(7 MARKS) CHAPTER WISE

LONG-TYPE QUESTIONS WITH ANSWERS

CORRELATION & REGRESSION

1. Define correlation coefficient and derive its limits.

Ans. Correlation coefficient may be defined as the degree of linear relationship between two variables. It is denoted by r_{xy} and is computed by the formula:

$$r_{xy} = \frac{Cov.(x,y)}{\sigma_x \sigma_y}$$

$$Cov.(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

In order to derive the limits of correlation coefficient, let us consider an expression

$$S = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_i - \bar{x}}{\sigma_x}\right) + \left(\frac{y_i - \bar{y}}{\sigma_y}\right)\right]^2$$

S represents the sum of squares of real numbers; so it is non-negative.

$$\text{Thus } S \geq 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{\sigma_y}\right)^2 + 2 \times \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right) \geq 0$$

$$\Rightarrow \sigma_x^2 \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sigma_y^2 \frac{1}{n} \sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{2}{\sigma_x \sigma_y} \times \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \geq 0$$

$$\Rightarrow \frac{\sigma_x^2}{\sigma_x^2} + \frac{\sigma_y^2}{\sigma_y^2} + \frac{2Cov(x,y)}{\sigma_x \sigma_y} \geq 0$$

$$\Rightarrow 1 + 1 + 2r_{xy} \geq 0 \Rightarrow 2r_{xy} \geq -2 \Rightarrow r_{xy} \geq -1$$

$$\Rightarrow -1 \leq r_{xy} \cdots (i)$$

Similarly let $S' = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_i - \bar{x}}{\sigma_x}\right) - \left(\frac{y_i - \bar{y}}{\sigma_y}\right)\right]^2$

Then clearly

$$S' \geq 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{\sigma_y}\right)^2 - 2 \times \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right) \geq 0$$

$$\Rightarrow \sigma_x^2 \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sigma_y^2 \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{2}{\sigma_x \sigma_y} \times \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \geq 0$$

$$\Rightarrow \frac{\sigma_x^2}{\sigma_x^2} + \frac{\sigma_y^2}{\sigma_y^2} - \frac{2Cov(x,y)}{\sigma_x \sigma_y} \geq 0$$

$$\Rightarrow 1 + 1 - 2r_{xy} \geq 0 \Rightarrow -2r_{xy} \geq -2 \Rightarrow r_{xy} \leq 1 \cdots (ii)$$

Combining the results (*i*) and (*ii*) it is concluded that $-1 \leq r_{xy} \leq 1$.

2. Discuss the effect of change of origin and scale on correlation coefficient.

Ans. Let x and y be two variables having correlation coefficient r_{xy}.

On applying change of origin and scale let the new variables be u and v; where $u = \frac{x - A}{h}$ and $v = \frac{y - B}{k}$

A, B, h & k are constants and $h, k \neq 0$.

Then $x = A + hu$ and $y = B + kv$

$\bar{x} = A + h\bar{u}$ and $\bar{y} = B + k\bar{v}$, $\sigma_x = h\sigma_u$ and $\sigma_y = k\sigma_v$

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(A + hu_i - A - h\bar{u}\right)\left(B + kv_i - B - K\bar{v}\right)$$

$$= hk\frac{1}{n}\sum_{i=1}^{n}\left(u_i - \bar{u}\right)\left(v_i - \bar{v}\right) = hkCov(u,v)$$

$$r_{xy} = \frac{\text{cov}(x,y)}{\sigma_x\sigma_y} = \frac{hk\,\text{cov}(u,v)}{h\sigma_u k\sigma_v} = \frac{\text{cov}(u,v)}{\sigma_u\sigma_v} = r_{uv}$$

Hence it is proved that correlation coefficient is independent of change of origin and scale.

3. Define a regression coefficient and describe the properties of regression coefficients.

Ans. A regression coefficient may be defined as the increment in the value of the dependent variable corresponding to unit increment in the value of the independent variable. Hence there are two regression coefficients named as b_{yx} = regression coefficient of y on x and b_{xy} = regression coefficient of x on y.

Properties of regression coefficients:

(i) $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

(ii) Both the regression coefficients and the correlation coefficient always have the same sign.

(iii) The geometric mean of the regression coefficients is the correlation coefficient.

(iv) The product of the regression coefficients always lies between 0 and 1.

(v) If the regression coefficients are positive then

$$\frac{b_{yx} + b_{xy}}{2} > r$$

(vi) If one regression coefficient is zero then the other is zero and the correlation coefficient is also zero.

4. What is the need for rank correlation method? Derive the formula for Spearman's Rank correlation coefficient for n pairs of observations without any repetition of ranks.

Ans. The method of rank correlation is useful for determination of correlation between attributes which are characteristics not capable of being measured directly. For example qualification and work efficiency are expected to be related but neither qualification nor work efficiency is measurable. So in such cases the correlation between these two attributes can be studied through the method rank correlation.

Spearman's rank correlation coefficient may be defined as the simple correlation coefficient between two sets of ranks assigned to the observations on the basis of the attributes under consideration. Assuming that no two observations have the same level of any attribute, the ranks of the n given observations will be 1, 2, 3, …, n for each attribute. However these ranks of the observations may not be in the same order.

Let x_i = rank of the i^{th} observation according to attribute x and y_i = rank of the i^{th} observation according to attribute y. Then Spearman's rank correlation coefficient will be the simple correlation coefficient between the pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

It is clear that

$$\sum_{i=1}^{n} x_i = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$$

$$= \sum_{i=1}^{n} y_i \Rightarrow \bar{x} = \frac{1}{n} \times \frac{n(n+1)}{2} = \frac{(n+1)}{2} = \bar{y}$$

$$\sum_{i=1}^{n} x_i^2 = 1^2 + 2^2 + 3^2 + \cdots + n^2$$

$$= \frac{n(n+1)(2n+1)}{6} = \sum_{i=1}^{n} y_i^2$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\bar{x}\right)^2 = \frac{1}{n} \times \frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2$$

$$= \left(\frac{n+1}{2}\right)\left(\frac{4n+2-3n-3}{6}\right)$$

$$= \left(\frac{n+1}{2}\right)\left(\frac{n-1}{6}\right) = \left(\frac{n^2-1}{12}\right) = \sigma_y^2$$

Let $d_i = x_i - y_i$

= rank difference for the i^{th} observation

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\left(x_i - y_i\right)^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} x_i y_i$$

$$= \frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)(2n+1)}{6} - 2\sum_{i=1}^{n} x_i y_i$$

$$\Rightarrow 2\sum_{i=1}^{n} x_i y_i = 2 \times \frac{n(n+1)(2n+1)}{6} - \sum_{i=1}^{n} d_i^2$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum_{i=1}^{n} d_i^2$$

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

$$= \frac{1}{n} \times \frac{n(n+1)(2n+1)}{6} - \frac{1}{2n}\sum_{i=1}^{n} d_i^2 - \left(\frac{n+1}{2}\right)^2$$

$$= \left(\frac{n+1}{2}\right)\left(\frac{4n+2-3n-3}{6}\right)$$

$$-\frac{1}{2n}\sum_{i=1}^{n} d_i^2 = \left(\frac{n^2-1}{12}\right) - \frac{1}{2n}\sum_{i=1}^{n} d_i^2$$

Hence Spearman's rank correlation coefficient

$$= R = \frac{\text{cov}(x,y)}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{\left(\frac{n^2-1}{12}\right) - \frac{1}{2n}\sum_{i=1}^{n} d_i^2}{\sqrt{\left(\frac{n^2-1}{12}\right)\left(\frac{n^2-1}{12}\right)}}$$

$$= \frac{\left(\frac{n^2-1}{12}\right)}{\left(\frac{n^2-1}{12}\right)} - \frac{\frac{1}{2n}\sum_{i=1}^{n} d_i^2}{\left(\frac{n^2-1}{12}\right)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

5. On the basis of a numerical example, show that r = 0 does not necessarily mean that the variables are independent.

Ans. Let us consider the following numerical example where $y = x^2$

x	–3	–2	–1	0	1	2	3
y	9	4	1	0	1	4	9

For this data: $n = 7, \sum x_i = 0, \sum y_i = 28,$

$$\sum x_i^2 = 28, \sum y_i^2 = 196, \sum x_i y_i = 0$$

Hence $r_{xy} = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}$

$$= \frac{0-0}{\sqrt{(7\times 28 - 0)(7\times 196 - 784)}} = 0$$

It is known that $y = x^2$ but it is calculated that $r_{xy} = 0$. This indicates that $r = 0$ does not necessarily mean that the variables are independent. It only indicates that the variables have no linear relationship.

6. Explain the method of least squares for fitting a regression line.

Ans: According to the method of least squares, a line of regression is a line passing through the scatter diagram in such a way that the sum of squares of deviations parallel to the axis representing the dependent variable is the minimum. Hence the line of regression of y on x is the line passing through the scatter diagram in such a way that the sum of squares of deviations parallel to the y-axis is the minimum.

