 3.026$$

**(v) The test statistic** is $t = \dfrac{\bar{d}-\mu}{S/\sqrt{n}} = \dfrac{\bar{d}}{S/\sqrt{n}} = \dfrac{1.4}{3.026/\sqrt{10}} = \dfrac{(1.4)(3.163)}{3.026} = 1.46$

$\therefore$ Calculated $|t| = 1.46$

Tabulated $t_{0.05}$ with $10 - 1 = 9$ degrees of freedom is 1.833

Since calculated $t < t_{005}$, we accept the Null Hypothesis $H_0$ and conclude that there is no significant difference in assembly times before and after training.

**13. Scores obtained in a shooting competition by 10 soldiers before and after intensive training are given below :**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Before** | **67** | **24** | **57** | **55** | **63** | **54** | **56** | **68** | **33** | **43** |
| **After** | **70** | **38** | **58** | **58** | **56** | **67** | **68** | **75** | **42** | **38** |

**Test whether the intensive training is useful at 0.05 level of significance.**

***Sol :***

Let us apply paired test.

Let p be the mean of population of differences.

**(i) Null Hypothesis $H_0$ :** $\mu_1 = \mu_2$ i.e., $\mu = 0$ there is no significant effect of the training.

**(ii) Alternative Hypothesis $H_1$:** $\mu_1 \neq \mu_2$ intensive training is useful.

**(iii) Level of significance,** $\alpha = 0.05$

**(iv) Computation :**

Differences $d_i$'s (before and after training) are –3, –14, –1, –3, 7, –13, –12, –7, –9, 5

$$\bar{d} = \frac{1}{10}\sum_{i=1}^{10} d_i = \frac{-50}{10} = -5$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{10}(d_i - \bar{d})^2$$

$$= \frac{1}{9}[(2)^2 + (-9)^2 + (4)^2 + (2)^2 + (12)^2 + (-8)^2 + (-7)^2 + (-2)^2 + (-4)^2 + (10)^2]$$

$$= \frac{1}{9}[4 + 81 + 16 + 4 + 144 + 64 + 49 + 4+16 + 100]$$

$$= \frac{482}{9} = 53.5555$$

$$\therefore S = 7.32$$

(v) **The test statistic is** $t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{-5-0}{7.32/\sqrt{10}} = 2.16$

Tabulated $t_{0.05}$ with $10 - 1 = 9$ degrees of freedom is 1.83.

Since calculated t > t tabulated t, we reject the Null Hypothesis and conclude that the intensive training is useful.

## 5.8 F-TEST FOR EQUALITY OF TWO POPULATION VARIANCES

**Q24. Define F-test.**

*Ans :*

F test is a statistical test that is used in hypothesis testing to check whether the variances of two populations or two samples are equal or not. In an f test, the data follows an f distribution. This test uses the f statistic to compare two variances by dividing them. An f test can either be one-tailed or two-tailed depending upon the parameters of the problem.

The f value obtained after conducting an f test is used to perform the one-way ANOVA (analysis of variance) test. In this article, we will learn more about an f test, the f statistic, its critical value, formula and how to conduct an f test for hypothesis testing.

**Q25. Explain F-test for equality of two variances.**

*Ans :* **(Imp.)**

The F Test for Equality of Variances between two groups is a statistical test used to compare the variances of two samples to determine whether they are equal. It is based on the assumption that the samples are drawn from normally distributed populations.

**Steps in the F Test for Equality of Two Variances**

1. **Specify the null and alternative hypotheses.** The null hypothesis is usually that the variances of the two samples are equal, while the alternative hypothesis is that the variances of the two samples are not equal.
2. **Select two samples** from the populations and calculate the sample variances and sizes.
3. **Calculate the test statistic,** which is the ratio of the larger sample variance to the smaller sample variance.
4. **Determine the critical value of the test statistic** based on the significance level (alpha) of the test and the degrees of freedom for the numerator and denominator. The degrees of freedom for the numerator and denominator are calculated as the sample sizes minus 1.
5. **Compare the calculated test statistic to the critical value** to determine whether to reject or fail to reject the null hypothesis. If the calculated test statistic exceeds the critical value, the null hypothesis is rejected, and the alternative hypothesis is accepted.

**Conditions for the F Test for Equality of Two Variances**

To conduct a valid F test for the equality of two variances, the following conditions must be met:

1. The samples must be drawn randomly from the populations.
2. Each observation in each sample must be independent of the others.
3. The population distributions must approximate a normal distribution.

**Typical Null and Alternate Hypotheses in the F Test for Equality of Two Variances:**

The null hypothesis in an F test for equality of two variances is that the variances of the two samples are equal. This can be expressed as:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Where

$\sigma_1^2$ is the variance of the first sample and $\sigma_2^2$ is the variance of the second sample.