Fitting of line of regression of y on x for n pairs of values (x_i, y_i) $(i = 1, 2, 3, \ldots, n)$

Let the equation of the line of regression of y on x be $Y = a + bx$. Then the point on the line of regression of y on x corresponding to $x = x_i$ is Y_i. Where $Y_i = a + b x_i$

Thus the deviation parallel to the y-axis is of the form $(y_i - Y_i) = (y_i - a - b x_i)$

The sum of squares of these deviations =

$$S = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

The objective is to find the values of a and b such that S becomes the minimum. Hence applying the principle of minima,

$$\frac{\partial S}{\partial a} = 0 \Rightarrow -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad \ldots \text{(i)}$$

$$\frac{\partial S}{\partial b} = 0 \Rightarrow -2\sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \quad \ldots \text{(ii)}$$

On solving these equations, unique values of a and b are obtained. It is observed that these values of a and b yield negative values for the second order partial derivatives of S with respect to a and b. Thus these values of a and b provide the minimum value for S. So by substituting these values of a and b in the equation Y = a + bx, the required equation of the line of regression of y on x can be obtained. The value of b obtained by solving the normal equations (i) and (ii) is used for the line of regression of y on x; so it is denoted by b_{yx}. As obtained from the normal equations,

$$b_{yx} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{Cov(x,y)}{\sigma_x^2}$$

This is the slope of the line of regression of y on x.

By dividing through out by n it is further obtained from equation (i) that $\bar{y} = a + b\bar{x}$ which implies that the line of regression of y on x passes through $(\bar{x}, \bar{y})$.

So by applying the point slope form of equation of straight line, the equation of the line of regression of y on x is obtained as $y - \bar{y} = b_{yx}(x - \bar{x})$

Fitting of line of regression of x on y for n pairs of values (x_i, y_i) $(i = 1, 2, 3, \ldots, n)$

In a similar manner the equation of the line of regression of x on y can be determined by considering the deviations parallel to the x-axis. Thus equation of the line of regression of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$

Here $b_{xy} = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum y_i^2 - \left(\sum y_i\right)^2} = \dfrac{Cov(x,y)}{\sigma_y^2}$.

7. What is a scatter diagram? How can it be used to determine the correlation between two variables?

Ans. A scatter diagram may be defined as the graphical representation of a bivariate data. It is a set of 'n' no. of points on the coordinate plane such that the coordinates of the i^{th} point are (x_i, y_i) where x_i = value of the variable x for the i^{th} observation and y_i = value of the variable y for the i^{th} observation. ($i = 1, 2, 3, \ldots, n$)

Correlation between two variables can be studied from the nature and position of the scatter diagram.

(i) If the points of the scatter diagram shows a general tendency of being spread from the left hand bottom corner towards the right hand top corner of the plane, the variables are positively correlated i.e. the values of the two variables have a general tendency to change in the same direction.

(ii) If the points of the scatter diagram shows a general tendency of being spread from the left hand top corner towards the right hand bottom corner of the plane, the variables are negatively correlated i.e. the values of the two variables have a general tendency to change in opposite directions.

(iii) If the points of the scatter diagram are such that they neither show positive nor show negative correlation, then the variables are said to have Zero correlation. This means that the variables do not have any linear relationship. However Zero correlation can not be conclusively determined only by inspecting a scatter diagram which is one of the major drawbacks of this method for the study of correlation.

(iv) If the points of the scatter diagram lie on a straight line making an angle of 45° with the positive X-axis, the variables are said to be perfectly positively correlated. In such cases the increase in the value of one variable result at an equal amount of increase in the value of the other variable.

(v) If the points of the scatter diagram lie on a straight line making an angle of 135° with the positive X-axis, the variables are said to be perfectly negatively correlated. In such cases the increase in the value of one variable result at an equal amount of decrease in the value of the other variable.

8. Discuss the effect of change of origin and scale on the regression coefficient.

Ans. Let x and y be two variables with regression coefficient b_{yx}. Then $b_{yx} = \dfrac{Cov(x,y)}{\sigma_x^2}$

Applying change of origin let $p = x - A$ and $q = y - B$ where A and B are constants.

Then $x = A + p$ and $y = B + q$

$$\text{So } \bar{x} = \frac{1}{n}\sum x_i = \frac{1}{n}\sum(A + p_i)$$

$$= \frac{1}{n}\sum A + \frac{1}{n}\sum p_i = \frac{1}{n} \times nA + \bar{p}$$

$$= A + \bar{p} \text{ and } \bar{y} = B + \bar{q}$$

$$\sigma_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum(A + p_i - A - \bar{p})^2$$

$$= \frac{1}{n}\sum(p_i - \bar{p})^2 = \sigma_p^2 \text{ and } \sigma_y^2 = \sigma_q^2$$

$$Cov(x,y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n}\sum(A + p_i - A - \bar{p})(B + q_i - B - \bar{q})$$

$$= \frac{1}{n}\sum(p_i - \bar{p})(q_i - \bar{q}) = Cov(p,q)$$

$$\text{Hence } b_{yx} = \frac{Cov(x,y)}{\sigma_x^2} = \frac{Cov(p,q)}{\sigma_p^2} = b_{qp}$$

$$\text{and similarly } b_{xy} = \frac{Cov(x,y)}{\sigma_y^2} = \frac{Cov(p,q)}{\sigma_q^2} = b_{pq}$$

Thus it is proved that regression coefficients are independent of change of origin.

Applying change of origin and scale, let

$u = \frac{x-A}{h}$ and $v = \frac{y-B}{k}$ Where A, B, h and k are constants and h, k $\neq 0$

Then $x = A + hu$ and $y = B + kv$

So $\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{n}\sum(A + hu_i)$

$= \frac{1}{n}\sum A + \frac{1}{n}\sum hu_i = \frac{1}{n} \times nA + h\bar{u}$

$= A + h\bar{u}$ and $\bar{y} = B + k\bar{v}$

$\sigma_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum(A + hu_i - A - h\bar{u})^2$

$= \frac{1}{n}\sum h^2(u_i - \bar{u})^2 = h^2\sigma_u^2$ and $\sigma_y^2 = k^2\sigma_v^2$

$Cov(x,y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$

$= \frac{1}{n}\sum(A + hu_i - A - h\bar{u})(B + kv_i - B - k\bar{v})$

$= \frac{1}{n}\sum hk(u_i - \bar{u})(v_i - \bar{v}) = khCov(u,v)$

Hence $b_{yx} = \frac{Cov(x,y)}{\sigma_x^2} = \frac{hkCov(u,v)}{h^2\sigma_u^2} = \frac{k}{h}b_{vu}$

and similarly $b_{xy} = \frac{Cov(x,y)}{\sigma_y^2} = \frac{hkCov(u,v)}{k^2\sigma_v^2} = \frac{h}{k}b_{uv}$

Thus it is obtained that regression coefficients are not independent of change of scale. Hence it can be concluded that regression coefficients are independent of change of origin but not independent of change of scale.

9. Explain the properties of regression lines.

Ans. Properties of regression lines:

(i) The two regression lines intersect at the point $(\bar{x}, \bar{y})$.

(ii) The slope of the line of regression of y on x is b_{yx} and the slope of the line of regression of x on y is $\frac{1}{b_{xy}}$

(iii) The regression lines can never be two distinct parallel lines. However if they have the same slope then they will coincide. In such cases the variables have perfect correlation i.e. $r = \pm 1$.

(iv) The lines of regression will be perpendicular to each other only if $r = 0$ i.e. the variables have no linear relationship. In such cases the line of regression of y on x will be parallel to the x-axis and the line of regression of x on y will be parallel to the y-axis.

(v) The angle between the two regression lines can be determined by using the condition

$\tan\alpha = \frac{b_{yx}b_{xy} - 1}{b_{yx} + b_{xy}}$ (where α is the angle between the two lines of regression)

10. Equations of two regression lines are $2x + y = 220$ and $9x + 50y = 3720$. Determine the means of x and y. Also find the correlation coefficient and the variance of x given that the standard deviation of y is 12. Then estimate the value of y for $x = 50$.

Ans. Given $2x + y = 220$ ... (i)

$9x + 50y = 3720$ ... (ii)

Slope of equation (i) $= -\frac{1}{2}$ and

slope of equation (ii) $= -\frac{9}{50}$

It is known that the regression lines intersect at the point $(\bar{x}, \bar{y})$. Hence by solving the equations (i) and (ii) it is obtained that

$$\bar{x} = \frac{\begin{vmatrix} 220 & 1 \\ 3720 & 50 \end{vmatrix}}{\begin{vmatrix} 2 & 1 \\ 9 & 50 \end{vmatrix}} = \frac{220 \times 50 - 1 \times 3720}{2 \times 50 - 1 \times 9} = \frac{7280}{91} = 80$$

$$\bar{y} = \frac{\begin{vmatrix} 2 & 220 \\ 9 & 3720 \end{vmatrix}}{\begin{vmatrix} 2 & 1 \\ 9 & 50 \end{vmatrix}} = \frac{2 \times 3720 - 9 \times 220}{2 \times 50 - 1 \times 9} = \frac{5460}{91} = 60$$

Let equation (i) be the line of regression of y on x then its slope = $b_{yx} = -\frac{1}{2}$ and equation (ii) be the line of regression of x on y then its slope =

$\frac{1}{b_{xy}} = -\frac{9}{50} \Rightarrow b_{xy} = -\frac{50}{9}$

Hence $b_{yx} \times b_{xy} = \left(-\frac{1}{2}\right)\left(-\frac{50}{9}\right) > 1$ which is impossible because the product of the regression coefficients always lies between 0 and 1. Hence the assumption is incorrect. So the correct assumption is equation (ii) is the line of regression of y on x having slope $= b_{yx} = -\frac{9}{50}$ and equation (i) is the line of regression of x on y with $b_{xy} = -2$

Thus $b_{yx} \times b_{xy} = \left(-\frac{9}{50}\right)(-2)$

$$\Rightarrow r = \sqrt{b_{yx} \times b_{xy}} = -\frac{3}{5} = -0.6$$

Given $\sigma_y = 12$. It is known that

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \Rightarrow -\frac{9}{50} = -\frac{3}{5} \times \frac{12}{\sigma_x}$$

$$\Rightarrow \sigma_x = \frac{3 \times 12 \times 50}{9 \times 5} = 40 \Rightarrow \sigma_x^2 = 1600$$

The value of y for a given value of x can be estimated by using the equation of line of regression of y on x. So the estimated value of y for $x = 50$ will be obtained by putting $x = 50$ in equation (ii). Thus for $x = 50$, Estimated value of $y = 65.4$

Thus for the given problem:

Mean of $x = 80$, mean of $y = 60$, correlation coefficient $= r = -0.6$, $\sigma_x^2 = 1600$, and the estimated value of y for $x = 50$ is 65.4.

11. Fit the lines of regression for the following data and determine the following estimated values:

(a) Value of X for Y= 20

(b) Value of Y for X= 40

X =	10	25	30	55	60
Y =	75	60	65	40	30

Ans Equation of the line of regression of y on x is

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

Equation of the line of regression of x on y is

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

$$b_{yx} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \text{ and}$$

$$b_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum y_i^2 - \left(\sum y_i\right)^2}$$

$$\bar{x} = \frac{\sum x_i}{n} \text{ and } \bar{y} = \frac{\sum y_i}{n}$$

x	y	x^2	y^2	xy
10	75	100	5625	750
25	60	625	3600	1500
30	65	900	4225	1950
55	40	3025	1600	2200
60	30	3600	900	1800
$\Sigma x_i = 180$	$\Sigma y_i = 270$	$\Sigma x_i^2 = 8250$	$\Sigma y_i^2 = 15950$	$\Sigma x_i y_i = 8200$

$$b_{yx} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{5 \times 8200 - 180 \times 270}{5 \times 8250 - (180)^2}$$

$$= \frac{41000 - 48600}{41250 - 32400} = -\frac{7600}{8850} = -0.86$$

$$b_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum y_i^2 - \left(\sum y_i\right)^2} = \frac{5 \times 8200 - 180 \times 270}{5 \times 15950 - (270)^2}$$

$$= \frac{41000 - 48600}{79750 - 72900} = -\frac{7600}{6850} = -1.11$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{180}{5} 36 \text{ and } \bar{y} = \frac{\sum y_i}{n} = \frac{270}{5} = 54$$

Hence Equation of the line of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x}) \Rightarrow y - 54 = -0.86(x - 36)$$

$$\Rightarrow y = 54 + 30.96 - 0.86x \Rightarrow y = 84.96 - 0.86x$$

Equation of the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 36 = -1.11(y - 54)$$

$$\Rightarrow x = 36 + 59.94 - 1.11y \Rightarrow x = 95.94 - 1.11y$$

TIME SERIES ANALYSIS

12. What is meant by analysis of time series? Explain the different mathematical models used for analysis of time series.

Ans. It is known that the value of a time series at a point of time is the combined effect due to four components. Analysis of time series means

determination of the independent effect due to each component of the time series. Hence for analyzing a time series it becomes necessary to assume a mathematical model for it. A mathematical model for a time series is assuming the value of the time series at a point of time as a function of the values of the components of the time series at that point of time. The popularly used models for analysis of time series are:

(i) Multiplicative model: $y_t = T_t \times S_t \times C_t \times I_t$

(ii) Additive model: $y_t = T_t + S_t + C_t + I_t$

Where y_t = value of the time series at time t

T_t = effect due to trend at time t

S_t = effect due to seasonal variation at time t

C_t = effect due to cyclical fluctuation at time t

I_t = effect due to irregular fluctuation at time t

There can be other models such as:

$\log y_t = \log T_t + \log S_t + \log C_t + \log I_t$

$y_t = T_t \times S_t \times C_t + I_t$, $y_t = T_t + S_t \times C_t \times I_t$ etc.

13. Discuss the semi average method for fitting trend to a time series.

Ans. According to this method, the observations in the time series are divided into two groups each containing equal number of consecutive observations. However if the series contains odd number of observations, the middle observation is excluded and the rest of the series is divided into two equal groups. The arithmetic means of the two groups are determined separately. These two averages are called the semi averages. The semi average of each group is plotted against the mid point of time of the corresponding group. The two points so obtained are joined by a straight line to form the trend of the given time series. The trend values of different points of time are the ordinates of the points corresponding to the respective points of time given in the series.

Merits:

(i) The method does not require complicated mathematical calculations so it is easy to understand and simple to apply.

(ii) It is an objective method because every one gets the same trend for the same time series.

(iii) It is capable of being used for forecasting the trend values because the trend is a straight line and so can be extended further.

(iv) It uses data completely i.e. it can provide trend values for each point of time in the series.

Demerits:

(i) The method presupposes the existence of a linear trend in every time series which may not be true in all cases.

(ii) It is not capable of giving non-linear trends.

(iii) It has the demerit of being unduly influenced by larger values in the time series. So the trend is more likely to contain effects due to irregular fluctuations.

(iv) Removing the middle observation in case of data having odd number of observations is purely arbitrary.

Uses:

This method can be used to measure the trend of a time series with minimum effort particularly when it is known to have a linear trend.

14. Define a time series giving suitable examples. Explain its various components.

Ans. A time series may be defined as the chronological arrangement of occurrences of an event presented preferably at equal intervals of time. Thus a time series is a bivariate data in which time is the independent variable.

Examples of time series are: (i) Annual production of rice in Odisha for a period of 15 years. (ii) Monthly sales of a departmental store for three consecutive years. (iii) Quarterly profits of a business house for 5 consecutive years. (iv) Daily sales of newspaper for six months etc.

The value of a time series is the combined effect due to a large number of factors. Such factors can be classified into four major categories called the components of the time series. The components of a time series are: (i) Trend or long term movement, (ii) Seasonal variations (iii) Cyclical fluctuations (iv) Random component or Irregular fluctuations. The two components Seasonal variations and Cyclical fluctuations combined together is named as Short term fluctuations or periodic changes. A time series may contain any one or more of the four components.

Trend or long term movement: It is the component of a time series which presents the general tendency of the values of the time series to increase or decrease with respect to time. So trend of a time series

can be studied only if the data is observed over a long period of time. Due to this reason, trend is also called the long term movement. Trend in a time series mainly arises out of the effect due to such factors which either do not change or show a slow and gradual change which can be marked only if the data is studied over a long period of time. Examples of factors due to which trend occurs are: Soil fertility, Changes in the taste and habit of a population, Changes caused from scientific inventions and discoveries etc. Sometimes it is observed that the trend shows the behaviour of remaining more or less constant or fluctuating between two fixed values for example the average temperature, average rainfall of a specific place remain almost constant every year.

Seasonal variations: In the analysis of time series, a season means a specific time interval of less than one year time. Popularly used seasons during the analysis of time series are Months and Quarters. A year is divided into 12 months and a quarter means a period of three consecutive months.

It is observed that time series data presented season wise is likely to show a uniform change periodically occurring at a regular interval of one year or less time. Such variations in the time series are referred to as seasonal variations. Seasonal variations are precise, definite and are predictable for future. So its study is highly essential for running a business successfully.

Seasonal variations in a time series mainly occur due to two types of forces namely (i) Natural forces, (ii) Social customs and traditions.

Examples of seasonal variation caused from natural forces are: (a) increase in the sale of ice-cream, cold drinks etc. during summer season. (b) Increase in the sale of woolen clothes during winter season. (c) Decrease in the price of agricultural commodities during their respective harvesting seasons etc.

Examples of seasonal variation caused from social customs and traditions are: (a) increase in sale of fire crackers during Diwali, (b) Increase in the sale of greeting cards during period of New year, (c) Increase in the sale of copies, books etc during admission season etc.

Cyclical fluctuations: These are the variations observed in a time series at more or less regular interval but their period of occurrence is more than one year. In other words these are oscillatory movements in the time series with period of oscillation more than one year. A complete period of oscillation is called the period of the cycle. Cyclical fluctuations in trade occur due to the business cycle. So the study of cyclical fluctuation helps in running a business smoothly. The stages of a business cycle are: Boom – Recession – Depression – Recovery – Boom.

Boom is the period showing the maximum profit and Depression is the period showing the minimum profit level. The time interval between Boom and Depression is called Recession and the time interval between Depression and the next Boom is called Recovery. The period of time between two consecutive Booms is called the period of the cycle. Some times the period of the cycle is not perfectly regular; so in such cases the average of all the periods of the cycles experienced by the time series is considered as the period of the cycle.

Examples of factors responsible for cyclical variations are: (i) changes in fashion, (ii) changes caused from scientific and technological developments etc.