The alternate hypothesis is the opposite of the null hypothesis and is that the variances of the two samples are unequal. This can be expressed as :

$$Ha : \sigma_1^2 \neq \sigma_2^2$$

**Calculating Test Statistic:**

The F statistic in an F test for the equality of two variances is calculated as the ratio of the larger sample variance to the smaller sample variance. It is used to determine whether the difference between the two sample variances is statistically significant. The formula for calculating the F statistic is as follows:

$$F = \frac{s_1^2}{s_2^2}$$

Where

$S_1^2$ is the variance of the first sample and

$S_2^2$ is the variance of the second sample.

**Calculating Critical Values:**

The critical values for the F statistic in an F test for equality of two variances depend on the significance level of the test and the degrees of freedom for the numerator and denominator. The degrees of freedom for the numerator and denominator are calculated as the sample sizes minus 1. Using these two values (significance level and degrees of freedom), you can find out the value of the critical F statistic using F tables.

## PROBLEM

**14. In a laboratory experiment, two samples gave the following results :**

| Sample | Size | Samle Mean | Sum of Squares of Deviation from the Mean |
|---|---|---|---|
| 1 | 10 | 15 | 90 |
| 2 | 12 | 14 | 108 |

**Test the equality of sample variances at 5% level of significance.**

*Sol :*

Null Hypothesis : $H_0 : \sigma_1^2 = \sigma_2^2$

Variance for sample 1 $(S_1^2) = \frac{\Sigma(X-\bar{X})^2}{n_1-1} = \frac{90}{10-1} = \frac{90}{9} = 10$

Vairnace for sample 2 $(S_2^2) = \frac{\Sigma(X-\bar{X})^2}{n_2-1} = \frac{108}{12-1} = \frac{108}{11} = 9.82$

**Test Statistic :** $F = \frac{S_1^2}{S_2^2} = \frac{10}{9.82} = 1.018$

**Critical Value :** For $v_1$ (10 – 1) 9 and $v_2$ (12 – 1) 11 and at 5% level of significance table value of F = 2.90.

**Decision :** The computed value of F(1.018) is less than its table value 2.90. Hence, null hypothesis is correct and variances of both samples are equal.

**15. For a random sample of 10 pigs fed on diet A, the increases in weight in pounds in certain period were :**

**10, 6, 16, 17, 13, 12, 8, 14, 15, 9**

**For another random sample of 12 pigs fed on diet B, the increases in weight in pounds in the same periods were :**

**7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17**

**Test that both the samples come from population having same variance. ($F_{.05}$ for $v_1 = 11$, $v_2 = 9$ is 3.112)**

*Sol :*

**(Imp.)**

**Null Hypothesis :** Both the samples come from population having same vairnace.

**Calculation of Variances**

| Sample I | | | Sample II | | |
|---|---|---|---|---|---|
| $X_1$ | $X_1 - \bar{X}_1$ | $(X_1 - \bar{X}_1)^2$ | $X_2$ | $X_2 - \bar{X}_2$ | $(X_2 - \bar{X}_2)^2$ |
| 10 | –2 | 4 | 7 | –8 | 64 |
| 6 | –6 | 36 | 13 | –2 | 4 |
| 16 | 4 | 16 | 22 | 7 | 49 |
| 17 | 5 | 25 | 15 | 0 | 0 |
| 13 | 1 | 1 | 12 | –3 | 9 |
| 12 | 0 | 0 | 14 | –1 | 1 |
| 8 | –4 | 16 | 18 | 3 | 9 |
| 14 | 2 | 4 | 8 | –7 | 49 |
| 15 | 3 | 9 | 21 | 6 | 36 |
| 9 | –3 | 9 | 23 | 8 | 64 |
| | | | 10 | –5 | 25 |
| | | | 17 | 2 | 4 |
| $\Sigma X_1 = 120$ | | 120 | $\Sigma X_1 = 180$ | | 314 |

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{120}{10} = 12$$

$$\bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{180}{12} = 15$$

$$S_1^2 = \frac{(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{120}{10-1} = \frac{120}{9} = 13.33$$

$$S_2^2 = \frac{(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{314}{12-1} = \frac{314}{11} = 28.55$$

**F-Statistic**

$$F = \frac{S_2^2}{S_1^2} = \frac{28.55}{13.33} = 2.142$$

**Critical Value :** 3.112 (given in the question)

**Decision :** The calculated value of F(2.142) is less than its critical value (3.112). Hence, null hypothesis is accepted and both the samples seem to come from population having same variance.

## 5.9 Chi-Square Test

**Q26. Explain briefly about Chi-Square Test.** **(Imp.)**

*Ans :*

The magnitude of discrepancy between the theory and observation is given by the quantity $\chi^2$ (a Greek letter, pronounced as "chi-square"). If $\chi = 0$, the observed and expected frequencies completely coincide. As the value of $\chi^2$ increases, the discrepancy between the observed and theoretical frequencies increases. Thus, $\chi^2$ affords a measure of the correspondence between theory and observation.