Irregular fluctuations: The fluctuations in a time series which are purely random, erratic, uncertain and unpredictable are termed as irregular fluctuations. Such fluctuations are caused due to two types of forces namely (i) Natural forces such as flood, earthquake, cyclone etc., (ii) socio-economic activities such as war, strike, lock-out, terrorist activities, communal riot etc. Since these fluctuations are unpredictable it is very difficult to isolate them.

These fluctuations sometimes become very effective and may give rise to seasonal as well as cyclical fluctuations. For example flood in the coastal districts of Odisha is experienced in the rainy season of every year. So it becomes a seasonal variation.

15. Describe the method for fitting a trend to a time series by the free hand curve.

Ans: It is the simplest method to measure trend in a time series. According to this method, the points of time are plotted along the x-axis and the corresponding values of the time series are plotted along the y-axis. The points obtained are joined serially by straight lines. The diagram so obtained is called the historigram. This diagram contains crests and troughs which are due to the effects of components other than the trend. So in order to eliminate them a free hand smooth curve is drawn through the historigaram which represents the

trend in the time series. A proper trend curve should satisfy the following conditions:

(i) It should be a smooth curve.

(ii) The number of points of the historigram on both sides of the trend curve should be equal.

(iii) The sum of deviations parallel to the y-axis should be zero.

(iv) The sum of squares of deviations parallel to the y-axis should be the minimum.

Merits:

(i) The method is easy to understand and does not require much knowledge of mathematics. So it can be adopted by any person drawing the trend.

(ii) It is capable of providing linear as well as non-linear trend.

(iii) It is a flexible method.

Demerits:

(i) It is not an objective method because the trend curve is drawn only through eye-inspection. So, different persons are likely to get different trend curves for the same data.

(ii) The suitability of the trend curve is highly dependent upon the experience, skill and understanding of the person drawing the trend.

(iii) Since the method is not based on mathematical approach; it may not be suitable for forecasting.

Uses:

Despite serious limitations, it is an indispensable method for the measurement of trend because it is the primary method that can suggest the type of trend that may be present in the time series so that a proper mathematical method can be adopted to find the suitable trend.

16. Explain the method of least squares for fitting a linear trend. Also fit a linear trend to a time series with 2n yearly observations.

Ans. Let the equation of the linear trend be $y = a + bt$

Where a and b are real numbers,

y = trend value of the time series

t = variable representing time

Then according to method of least squares, this line passes through the historigram in such a way that the sum of squares of deviations parallel to the y-axis becomes the minimum. Let the time series has n observations of the form (t_i, u_i). Hence the trend value corresponding to time t_i is $y_i = a + bt_i$

Let the sum of squares of deviations parallel to the y-axis be S.

$$\text{Then } S = \sum_{i=1}^{n}(u_i - y_i)^2 = \sum_{i=1}^{n}(u_i - a - bt_i)^2$$

Thus it is needed to find the values of a and b such that S becomes the minimum.

Applying the principle of minima,

$$\frac{\partial S}{\partial a} = 0 \Rightarrow -2\sum_{i=1}^{n}(u_i - a - bt_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} u_i = na + b\sum_{i=1}^{n} t_i \quad ...(i)$$

$$\frac{\partial S}{\partial b} = 0 \Rightarrow -2\sum_{i=1}^{n} t_i(u_i - a - bt_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} u_i t_i = a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} t_i^2 \quad ...(ii)$$

Equations (i) and (ii) are called the normal equations for fitting a linear trend. On solving these equations, the values of a and b are obtained and by putting these values of a and b in the 2nd order derivatives of S with respect to a and b the results become positive. So these values of a and b provide the minimum value of S and hence by putting them in the trend equation, the required linear trend is obtained.

If the values of t_i are assigned in such away that

$$\sum_{i=1}^{n} t_i = 0 \text{ then } a = \frac{\sum_{i=1}^{n} u_i}{n} \text{ and } b = \frac{\sum_{i=1}^{n} u_i t_i}{\sum_{i=1}^{n} t_i^2}$$

Thus the equation of the linear trend for a time series having n observations of the form (t_i, u_i) is

$$y = \frac{\sum_{i=1}^{n} u_i}{n} + \left(\frac{\sum_{i=1}^{n} u_i t_i}{\sum_{i=1}^{n} t_i^2}\right) t$$

The trend values corresponding to different values of t are the values of y obtained by putting respective values of t in the trend equation.

While fitting a linear trend to a time series with 2n yearly observations, t = 2× (year-middle year in the series) so that $\sum_{i=1}^{n} t_i = 0$.

Then $t = -(2n-1), -(2n-3), -(2n-3), -(2n-5)$, ... $-1, 1, 3, 5, ...(2n-1)$

The values of Σt_i, Σu_i, $\Sigma u_i t_i$, Σt_i^2 are computed from the given time series and by solving the normal equations the values of a and b can be determined. These values are substituted in the equation $y = a + bt$ to get the required equation of the linear trend.

17. Explain moving average method for fitting trend to a time series.

Ans. According to the moving average method, the observations in the series are divided into overlapping groups of equal number of consecutive observations in such a way that there should be difference of exactly one observation between two consecutive groups. The number of observations taken in each group is called the period of moving average. The process of determination of the trend values differ from the period of moving average becoming odd or even.

Odd period moving average:

Let the period of moving average be odd (say) 5 years, then the overlapping groups of observations are:

1st group = 1st year to 5th year, 2nd group = 2nd year to 6th year and so on.

The average of each group called the moving average represents the trend value for the mid point of time of the corresponding group.

Thus average of the 1st group = trend value of the 3rd year

Average of the 2nd group = trend value of the 4th year and so on.

The moving averages are plotted on the graph against the points of time to which they correspond. The points obtained are joined serially by a free hand smooth curve to provide the trend of the time series.

Even period moving average:

Let the period of moving average be even (say) 4 years then the overlapping groups of observations are:

1st group = 1st year to 4th year, 2nd group = 2nd year to 5th year and so on. The averages of these groups called the 4 yearly moving averages do not correspond to specific years, so they do not represent the trend values. Hence if the period of moving average is even, the trend values are computed by the method of centering. The centered moving averages are obtained by averaging overlapping 4 yearly moving averages taking two consecutive figures at a time. These centered moving averages are the trend values for the mid point of time of the respective two consecutive 4 yearly moving averages i.e.

Trend value of the 3rd year = Centered moving average of the 1st and 2nd 4 yearly moving averages.

Trend value of the 4th year = Centered moving average of the 2nd and 3rd 4 yearly moving averages and so on.

The trend values are plotted on the graph against the points of time to which they correspond. The points obtained are joined serially by a free hand smooth curve to provide the trend of the time series.

Merits:

(i) The method is simple to understand and easy to use because it does not require complicated mathematical calculations.

(ii) It is an objective method because every one gets the same trend for the same time series.

(iii) It is capable of giving linear as well as non-linear trends.

(iv) The method is most suitable for measurement of trend in a time series having a cyclical fluctuation occurring at regular intervals. In such cases the period of the time series should be taken equal to the period of the cycle.

Demerits:

(i) The method fails to provide trend values for all the points of time given in the series.

(ii) If the period of the cycle is not perfectly regular, the trend is likely to contain some effect due to cyclical fluctuations.

(iii) It may not be always suitable for forecasting the trend values because the trend is a free hand curve and is not expressed as a function of time.

Uses:

The method can be used for measurement of trend in time series relating to business which is likely to contain cyclical fluctuations.

18. Fit a quadratic trend to a time series with $2n+1$ observations at equal intervals.

Ans. Let the equation of the quadratic trend be $y = a + bt + ct^2$.

Let the $(2n+1)$ yearly observations be (t, u)

Where t_i = Year - middle year of the series. Then $t = -n, -(n-1), \ldots -1, 0, 1, 2, \ldots n$

The three normal equations for fitting the quadratic trend are :

$$\sum_{i=1}^{n} u_i = na + b\sum_{i=1}^{n} t_i + c\sum_{i=1}^{n} t_i^2 \cdots (i)$$

$$\sum_{i=1}^{n} u_i t_i = a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} t_i^2 + c\sum_{i=1}^{n} t_i^3 \cdots (ii)$$

$$\sum_{i=1}^{n} u_i t_i^2 = a\sum_{i=1}^{n} t_i^{2i} + b\sum_{i=1}^{n} t_i^3 + c\sum_{i=1}^{n} t_i^4 \cdots (iii)$$

Here $\sum_{i=1}^{n} t_i = 0$ and $\sum_{i=1}^{n} t_i^3 = 0$. So the normal equations become:

$$\sum_{i=1}^{n} u_i = na + c\sum_{i=1}^{n} t_i^2 \cdots (iv)$$

$$\sum_{i=1}^{n} u_i t_i = b\sum_{i=1}^{n} t_i^2 \cdots (v)$$

$$\sum_{i=1}^{n} u_i t_i^2 = a\sum_{i=1}^{n} t_i^2 + c\sum_{i=1}^{n} t_i^4 \cdots (vi)$$

The values of a, b and c can be obtained by solving the equations (iv), (v) and (vi). These values can be substituted in the equation $y = a + bt + ct^2$ to obtain the equation of the required quadratic trend.

19. Draw comparison between:

(a) Least square method and semi average method

Ans. (i) Least square method can provide linear as well as non-linear trends but semi average method can give only linear trend.

(ii) The trend obtained by the least square method completely eliminates the effect due to other components but trend by semi average method is likely to be affected by irregular fluctuations and cyclical fluctuations.

(iii) Least square method can be used to determine the effects of components other than that of trend but semi average method cannot.

(iv) Semi average method does not use the middle observation when the series has odd number of observations but in least square method all the observations are taken into consideration at the time of fitting the trend.

(v) Semi average method is easier to understand as compared to the method of least squares.

(b) Least square method and moving average method

Ans. (i) Moving average method does not provide the trend values for all the points of time given in the series but least square method gives the trend values for all the points of time.

(ii) Trend obtained by the least square method is in the form of a function of time; so it is more suitable for forecasting the trend values for future points of time whereas moving average method may not always be dependable for future forecasting.

(iii) Moving average method in comparison to the least square method is more suitable for measurement of trend in a time series having a regular period of cyclical fluctuation.

(c) Least square method and free hand curve method

Ans. (i) The trend obtained by the free hand curve method is highly subjective in nature but that obtained by the least square method is objective.

(ii) Free hand curve method is an eye-inspection method but least square method is logical.

(iii) Trend obtained by the least square method is in the form of a function of time; so it is more capable of forecasting the trend values for future points of time whereas free hand curve method may not be dependable for future forecasting.

(d) Moving average method and semi average method

Ans. (i) Moving average method can provide linear as well as non-linear trends but semi average method can give only linear trend.

(ii) Moving average method does not provide the trend values for all the points of time given in the series but semi average method gives the trend values for all the points of time.

(iii) As compared to moving average method, semi average method is more dependable for forecasting since the trend is a straight line which can be extended further.

(e) Moving average method and free hand curve method

Ans. (i) Free hand curve method is not objective but moving average method is an objective method.

(ii) Free hand curve method is an eye-inspection method but moving average method has logical basis.

(f) Semi average method and free hand curve method

Ans. (i) Free hand curve method is not objective but semi average method is an objective method.

(ii) Free hand curve method can provide linear as well as non-linear trends but semi average method can give only linear trend.

(iii) Free hand curve method is not suitable for forecasting but semi average method can be used for forecasting.

INDEX NUMBER

20. Explain the various types of index numbers.

Ans. The different types of index numbers can be broadly classified into four major categories on the basis of the characteristic whose relative change is under consideration.

They are:

(i) **Price index numbers :** which are further divided into two types such as wholesale price index numbers and Retail price index numbers. A price index number measures the relative change in the price level of a group of commodities with respect to time or space or due to any other characteristic.

(ii) **Quantity index numbers :** A Quantity index number measures the relative change in the level of the quantity consumed sold, produced, exported etc with respect to time or space or due to any other characteristic.

(iii) **Value index numbers :** Such an index number measures the relative change in the value of a group of commodities with respect to time or space or due to any other characteristic. Here value of a commodity means the total amount of money spent in the purchase of the commodity.

So Value = Price × Quantity

(iv) **Special purpose index numbers :** The index numbers which do not fall in any of the three categories mentioned above are called special purpose index numbers. For example the index number showing the relative change in the level of the business activities of a place with respect to time is a special purpose index number.

21. Define an index number with suitable examples and describe its uses.

Ans. An index number may be defined as the ratio representing the relative change in the level of a phenomenon or a group of phenomena with respect to time or space or due to any other characteristic. It is always a pure number which is free from units. Examples of index numbers are: (i) wholesale price index number, (ii) cost of living index number etc.

Uses:

Some of the important uses of index numbers are given below:

(i) They help in studying trends and tendencies of time series. Such studies are highly helpful for making plans for future.

(ii) Index numbers are used by the Government and other organisatios to frame policies regarding fixation of D.A and other allowances to be paid to the employees.

(iii) They are used by the Government to take decisions regarding taxes to be levied.

(iv) They are used by the trade unions for wage negotiations.

(v) Used to find the purchasing power of money.

(vi) Used to determine the real wage.

(vii) Used for deflating various values like nominal wage, nominal sales, nominal exports etc.

(viii) They act as economic barometers to measure the pressure of the economy on the people of a state.

22. What is meant by a Base Year? How the Base Year can be selected for construction of an index number?

Ans. The base year is that year from among the past years with whose price level the current year's price level is to be compared. Index numbers can be constructed by two methods on the basis of the selection of the base year. They are (i) Fixed base method and (ii) Chain base method.

According to the fixed base method, a specific year from among the past years is taken as the base year and the price level of all other years are compared with the price level of that base year. Hence an ideal base year

under fixed base method should have the following characteristics.

(a) It should be a normal year from economic point of view i.e the price level should be neither very high nor very low during the base year.

(b) It should be free from irregular fluctuations such as flood, cyclone, earthquake, war etc which are likely to cause economic instability.

(c) The base year should not be far away from the current year because in such cases some commodities might have become absolete or some new commodities might have entered into use due to scientific developments.

Thus fixed base method can be used to compare the price level of a year with that of a recent past year.

According to chain base method, the year preceding the current year is taken as the base year. Hence this method can be useful for comparing the price level of a year with a distant past year.

23. Draw comparison between:

(a) Unweighted and Weighted index numbers.

Ans. The formulae for the construction of a price index number can be broadly classified into two categories:

(I) Unweighted index number formulae in which all the commodities are assumed to have equal importance. So for computing an unweighted index number, no specific weight is assigned to any commodity.

(i) Unweighted Aggregative Method:

$$P_{01} = \frac{\sum_{i=1}^{n} p_{1i}}{\sum_{i=1}^{n} p_{0i}} \times 100$$

(ii) Unweighted A.M of relatives method:

$$P_{01} = \left[\frac{1}{n}\sum_{i-1}^{n}\left(\frac{p_{1i}}{p_{oi}}\right)\right] \times 100$$

(iii) Unweighted G.M of relatives method:

$$P_{01} = \left[Anti\log\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p_{1i}}{p_{oi}}\right)\right\}\right] \times 100$$

(II) Weighted index number formulae in which different commodities are assigned with specific commodities in accordance with their relative importance.

(i) Weighted Aggregative method:

$$P_{01} = \frac{\sum_{i=1}^{n} p_{1i} w_i}{\sum_{i=1}^{n} p_{0i} w_i} \times 100$$

(ii) Laspeyre's method: $P_{01}^{La} = \dfrac{\sum_{i=1}^{n} p_{1i} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}} \times 100$

(iii) Paasche's method: $P_{01}^{Pa} = \dfrac{\sum_{i=1}^{n} p_{1i} q_{1i}}{\sum_{i=1}^{n} p_{0i} q_{1i}} \times 100$

(iv) Fisher's Ideal method:

$$P_{01}^{Id} = \sqrt{P_{01}^{La} \times P_{01}^{Pa}} = \left[\sqrt{\frac{\sum_{i=1}^{n} p_{1i} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}} \times \frac{\sum_{i=1}^{n} p_{1i} q_{1i}}{\sum_{i=1}^{n} p_{0i} q_{1i}}}\right] \times 100$$

(v) Weighted A.M of relatives method:

$$P_{01} = \left[\frac{\sum_{i=1}^{n} w_i\left(\frac{p_{1i}}{p_{0i}}\right)}{\sum_{i=1}^{n} w_i}\right] \times 100$$

(vi) Weighted G.M of relatives method:

$$P_{01} = \left[Anti\log\left\{\frac{\sum_{i=1}^{n} w_i \log\left(\frac{p_{1i}}{p_{0i}}\right)}{\sum_{i=1}^{n} w_i}\right\}\right] \times 100$$

(b) Aggregative method and Average of relatives method.

Ans. According to the aggregative method the ratio is determined between the aggregate of the current year's figure and the aggregate of the base year's figure. The

formulae of index number formulae under the aggregative method are:

(i) Unweighted Aggregative Method:

$$P_{01} = \frac{\sum_{i=1}^{n} p_{1i}}{\sum_{i=1}^{n} p_{0i}} \times 100$$

(ii) Weighted Aggregative method:

$$P_{01} = \frac{\sum_{i=1}^{n} p_{1i} w_i}{\sum_{i=1}^{n} p_{0i} w_i} \times 100$$

(iii) Laspeyre's method:

$$P_{01}^{La} = \frac{\sum_{i=1}^{n} p_{1i} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}} \times 100$$

(iv) Paasche's method:

$$P_{01}^{Pa} = \frac{\sum_{i=1}^{n} p_{1i} q_{1i}}{\sum_{i=1}^{n} p_{0i} q_{1i}} \times 100$$

The formulae of index number under the average of relatives method are determined by taking the average of the price relatives. They are:

(i) Unweighted A.M of relatives method:

$$P_{01} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{p_{1i}}{p_{oi}}\right] \times 100$$

(ii) Unweighted G.M of relatives method:

$$P_{01} = \left[Anti\log\left\{\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{1i}}{p_{oi}}\right)\right\}\right] \times 100$$

(iii) Weighted A.M of relatives method:

$$P_{01} = \left[\frac{\sum_{i=1}^{n} w_i\left(\frac{p_{1i}}{p_{0i}}\right)}{\sum_{i=1}^{n} w_i}\right] \times 100$$

(iv) Weighted G.M of relatives method:

$$P_{01} = \left[Anti\log\left\{\frac{\sum_{i=1}^{n} w_i \log\left(\frac{p_{1i}}{p_{0i}}\right)}{\sum_{i=1}^{n} w_i}\right\}\right] \times 100$$

(c) Fixed base method and Chain base method.