**Definition**

If a set of events $A_1, A_2,.....,A_n$ are observed to occur with frequencies $O_1, O_2,.....,O_n$ respectively and according to probability rules $A_1, A_2,...., A_n$ are expected to occur with frequencies $E_1, E_2,....,E_n$ respectively with $O_1, O_2,.....,0_n$ are called observed frequencies and $E_1, E_2,....,E_n$ are called expected frequencies.

If $O_i$ (i = 1, 2,...,n) is a set of observed (experimental) frequencies and $E_i$ (i= 1, 2,...,n) is the corresponding set of expected (theoretical) frequencies, then $\chi^2$ is defined as $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)}{E_i}$ with (n – l) degrees of freedom.

$\chi^2$ is used to test whether differences between observed and expected frequencies are significant.

**Note :**

If the data is given in a series of 'n' numbers then degrees of freedom = n – 1.

In case of Binomial distribution, d.f. = n – 1.

In case of Poisson distribution, d.f. = n – 2

In case of Normal distribution, d.f. = n – 3

Chi-square Distribution is an important continuous probability distribution and it is used in both large and small tests. In chi-square tests, $\chi^2$ - distribution is mainly used

(i) To test the goodness of fit,

(ii) To test the independence of attributes,

(iii) To test if the population has a specified value of the variance $\sigma^2$.

### 5.9.1 Test for Single Variance

**Q27. Explain chi-square test for Single Variance.**

*Ans :*

A test of a single variance assumes that the underlying distribution is normal. The null and alternative hypotheses are stated in terms of the population variance (or population standard deviation). The test statistic is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

where:

n is the the total number of data

$s^2$ is the sample variance

$\sigma^2$ is the population variance

You may think of s as the random variable in this test. The number of degrees of freedom is df

n – 1. A test of a single variance may be right-tailed, left-tailed, or two-tailed. The next example will show you how to set up the null and alternative hypotheses. The null and alternative hypotheses contain statements about the population variance.

## PROBLEMS

16. **A statistician wishes to test the claim that the standard deviation of the weights of firemen is less than 25 pounds. She selected a random sample of 20 firemen and found s = 23.2 pounds. Assuming that the weights of firemen are normally distributed, test the claim of the statistician at the 0.05 level of significance.**

*Sol :*

Given that the sample size n = 20 and sample standard deviation s = 23.2

**Step 1 : Hypothesis Problem**

The hypothesis testing problem is

$H_0 : \sigma = 25$ against

$H_1 : \sigma < 25$ (left-tailed).

**Step 2 : Test Statistic**

The test statistic for testing above hypothesis testing problem is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

The test statistic follows a chi-square distribution with n – 1 degrees of freedom.

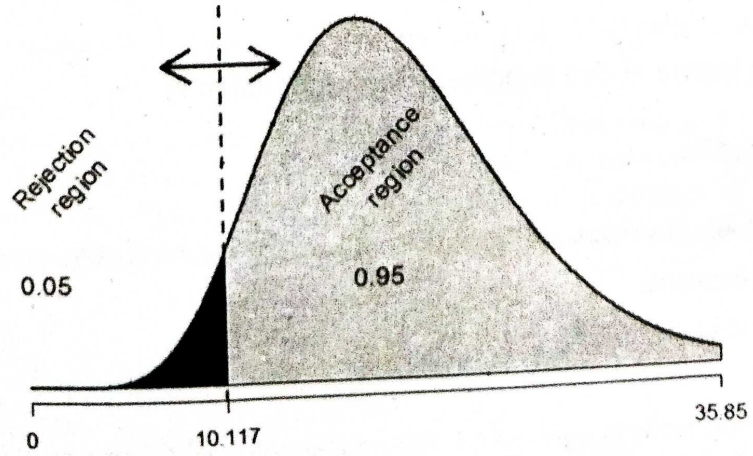
**Step 3 : Level of Significance**

The significance level is a = 0.05.

**Step 4 : Critical Value**

As the alternative hypothesis is left-tailed, the critical value of $\chi^2$ for $\alpha = 0.05$ level of significance and n – 1 = 19 degrees of freedom si 10.117 (from $\chi^2$ statistical table).

**Critical Region for Left-tailed test**




The rejection region (i.e., critical region) is $\chi^2 < 10.117$.

**Step 5 : Test Statistic**

The test statistic under the null hypothesis is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$$= \frac{(20-1)*(23.2)^2}{(25)^2}$$

$$= 16.362$$

**Step 6 : Decision (Traditional Approach)**

The test statistic is $\chi^2 = 16.362$ which falls outside the critical region, we fail to reject the null hypothesis.

(OR)

**Step 6 : Decision (p-value Approach)**

This is a left-tailed test, so the p-value is the area to the left of the test statistic ($\chi^2 = 16.362$) is p-value = 0.367.

The p-value is 0.367 which is greater than the significance level of a = 0.05, we fail to reject the null hypothesis.