Ans. Under the fixed base method, a specific year from among the past years is taken as the base year for comparing the values of all other years. This method is more suitable when the time interval between the current year and the base year is not large.

Chain base method is that in which the year preceding the current year is taken as the base year. So this method is suitable for computing the relative change with respect to a distant past year.

(d) Laspeyre's method and Paasche's method.

Ans. (i) According to Laspeyre's method, weight of a commodity is the quantity of the commodity consumed in the base year but according to Paasche's method, weight of a commodity is the quantity of the commodity consumed in the current year. (ii) If prices show an increasing trend, Laspeyre's method shows an upward bias i.e it overestimates the relative change whereas Paasche's method shows a downward bias i.e it underestimates the relative change. (iii) If the base year remains fixed, Laspeyre's method needs data on quantity only for once but even for fixed base method, Paasche's method requires quantity data every time.

24. What are the problems in the construction of an index number?

Ans. A number of problems are likely to occur during the construction of an index number which need to be handled carefully to avoid misleading conclusions. Such problems are:

(i) **Selection of the base period :** The base period is the time period with which the comparison is to be made. Generally it is taken as a period of one year from among the past years and hence it is also called the base year. In order to get a suitable measure of the relative change, it is highly important to select a proper base year free from economic abnormalities and irregular fluctuations. It is a much difficult task to select such a year which is absolutely normal from economic point of view. So selection of base period is considered to be a problem in the construction of index number.

(ii) **Selection of commodities :** Ideally, all commodities in use should be taken into consideration for the construction of index number but practically it is not possible to take into account all the commodities. Hence

representative commodities are chosen at the time of constructing the index number. Since the use of commodities differ according to place, age, sex and consumption pattern, the selection of representative commodities becomes a problem in the construction of an index number.

(iii) **Selection of places for collection of price quotations :** Price quotations for the selected commodities should be collected from all the places where the commodities are sold. But again it is not practicable. So some representative markets need to be selected for collecting the price quotations. It is not easy to select such representative places. So it is considered a problem in the construction of index number.

(iv) **Assignment of weights :** All the commodities do not have equal importance. Hence it is necessary to assign rational weights to the selected commodities in accordance with their importance of use. Assigning such weights is not easy because the importance of the commodities also depends on the consumption pattern.

25. Explain the reasons due to which Fisher's index is called "Ideal".

Ans. Fisher's index has the following merits for being called the ideal index:

(i) It uses the geometric mean which is the most suitable average for construction of index number.

(ii) Being the average it is likely to cancel out the overestimation due to Laspeyre's method and the underestimation due to Paasche's method.

(iii) It satisfies maximum number of tests for the adequacy of an index number formula.

26. What is splicing? Explain it on the basis of a hypothetical numerical example.

Ans. When two overlapping series of index numbers are known, one series of index numbers can be computed with respect to the base of the other series. The process is called splicing.

Example: The following table presents two overlapping series A and B of index numbers and series – B is spliced into series – A.

Year	Series – A (Base = 2010)	Series – B (Base = 2015)	Spliced series – A
2010	100		100
2011	120		120
2012	125		125
2013	150	100	150
2014		110	$110 \times \frac{150}{100} = 165$
2016		120	$120 \times \frac{150}{100} = 180$
2017		140	$140 \times \frac{150}{100} = 210$

27. Write a note on the tests of adequacy of an index formula and also verify them for various index number formulae.

Ans. There are four tests of adequacy of an index number formula. They are:

(i) **Unit test:** According to this test, an index number formula should be independent of the units in which price and quantity are measured. All index number formulae except the unweighted aggregative method satisfy this test.

(ii) **Time reversal test:** As suggested by Prof. Irving Fisher, an index number formula should allow the interchange of the two time periods of comparison, base period and current period giving the same ratio of comparison. Mathematically, the index number formula for which $P_{01} \times P_{10} = 1$, satisfies the time reversal test.

(iii) **Factor reversal test:** Fisher suggested that a good index number formula should allow the interchange of the factors price and quantity without giving inconsistent results.

Mathematically, if an index number formula is such that $P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$ = True value ratio then the formula is said to satisfy factor reversal test.

(iv) **Circular test:** It is an extension over the time reversal test where the number of periods of comparison is more than two. Let the periods of comparison be 0, 1, 2, ..., n

If an index number formula satisfies the circular test, then $P_{01} \times P_{12} \times P_{23} \times .... \times P_{n0} = 1$

While examining an index number formula to satisfy a test, it should not be expressed as percentage.

28. Describe the various steps in the construction of a price index number. Also write the formulae for constructing a price index number along with the data required for their computation.

Ans. Various steps im the construction of a price index number are as follows:

(i) Purpose of the index: This is the most important step in the construction of a price index number. It means to specify the particular class of people for whom the index number will be used. For example lower income group, higher income group, Government employees, agricultural labourers, industrial workers etc. It is observed that the consumption pattern differs from one class of people to another resulting in the use of different commodities, their qualities and their relative importance. So a clearly mentioned purpose helps in the selection of commodities and assigning proper weights to the commodities.

(ii) Selection of base period : The base period is the time period with which the comparison is to be made. Generally it is taken as a period of one year from among the past years and hence it is also called the base year. Selection of the base year can be done by two methods such as:

(I) Fixed base method : According to the fixed base method, a specific year from among the past years is taken as the base year and the price level of all other years are compared with the price level of that base year. Hence an ideal base year under fixed base method should have the following characteristics.

(a) It should be a normal year from economic point of view i.e the price level should be neither very high nor very low during the base year.

(b) It should be free from irregular fluctuations such as flood, cyclone, earthquake, war etc which are likely to cause economic instability.

(c) The base year should not be far away from the current year because in such cases some commodities might have become obsolete or some new commodities might have entered into use due to scientific developments.

Thus fixed base method can be used to compare the price level of a year with that of a recent past year.

(II) Chain base method : According to chain base method, the year preceding the current year is taken as the base year. Hence this method can be useful for comparing the price level of a year with a distant past year.

The selection of the base period depends on the purpose of the index, availability of data and resources.

(iii) Selection of commodities: For the purpose of constructing a price index number, it is necessary to select some representative commodities keeping in view the following points:

(a) The commodities selected should be representative of the necessity, taste, habit, customs and tradition of the class of people selected for the purpose.

(b) They should be stable in quality.

(c) All qualities of a commodity which are in common use by the group of people under consideration should be included in the selection.

(d) There is no fixed rule regarding the number of commodities to be selected. But the number of commodities should be neither very large nor very small. A large number of commodities may lead to difficulty in data management and a very small number of commodities may not be able to reveal the proper relative change. However number of commodities to be selected should be decided on the basis of availability of data and resources.

(iv) Collection of price quotations: Price quotations need to be collected for the selected commodities from representative places. In order to construct a wholesale price index number, price quotations are collected from wholesale markets and for constructing a retail price index number, price quotations need to be collected from retail markets. The following points should be taken into account while collecting the price quotations.

(a) They should be related to the specific quality of the commodity selected.

(b) They should be in the form of price per unit quantity and not in the form of quantity per unit of money.

(c) The price quotations should take into account the discounts allowed on cash payment and the interest charged on late payment.

(d) The price quotations should include controlled or subsidized prices if any.

After collecting the price quotations for a commodity from all the selected places, they are averaged to find the average price of the commodity.

(v) Assigning weights: Weight of a commodity in the context of construction of index number refers to a numerical value of the relative importance of the commodity in the group of commodities selected. It is known that all commodities used do not have equal importance. The importance of a commodity depends on the taste, habit, etc of the group of people under consideration. For example the importance of wheat is more as compared to that of rice in the northern states of India. But rice is more important in the southern and eastern parts of India. Thus it is essential to assign proper and rational weights to the selected commodities in accordance with their relative importance. Such assignment of weights can be done by two processes such as: (a) Quantity weights- where the weight of a commodity is the quantity of the commodity consumed. (b) Value weights- where the weight of a commodity is the value of the commodity i.e the amount of money spent in purchasing the commodity.

Selection of the method of assigning weights depends on the availability of data and resources.

(vi) Selection of average: It is known that an index number is a specialized type of average. So selection of a proper average plays an important role in the construction of the index number. Theoretically, any one of the five averages namely arithmetic mean, median, mode, geometric mean and harmonic mean can be used as an average; but while making a choice, it is observed that A.M, Median and Mode are absolute measures. Further A.M has a demerit of being unduly influenced by higher values and Median and Mode are not based on all observations. Hence A.M, Median and Mode are not suitable for construction of index number. Harmonic mean is a rate measurer and is unduly affected by small values in the data. So H.M is also not found to be suitable for constructing an index number. The only average left is the geometric mean which is a measure of relative change. It is also observed that the index number formula based on geometric mean satisfies maximum number of tests of adequacy of an index number formula. So geometric mean is considered as the most suitable average for the construction of an index number.

(vii) Selection of formula: The formulae for the construction of a price index number can be broadly classified into two categories:

(a) Unweighted index number formulae in which all the commodities are assumed to have equal importance. So for computing an unweighted index number, no specific weight is assigned to any commodity.

(b) Weighted index number formulae in which different commodities are assigned with specific weights in accordance with their relative importance.

Notations and data required :

P_{01} = Price index for the current year denoted by 1 with respect to the base year denoted by 0.

p_{0i} = Price of the i^{th} commodity in the base year.

p_{1i} = Price of the i^{th} commodity in the current year.

q_{0i} = Quantity of the i^{th} commodity in the base year.

q_{1i} = Quantity of the i^{th} commodity in the current year.

n = Number of commodities

w_i = Weight of the i^{th} commodity

Different Index Number Formulae :

(a) Unweighted Aggregative Method :

$$P_{01} = \frac{\sum_{i=1}^{n} p_{1i}}{\sum_{i=1}^{n} p_{0i}} \times 100$$

(b) Unweighted A.M. of relatives method :

$$P_{01} = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{p_{1i}}{p_{oi}}\right] \times 100$$

(c) Unweighted G.M. of relatives method :

$$P_{01} = \left[Anti\log\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p_{1i}}{p_{oi}}\right)\right\}\right] \times 100$$

(d) Weighted Aggregative method :

$$P_{01} = \frac{\sum_{i=1}^{n} p_{1i} w_i}{\sum_{i=1}^{n} p_{0i} w_i} \times 100$$

(e) Laspeyre's method : $P_{01}^{La} = \dfrac{\sum_{i=1}^{n} p_{1i} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}} \times 100$

(f) Paasche's method : $P_{01}^{Pa} = \dfrac{\sum_{i=1}^{n} p_{1i} q_{1i}}{\sum_{i=1}^{n} p_{0i} q_{1i}} \times 100$

(g) Fisher's Ideal method :

$$P_{01}^{Id} = \sqrt{P_{01}^{La} \times P_{01}^{Pa}} = \left[\sqrt{\frac{\sum_{i=1}^{n} p_{1i} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}} \times \frac{\sum_{i=1}^{n} p_{1i} q_{1i}}{\sum_{i=1}^{n} p_{0i} q_{1i}}}\right] \times 100$$

(h) Weighted A.M. of relatives method :

$$P_{01} = \left[\frac{\sum_{i=1}^{n} w_i \left(\frac{p_{1i}}{p_{0i}}\right)}{\sum_{i=1}^{n} w_i}\right] \times 100$$

(i) Weighted G.M. of relatives method :

$$P_{01} = \left[Anti\log\left\{\frac{\sum_{i=1}^{n} w_i \log\left(\frac{p_{1i}}{p_{0i}}\right)}{\sum_{i=1}^{n} w_i}\right\}\right] \times 100$$

SAMPLING TECHNIQUES

29. What is Simple Random Sampling and what are the different methods by which a simple random sample can be drawn.

Ans. Simple Random Sampling (SRS) is that method of sampling where all the units of the population are assumed to be homogeneous and every unit of the population is assigned with the equal probability of being included in the sample. There are two types of Simple Random Samples namely Simple Random Sample with Replacement (SRSWR) and Simple Random Sample without Replacement (SRSWOR). In case of SRSWR a unit of the population can be enumerated in the sample more than once but in SRSWOR a population unit is allowed to be enumerated only once in the sample.

The different methods by which a simple random sample can be drawn are: (i) Lottery method (ii) By using a Random Number Table. While selecting a simple random sample, the first step is to assign a serial number to every unit of the population. Thus if a population consists of N units the they should be assigned with serial numbers 1, 2, . . . , N.

(i) **Lottery method:** According to this method, each of the serial numbers 1, 2, ... , N is written in a piece of paper and all these paper pieces are folded identically in such a way that the number written on them should not be visible without unfolding. Proper care should be taken to see that all the paper pieces should be identical in shape, size, colour quality etc. Such folded paper pieces called lots are kept in a container and the number of units required for the sample are drawn one after another either with replacement or without replacement depending on the type of sample (SRSWR or SRSWOR) required.

(ii) **Using a random number table:** A Random Number Table (RNT) is an arrangement of the digits 0, 1, 2, ..., 9 in rows and columns in such a way that each digit has an equal frequency of occurrence and no fixed relation exists between rows or columns or diagonals etc. Standard RNT available are Tippett's RNT, Fisher's RNT etc. Using any RNT,

random numbers having number of digits equal to that in N are selected starting from any place of the table and then moving continuously in some fixed direction. A random number can begin with zero but random numbers with all zeros and those random numbers larger than the largest multiple of N having equal number of digits as in N should be excluded from the selection. However if no other multiple of N with equal number of digits as in N are there then no random number is excluded from the selection. The selection of units is done by adopting the following procedure:

If the random number selected is smaller than or equal to N, then the population unit bearing that serial number is included in the sample.

If the random number selected is larger than N but is not under the exclusion set, it is divided by N and the population unit bearing the serial number as the remainder is included in the sample. However if the remainder is zero then the N^{th} unit is included in the sample. The inclusion of the population units depends on the type of sample (SRSWR or SRSWOR) required. The process continues until required number of units are included in the sample.

30. Draw a comparison between sampling and complete enumeration.

Ans. (i) Complete enumeration is the process of collecting data from every unit of the population; but in sampling data is collected from only some selected units. So in complete enumeration or census, the number of units from which data is to be collected is sufficiently larger than that in sampling.

(ii) Census requires more time but sampling is speedier.

(iii) Census is costly in comparison to sampling.

(iv) Census needs to engage a larger manpower as compared to that required for sampling.

(v) Sampling can have a better administrative control in comparison to Census.

(vi) Since the volume of data to be collected is less in sampling, it is capable of giving a wider scope whereas census provides only a limited scope of study.

(vii) Census cannot be conducted for experiments having a chance of damage of a machine or loss of a life in the course of experimentation. In such cases the only alternative method is sampling.

(viii) The results obtained by the census method are to be accepted as the true results but in sampling method precision of the estimate can be determined.

(ix) Sampling is likely to contain both sampling error and non-sampling error but census is affected by only non-sampling error.

31. Prove that in simple random sampling, sample mean is an unbiased estimate of population mean.

Ans. Simple mean $= \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Let us define the variable α_i as

$$\alpha_i = \begin{cases} 1, \text{if } Y_i \text{ is included in the simaple} \\ 0, \text{if } Y_i \text{ is not included in the simple} \end{cases}$$

Then $\sum_{i=1}^{n} y_i = \sum_{i=1}^{N} \alpha_i Y_i$

Further $E(\alpha_i) = 1 \times \frac{n}{N} + 0 \times \left(1 - \frac{n}{N}\right) = \frac{n}{N}$

[Because it is known that the probability of inclusion of Yi in the sample $= \frac{n}{N}$]

$$E(\bar{y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = E\left(\frac{1}{n}\sum_{i=1}^{N} \alpha_i Y_i\right) = \frac{1}{n}\sum_{i=1}^{N} Y_i E(\alpha_i)$$

$$= \frac{1}{n}\sum_{i=1}^{N} Y_i \frac{n}{N} = \frac{1}{n} \times \frac{n}{N}\sum_{i=1}^{N} Y_i = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y}$$

= Population mean.

32. Explain the meaning of the sampling distribution of a statistic. Also construct the sampling distribution of mean for a simple random sample of 3 units selected without replacement from a population of 5 units.

Ans. The sampling distribution of a statistic may be defined as the set of the values of the statistic that can be computed from all possible samples of some particular type from a population. For example let the

population size be N and a simple random sample of n units is to be drawn from it without replacement. Then the number of possible sample that can be drawn = ${}^{N}C_{n}$. Let the statistic be sample mean. Then sample mean has ${}^{N}C_{n}$ possible values. This set of ${}^{N}C_{n}$ possible values of the sample mean is called the sampling distribution of mean.

Let the population units be Y_1, Y_2, Y_3, Y_4, Y_5. Taking simple random samples from the population without replacement, the number of possible samples = ${}^{5}C_{3} = 10$. The following table presents the sampling distribution of mean.

Sampling Distribution of Mean

Sample Number	Units included	Sample mean
1	Y_1, Y_2, Y_3	$\frac{Y_1+Y_2+Y_3}{3}$
2	Y_1, Y_2, Y_4	$\frac{Y_1+Y_2+Y_4}{3}$
3	Y_1, Y_2, Y_5	$\frac{Y_1+Y_2+Y_5}{3}$
4	Y_1, Y_3, Y_4	$\frac{Y_1+Y_3+Y_4}{3}$
5	Y_1, Y_3, Y_5	$\frac{Y_1+Y_3+Y_5}{3}$
6	Y_1, Y_4, Y_5	$\frac{Y_1+Y_4+Y_5}{3}$
7	Y_2, Y_3, Y_4	$\frac{Y_2+Y_3+Y_4}{3}$
8	Y_2, Y_3, Y_5	$\frac{Y_2+Y_3+Y_5}{3}$
9	Y_2, Y_4, Y_5	$\frac{Y_2+Y_4+Y_5}{3}$
10	Y_3, Y_4, Y_5	$\frac{Y_3+Y_4+Y_5}{3}$

32. Derive the variance of sample mean in simple random sampling.

Ans. In simple random sampling

Sample mean $= \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and

Population mean $= \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

Further, it is known that sample mean is an unbiased estimate of the population mean i.e.