---

**17. An engineer is investing the amount of standard deviation in the time it takes a 3D printer to make a particular part. The engineer believes that the standard deviation in the time it takes to make the part is more than 2. Test this at a = 0.01 level of significance, using 11 sample times taken while the printer was making these parts. The sample standard deviation is 2.3.**

***Sol :***

Given that the sample size n = 11 and sample standard deviation s = 2.3.

**Step 1 : Hypothesis Problem**

The hypothesis testing problem is

$H_0 : \sigma = 2$ against

$H_1 : \sigma > 2$ (right-tailed)

**Step 2 : Test Statistic**

The test statistic for testing above hypothesis testing problem is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

The test statistic follows chi-square distribution with n – 1 degrees of freedom.

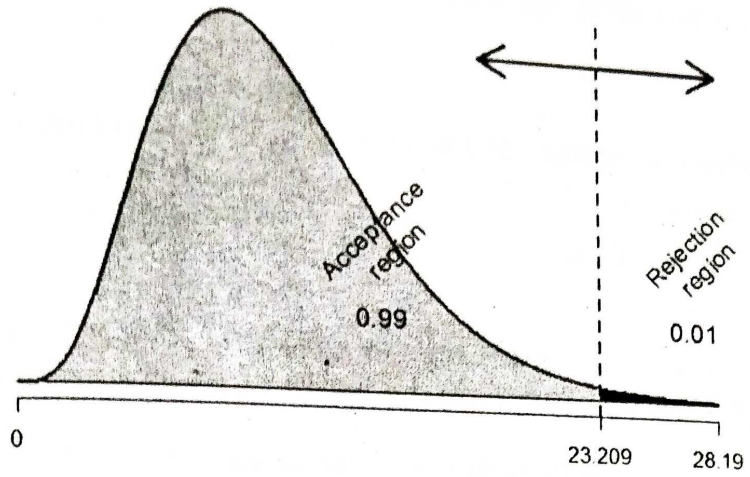
**Step 3 : Level of Significance**

The significance level is a = 0.01.

**Step 4 : Critical Value**

As the alternative hypothesis is right-tailed, the critical value of $\chi^2$ for a = 0.01 level of significance and n – 1 = 10 degrees of freedom is 23.209 (from $\chi^2$ statistical table)

**Critical Region for Right-tailed test**




The rejection region (i.e., critical region is $\chi^2 > 23.209$.

**Step 5 : Test Statistic**

The test statistic under the null hypothesis is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$$= \frac{(11-1)*(2.3)^2}{(2)^2}$$

$$= 13.225$$

**Step 6 : Decision (Traditional approach)**

The test statistic is $\chi^2 = 13.225$ which falls outside the critical region, we fail to reject the null hypothesis.

OR

**Step 6 : Decision (p-value Approach)**

This is a right-tailed test, so the p-value is the area to the left of the test statistic ($\chi^2 = 13.225$) is p-value = 0.2114.

The p-value is 0.2114 which is greater than the significance level of a = 0.01, we fail to reject the null hypothesis.

18. **A cigarette manufacturer wishes to test the claim that the variance of nicotine of its cigarettes is 0.644. Nicotine content is measured in milligrams and is assumed normally distributed. A sample of 20 cigarettes has a standard deviation of 1.00 milligram. At $\alpha$ = 0.01, is there enough evidence to reject the manufacturer's claim ?**

*Sol :*

Given that the sample size n = 20 and sample standard deviation s = 1.

**Step 1 Hypothesis Problem**

The hypothesis testing problem is $H_0 : \sigma^2 = 0.644$ against $H_1 : \sigma^2 \neq 0.644$ (two - tailed)

**Step 2 Test Statistic**

The test statistic for testing above hypothesis testing problem is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

The test statistic follows chi - square distribution with n – 1 degrees of freedom.

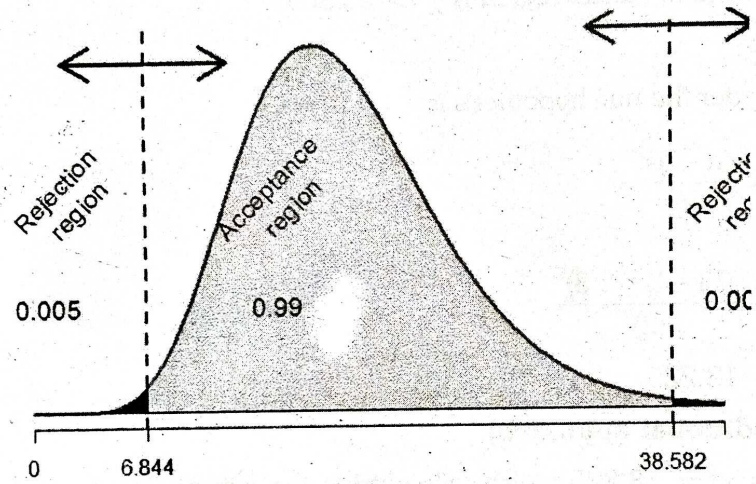
**Step 3 Level of significance**

The significance level is $\alpha = 0.01$

**Step 4 Critical Value**

As the alternative hypothesis is two-tailed, the critical values of $\chi^2$ for $\alpha = 0.01$ level of significance and $n - 1 = 19$ degrees of freedom are 6.844 and 38.582 (from $\chi^2$ statistical table).

Critical Region for Two-tailed test




The rejection region (i.e. critical region) is $\chi^2 < 6.844$ or $\chi^2 > 38.582$.

**Step 5 Test Statistic**

The test statistic under the null hypothesis is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$$= \frac{(20-1)*(1)^2}{(0.8024961)^2}$$

$$= 29.503$$

**Step 6 Decision (p - value Approach)**

This is a two - tailed test, so the p - value is the area to the left of the test statistic ($\chi^2 = 29.503$) is p-value = 0.0585.

The p-value is 0.0585 which is greater than the significance level of $\sigma = 0.02$, we fail to reject the null hypothesis.

### 5.9.2 Test of Independence of Attributes

**Q28. Explain the Chi-Square Test for independence of attributes.**

**(Imp.)**

*Ans :*

An attribute means a quality or characteristic. Example of attributes are drinking, smoking, blindness, honesty, beauty etc.

An attribute may be marked by its presence (position) or absence in a number of a given population. Let the observations be classified according to two attribute and the frequencies $O_i$ in the different categories be shown in a two-way table called contingency table. We have to test on the basis of cell frequencies whether the two attributes are independent or not. We take the Null - Hypothesis $H_0$ that there is no association between the attributes i.e., we assume that the two attributes are independent. The expected frequencies ($E_i$) of any cell $= \dfrac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$

The test statistic $\chi^2 = \sum_i \left[\dfrac{(O_i - E_i)^2}{E_i}\right]$ approximately follows Chi-square distribution with d.f. = (No. of rows – 1) × (No. of columns – 1)

If the calculated value of $\chi^2$ is less than the table value at a specified level (generally 5%) of significance, the hypothesis holds good i.e., the attributes are independent and do not bear any association. On the other hand, if the calculated value of $\chi^2$ is greater than the table value at a specified level of significance, we say that the results of the experiment do not support the hypothesis, in other words, the attributes are associated.

Let us consider two attributes A and B. A is divided into two classes and B is divided into two classes. The various cell frequencies can be expressed in the following table known as 2 × 2 contingency table.

| | A | a | b |
|---|---|---|---|
| | B | c | d |
| a | b | | a + b |
| c | d | | c + d |
| a + c | b + d | | N = a + b + c + d |

The expected frequencies are given by

$E(a) = \dfrac{(a+c)(a+b)}{N}$ $\quad$ $E(b) = \dfrac{(b+d)(a+b)}{N}$ $\quad$ a + b

$E(c) = \dfrac{(a+c)(c+d)}{N}$ $\quad$ $E(d) = \dfrac{(b+d)(c+d)}{N}$ $\quad$ c + d

a + c $\quad$ b + d $\quad$ N = a + b + c + d

The value of $y^2$ is given by $\chi^2$ is given $\chi^2 = \dfrac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ where

N = a + b + c + d with d.f. = (2 – 1)(2 – 1) = 1. We use this formula when the expected frequencies are in fractions (or decimals).

## PROBLEMS

**19. On basis of information given below about the treatment of 200 patients suffering from a disease, state whether the new treatment is comparatively superior to the conventional treatment.**

| | Favourable | Not favourable | Total |
|---|---|---|---|
| New | 60 | 30 | 90 |
| Conventional | 40 | 70 | 110 |

*Sol :*

**Null Hypothesis $H_0$ :** No difference between new and conventional treatment (or) New and conventional treatment are independent.

The number of degrees of freedom is $(2 - 1)(2 - 1) = 1$

Expected frequencies are given in the table :

$$\text{Expected frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

$\frac{90 \times 100}{200} = 45$ $\quad$ $\frac{90 \times 100}{200} = 45$ $\quad$ 90

$\frac{100 \times 110}{200} = 55$ $\quad$ $\frac{100 \times 110}{200} = 55$ $\quad$ 110

100 $\quad$ 100 $\quad$ 200

**Calculation of $\chi^2$ :**

| Observed Frequency ($O_i$) | Expected Frequency ($E_i$) | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|
| 60 | 45 | 225 | 5 |
| 30 | 45 | 225 | 5 |
| 40 | 55 | 225 | 4.09 |
| 70 | 55 | 225 | 4T)9 |
| 200 | 200 | | 18.18 |

$$\therefore \quad \chi^2 = \sum \frac{(O - E)^2}{E} = 18.18$$

Tabulated $\chi^2$ for 1 d.f. at 5% level of significance is 3.841.

Since calculated $\chi^2$ > tabulated $\chi^2$ we reject the null hypothesis $H_0$ i.e., new and conventional treatment are not independent. The new treatment is comparatively. superior to conventional treatment.

**20. The following table gives the classification of 100 workers according to sex and nature of work. Test whether the nature of work is independent of the sex of the worker.**

| | Stable | Unstable | Total |
|---|---|---|---|
| Males | 40 | 20 | 60 |
| Females | 10 | 30 | 40 |
| Total | 50 | 50 | 100 |

*Sol :*

**Null Hypothesis $H_0$ :** The nature of work is independent of the sex of the workers. Expected frequencies are given in the table :

| | | |
|---|---|---|
| $\frac{50\times60}{100} = 30$ | $\frac{50\times60}{100} = 30$ | 60 |
| $\frac{50\times40}{100} = 20$ | $\frac{50\times40}{100} = 20$ | 40 |
| 50 | 50 | 100 |

**Calculation of $\chi^2$ :**

| $O_i$ | $E_i$ | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|
| 40 | 30 | 100 | 3.333 |
| 20 | 30 | 100 | 3.333 |
| 10 | 20 | 100 | 5.000 |
| 30 | 20 | 100 | 5.000 |
| 100 | 100 | $\sum\frac{(O_i - E_i)^2}{E_i}$ | 16.66 |

$$\chi^2 = \sum\frac{(O_i - E_i)^2}{E_i} = 16.66$$

$\therefore$ Calculated $\chi^2 = 16.66$

Tabulated value of $\chi^2$ for $(2 - 1)(2 - 1) = 1$ d.f. at 5% level of significance is 3.84.

Since calculated $\chi^2$ > tabulated $\chi^2$, we reject the null hypothesis $H_0$, i.e,, the nature of work is not independent of the sex of the workers.

i.e., there is difference in the nature of work on the basis of sex..

**21. In an anti malarial campaign in a certain area quinine was administered to 812 persons. Out of a total population of 3248 persons the number of fever cases in shown below :**

| | Fever | No Fever | Total |
|---|---|---|---|
| Quinine | 20 | 792 | 812 |
| No Quinine | 220 | 2216 | 2436 |
| Total | 240 | 3008 | 3248 |

**Discuss the usefulness of quinine in checking malaria.**

**Sol :** **(Imp)s**

It is a $\chi^2$ test

**Null Hypothesis**

$H_0$ = Quinine is not effective in checking malaria

$H_a$ = Quinine is effective in checking malaria

**Computing Test Statistic**

$$\chi^2 = \Sigma\left(\frac{(O_{ij} - E_{ij})^2}{E_q}\right)$$

$$E_{ij} = \frac{RT \times CT}{GT}$$

**Calculation of Expected Frequencies**

| Treatment | Fever | No Fever | Total |
|---|---|---|---|
| Quinine | $\frac{812 \times 240}{3.248} = 60$ $E_{11}$ | $\frac{812 \times 3,008}{3,248} = 752$ $E_{12}$ | 812 |
| No Quinine | $\frac{2,436 \times 240}{3,248} = 180$ $E_{21}$ | $\frac{2,436 \times 3,008}{3,248} = 2,256$ $E_{22}$ | 2,436 |
| **Total** | **240** | | **3,248** |

**Calculation of $x^2$**

| Group | $Q_{ij}$ | $E_{ij}$ | $Q_{ij} - E_{ij}$ | $(Q_{ij} - E_{ij})^2$ | $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ |
|---|---|---|---|---|---|
| | (1) | (2) | (3) = (1) – (2) | (4) | (5) = (4)/(2) |
| 11 | 20 | 60 | – 40 | 1,600 | 26.667 |
| 12 | 792 | 752 | 40 | 1,600 | 2.128 |
| 21 | 220 | 180 | 40 | 1,600 | 8.889 |
| 22 | 2,216 | 2,256 | – 40 | 1,600 | 0.709 |
| | | | | Total | 38.393 |

$$\therefore \chi^2 = \sum\left(\frac{O_{ij} - E_{ij}}{E_{ij}}\right) = 38.393$$

Level of significance, $\alpha = 0.05$ (Assumed)

Degree of freedom $= (c - 1)(r - 1) = (2 - 1)(2 - 1)$

$= 1 \times 1 = 1$

Table of $\chi^2$ at 1 d,f and 0.05 is 3.84

Since calculated $\chi^2 \geq \chi^2_{tab}$, we reject null hypothesis. Hence, quinine is effective in checking malaria.

# Short Question and Answers

**1. Correlation.**

*Ans :*

**Meaning**

Correlation is the study of the linear relationship between two variables. When there is a relationship of 'quantitative measure between two set of variables, the appropriate statistical tool for measuring the relationship and expressing each in a precise way is known as correlation.

**For example,** there is a relationship between the heights and weights of persons, demand and prices of commodities etc.

Correlation analysis is the statistical tool we can use to describe the degree to which one variable is linearly related to another.

**Definitions**

(i) **According to L.R. Connor** "If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated."

(ii) **According to A.M. Tuttle** "Correlation is an analysis of covariation between two or more variables".

(iii) **According to Croxton and Cowden** "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

**2. Positive correlation**

*Ans :*

If the values of two variables deviate in the same direction i.e., if increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, the corresponding correlation is said to be positive or direct.

**3. What is Scatter Diagram?**

*Ans :*

Scatter diagram method is the simplest way of diagrammatic representation of a bivariate distribution and helps in ascertaining the correlation between the two variables under study i.e., it portraits the relationship between these two variables graphically.

**4. Karl Pearson's Coefficient of Correlation.**

*Ans :*

Karl Parson's Coefficient of Correlation is arrived at with the help of a statistical formula that takes into account the mean and standard deviation of the two variables, the number of such observations and the covariance between them. Since Karl Pearson's coefficient of correlation is a number, it can describe the strength of the correlation in greater detail and more objectively. A value of –1 signifies "absolute" negative correlation, a value between –1 and –0.5 signifies strong negative correlation, a value between –0.5 and –0.25 signifies moderate negative correlation and a value between –0.25 and 0 signifies weak negative correlation. Similarly, a value of +1 signifies "absolute" positive correlation, a value between +1 and +0.5 signifies strong positive correlation, a value between +0.5 and +0.25 signifies moderate positive correlation and a value between +0.25 and 0 signifies weak positive correlation.

**5. Define Regression?**

*Ans :*

**Meaning**

Regression analysis which confines itself to a study of only two variables is called simple regression. The regression analysis which studies more than two variables at a time is called multiple regression. In the simple regression analysis there are two variables-one of which is known as 'independent variable' or 'regressor' or 'predictor'. On the basis of the values of this variable the values

of the other variable are predicted. The other variable whose values are predicted is called the 'dependent' or 'regressed' variable.

**Definitions**

i) "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data."

ii) **According to Morris Hamburg** The term 'regression analysis' refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process."

iii) **According to Taro Yamane** "One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related causally, is regression analysis."

## 6. Simple Linear Regression.

*Ans :*

Linear regression is a form of regression which is used for modeling the relationship between scalar variables like X and F under linear regression, linear functions are used to model the data and the unknown parameters, of models are estimated from the data. Hence, these models are known as linear models.

Linear models more commonly refers to those models, where the conditional mean of variable 'F for a given value of variable X will be an affine function of X. A linear regression may also refer to a model, where median or other quantile of the conditional distribution of 'F for a given value of 'X is termed as linear function of X. Similar, to all types of regression analysis, linear regression also aims on the conditional probability distribution of 'F for a given 'X, instead of joint probability distribution of 'F and X.

## 7. Define Hypothesis.

*Ans :*

**Introduction**

The term 'hypothesis' is derived from the ancient Greek word, 'hypothesis' that means 'to put under' (or) 'to suppose'. Hypothesis is also a combination of two words 'Hypo, Thesis where 'Hypo' means tentative or subject to verification and 'Thesis' a statement based on concepts, theories and past experiences about the solution of the problem. The term hypothesis literally means an assumption or a supposition about the state of affairs of a certain thing or phenomena or facts or variable or situation. Thus, "hypothesis is perceived as a proposition or set of propositions set forth as an explanation for occurrence of some specified group of phenomenon either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts. Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variables"

**Definitions**

The view point of various thinkers has been presented as under :

1. **According to Good, Barr and Scates** "Hypothesis is a statement temporarily accepted as true in the light of what is at the time, known about a phenomenon, and it is employed as a basis for action in the search for new truth. When hypothesis is fully established, it may take the form of facts, principles and theories".

2. **According to Lundberg** "Hypotheses is a tentative generalization, the validity of which is remains to be tested. In the most elementary stage the hypothesis may be any hunch, guess, imaginative idea which become base for further investigation".

3. **According to Best** "Hypothesis id a shrewd guess or inference that is formulated and provisionally adopted to explain observed facts or conditions and to guide in further investigation".

## 8. What is Small sample test ?

*Ans :*

Small sample size referred to size of sample which is less than 30. In case of small sample size the z-test is not appropriate test statistic as the assumptions on which it is based do not hold good

in case of small sample. The theoretical work on t-distribution was done by W.S. Gosset (1876-1937) under the pen name "student' as he was the employee of the company Guinness & Sons, a Dublin bravery, Ireland, which did not allowed it employees to publish research findings under their own names. The t-distribution is used when sample size is less than 30 and the population standard deviation is not known.

**9. What are the properties of t-distribution?**

*Ans :*

1. The shape of t-distribution is bell-shaped, which is similar to that of a normal distribution and is symmetrical about the mean.
2. The t-distribution curve is also asymptotic to the t-axis, i.e., the two tails of the curve on both sides of t = 0 extends to infinity.

**10. Paired t-test.**

*Ans :*

Paired observations arise in many practical situations where each homogeneous experimental unit receives both population conditions. As a result, each experimental unit has a pair of observations, one for each population.

For instance, to test the effectiveness of "drug" some 11 persons blood pressure is measured "before" and "after" the intake of certain drug. Here the individual person is the experimental unit and the two populations are blood pressure "before" and "after" the drug is given. Thus for each observation in one sample, there is a corresponding observation in the other sample pertaining to the same character. Hence the two samples are not independent.

Consider another example. Suppose a business concern is interested to know whether a particular media of promoting sales of a product is really effective or not. In this case we have to test whether the average sales before and after the sales promotion are equal.

**11. Define F-test.**

*Ans :*

F test is a statistical test that is used in hypothesis testing to check whether the variances of two populations or two samples are equal or not. In an f test, the data follows an f distribution. This test uses the f statistic to compare two variances by dividing them. An f test can either be one-tailed or two-tailed depending upon the parameters of the problem.

The f value obtained after conducting an f test is used to perform the one-way ANOVA (analysis of variance) test. In this article, we will learn more about an f test, the f statistic, its critical value, formula and how to conduct an f test for hypothesis testing.

# Exercises Problems

1. Calculate Correlation for the data given below:

| X | 42 | 36 | 48 | 43 | 55 | 52 | 38 |
|---|---|---|---|---|---|---|---|
| Y | 132 | 120 | 140 | 143 | 142 | 148 | 122 |

**(Ans: r = 0.856)**

2. Calculate Co-efficient of Correlation and interpret the value:

| Marks in QT | 57 | 42 | 40 | 38 | 42 | 45 | 42 | 44 | 40 | 46 | 44 | 43 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Marks in IOM | 10 | 26 | 30 | 41 | 29 | 27 | 27 | 19 | 18 | 19 | 31 | 29 |

**(Ans: r = – 0.7285).**

3. Calculate the coefficient of correlation and probable error from the following - data:

| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 20 | 16 | 14 | 10 | 10 | 9 | 8 | 7 | 6 | 5 |

**(Ans: r = – 0.95; P.E. = 0.0208)**

4. Fit a straight line regression equation of Yon X from the following data.

| X | 10 | 12 | 13 | 16 | 17 | 20 | 25 | 29 |
|---|---|---|---|---|---|---|---|---|
| Y | 10 | 12 | 24 | 27 | 29 | 33 | 37 | 42 |

**(Ans: Y=1.6 X – 1.65)**

5. Find the two regression equations from the following data:

| X | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Y | 2 | 3 | 5 | 4 | 6 |

**(Ans: X = 0.9 Y – 0.6; Y = 0.9 X +1.3)**

# Choose the Correct Anaswers

1. Which of the following are types of correlation? [d]
   (a) Positive and Negative
   (b) Simple, Partial and Multiple
   (c) Linear and Nonlinear
   (d) All of the above

2. Which of the following is true for the coefficient of correlation? [c]
   (a) The coefficient of correlation is not dependent on the change of scale
   (b) The coefficient of correlation is not dependent on the change of origin
   (c) The coefficient of correlation is not dependent on both the change of scale and change of origin
   (d) None of the above

3. Which of the following statements is true for correlation analysis? [c]
   (a) It is a bivariate analysis
   (b) It is a multivariate analysis
   (c) It is a univariate analysis
   (d) Both a and c

4. If the values of two variables move in the same direction, [d]
   (a) The correlation is said to be non-linear
   (b) The correlation is said to be linear
   (c) The correlation is said to be negative
   (d) The correlation is said to be positive

5. If the values of two variables move in the opposite direction, [d]
   (a) The correlation is said to be linear
   (b) The correlation is said to be non-linear
   (c) The correlation is said to be positive
   (d) The correlation is said to be negative

6. Which of the following techniques is an analysis of the relationship between two variables to help provide the prediction mechanism? [c]
   (a) Standard error
   (b) Correlation
   (c) Regression
   (d) None of the above

7. Which of the following statements is true about the arithmetic mean of two regression coefficients? [d]
   (a) It is less than the correlation coefficient
   (b) It is equal to the correlation coefficient
   (c) It is greater than or equal to the correlation coefficient
   (d) It is greater than the correlation coefficient

8. What is the meaning of the testing of the hypothesis? [b]
   (a) It is a significant estimation of the problem
   (b) It is a rule for acceptance or rejection of the hypothesis of the research problem
   (c) It is a method of making a significant statement
   (d) None of the above

9. Which of the following statements is true about the null hypothesis? [a]

(a) Any wrong decision related to the null hypothesis results in two types of errors

(b) Any wrong decision related to the null hypothesis results in one type of an error

(c) Any wrong decision related to the null hypothesis results in four types of errors

(d) Any wrong decision related to the null hypothesis results in three types of errors

10. Which of the following statements is true about the type two error? [a]

(a) Type two error means to accept an incorrect hypothesis

(b) Type two error means to reject an incorrect hypothesis

(c) Type two error means to accept a correct hypothesis

(d) Type two error means to reject a correct hypothesis