Sample mean $= E(\bar{y}) = \bar{Y}$

Let us define

$$\alpha_i = \begin{cases} 1 \text{ if } Y_i \text{ is included in the sample} \\ 0 \text{ if } Y_i \text{ is not included in the sample} \end{cases}$$

So Variance of Sample mean $= V(\bar{y})$

$$= E[(\bar{y}) - E(\bar{y})^2] = E[(\bar{y}) - \bar{Y}]^2$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n} y_i - \bar{Y}\right]^2 = \frac{1}{n^2} E\left[\sum_{i=1}^{n} y_i - n\bar{Y}\right]^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n} (y_i - \bar{Y})\right]^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n} (y_i - \bar{Y})^2\right] + \frac{1}{n^2} E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} (y_i - \bar{Y})(y_j - \bar{Y})\right]$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{N} \alpha_i (y_i - \bar{Y})^2\right]$$

$$+ \frac{1}{n^2} E\left[\sum_{i=1}^{N}\sum_{j\neq i}^{N} (y_i - \bar{Y})(y_j - \bar{Y}) E(\alpha_i \alpha_j)\right]$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{N} (y_i - \bar{Y})^2 E(\alpha_i)\right]$$

$$+ \frac{1}{n^2} E\left[\sum_{i=1}^{N}\sum_{j\neq i}^{N} (y_i - \bar{Y})(y_j - \bar{Y}) E(\alpha_i \alpha_j)\right]$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{N} (y_i - \bar{Y})^2 \left\{1 \times \frac{n}{N} + 0 \times \left(1 - \frac{n}{N}\right)\right\}\right]$$

$$+ \frac{1}{n^2} E\left[\sum_{i=1}^{N}\sum_{j\neq i}^{N} (y_i - \bar{Y})(y_j - \bar{Y})\right.$$

$$\left.\left\{1 \times \frac{n(n-1)}{N(N-1)} + 0 \times \left(1 - \frac{n(n-1)}{N(N-1)}\right)\right\}\right]$$

$$= \frac{1}{nN}\left[\sum_{i=1}^{N} (y_i - \bar{Y})^2\right] + \frac{1}{n^2}$$

$$\times \frac{n(n-1)}{N(N-1)}\left[\sum_{i=1}^{N}\sum_{j\neq i}^{N} (y_i - \bar{Y})(y_j - \bar{Y})\right]$$

$$=\frac{1}{nN}\left[\sum_{i=1}^{N}(y_i-\bar{Y})^2\right]+\frac{(n-1)}{nN(N-1)}\left[\sum_{i=1}^{N}\sum_{j\neq i}^{N}(y_i-\bar{Y})(y_j-\bar{Y})\right]$$

$$=\frac{1}{nN}\left[\sum_{i=1}^{N}(y_i-\bar{Y})^2+\frac{(n-1)}{nN(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(y_i-\bar{Y})(y_j-\bar{Y})\right]$$

$$=\frac{1}{nN}\left[\sum_{i=1}^{N}(y_i-\bar{Y})^2-\frac{(n-1)}{(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})^2+\frac{(n-1)}{(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})^2+\frac{(n-1)}{(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(y_i-\bar{Y})(y_j-\bar{Y})\right]$$

$$=\frac{1}{nN}\left[\left(1-\frac{n-1}{N-1}\right)\sum_{i=1}^{N}(y_i-\bar{Y})^2+\frac{n-1}{N-1}\left\{\sum_{i=1}^{N}(y_i-\bar{Y})^2+\sum_{i=1}^{N}\sum_{j\neq 1}^{N}(y_i-\bar{Y})(y_j-\bar{Y})\right\}\right]$$

$$=\frac{1}{nN}\left[\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{N}(y_i-\bar{Y})^2+\frac{n-1}{N-1}\left\{\sum_{i=1}^{N}(y_i-\bar{Y})\right\}^2\right]$$

$$=\frac{1}{nN}\left[\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{N}(y_i-\bar{Y})^2+\frac{n-1}{N-1}\left\{\sum_{i=1}^{N}y_i-\sum_{i=1}^{N}\bar{Y}\right\}^2\right]$$

$$=\frac{1}{nN}\left[\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{N}(y_i-\bar{Y})^2+\frac{n-1}{N-1}\{N\bar{Y}-N\bar{Y}\}^2\right]$$

$$=\frac{N-n}{nN}\times\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2$$

$$=\left(\frac{N-n}{nN}\right)S^2=\frac{S^2}{n}\left(\frac{N-n}{N}\right)$$

$$=\frac{S^2}{n}\left(\frac{N}{N}-\frac{n}{N}\right)=\frac{S^2}{n}(1-f)$$

33. In simple random sampling, prove that s^2 is an unbiased estimate of S^2.

Ans. In simple random sampling

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \text{ and}$$

$$S^2=\frac{1}{N-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

To prove that $E(s^2) = S^2$

$$E(S^2)=E\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

$$=\frac{1}{n-1}E\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

$$=\frac{1}{n-1}E\left[\sum_{i=1}^{n}(y_i-\bar{Y}+\bar{Y}-\bar{y})^2\right]$$

$$=\frac{1}{n-1}E\left[\sum_{i=1}^{n}\left[(y_i-\bar{Y})-(\bar{y}-\bar{Y})\right]^2\right]$$

$$=\frac{1}{n-1}\left[E\left\{\sum_{i=1}^{n}(y_i-\bar{Y})^2\right\}+E\left\{\sum_{i=1}^{n}(y_i-\bar{Y})^2\right\}-2E\left\{\sum_{i=1}^{n}(y_i-\bar{Y})-(\bar{y}-\bar{Y})\right\}\right]$$

$$=\frac{1}{n-1}\left[E\left\{\sum_{i=1}^{N}\alpha_i(Y_i-\bar{Y})^2\right\}+nE(\bar{y}-\bar{Y})^2-2E\left\{(y-\bar{Y})\sum_{i=1}^{n}(\bar{y}_i-\bar{Y})\right\}\right]$$

$$=\frac{1}{n-1}\left[E\left\{\sum_{i=1}^{N}\alpha_i(Y_i--\bar{Y})^2\right\}+nE(\bar{y}-\bar{Y})^2-2E\left\{(y-\bar{Y})n(y-\bar{Y})\right\}\right]$$

$$=\frac{1}{n-1}\left[E\left\{\sum_{i=1}^{N}\alpha_i(Y_i-\bar{Y})^2\right\}+nE(\bar{y}-\bar{Y})^2-2nE(\bar{y}-\bar{Y})^2\right]$$

$$=\frac{1}{n-1}\left[E\left\{\sum_{i=1}^{N}\alpha_i(Y_i-\bar{Y})^2\right\}-nE(Y_i-\bar{Y})^2\right]$$

Where $\alpha_i=\begin{cases}1 \text{ if } Y_i \text{ is included in the sample}\\ 0 \text{ if } Y_i \text{ is not included in the sample}\end{cases}$

and $E(\alpha_i)=1\times\frac{n}{N}+0\times\left(1-\frac{n}{N}\right)=\frac{n}{N}$

$$E(S^2)=\frac{1}{n-1}\left[\left\{\sum_{i=1}^{N}(Y_i-\bar{Y})^2E(\alpha_i)\right\}-nV(\bar{y})\right]$$

$$= \frac{1}{n-1}\left[\frac{n}{N}\left\{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2\right\} - n \times \frac{N-n}{Nn} S^2\right]$$

$$= \frac{1}{n-1}\left[\frac{n}{N}\left\{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2\right\} - \frac{N-n}{Nn} S^2\right]$$

$$= \frac{n}{N(n-1)} \times (N-1)S^2 - \frac{N-n}{N(n-1)} S^2$$

$$= S^2\left[\frac{n(N-n)}{N(n-1)} - \frac{N-n}{N(n-1)}\right]$$

$$= S^2\left[\frac{nN-n-n+n}{N(n-1)}\right] = S^2\left[\frac{nN-N}{N(n-1}\right]$$

$$= S^2\left[\frac{N(n-1)}{N(n-1)}\right] = S^2$$

Thus it is proved that s^2 is an unbiased estimate of S^2.

THEORITICAL DISTRIBUTIONS

34. Derive Poisson distribution as a limiting case of binomial distribution.

Ans. Poisson distribution is a limiting case of binomial distribution under the following conditions:

(i) $n \to \infty$ i.e the number of independent Bernoullian trials is large.

(ii) $p \to 0$ i.e the constant probability of success at each trial is very small.

(iii) $np = \lambda$ is a constant

Hence the probability generating function of a Poisson distribution will be:

$$P(X;\lambda) = \lim_{\substack{n\to\infty \\ p\to 0}} B(X;n,p) = \lim_{\substack{n\to\infty \\ p\to 0}} {}^nC_x p^x q^{n-x}$$

$$= \lim_{\substack{n\to\infty \\ p\to 0}} \frac{n!}{x!(n-x)!}\left(\frac{np}{n}\right)^x \left(1-\frac{np}{n}\right)^{n-x}$$

$$= \lim_{n\to\infty} \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!} \frac{\lambda^x}{n^x}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-x}$$

$$= \frac{\lambda^x}{x!}\lim_{n\to\infty}\left[\frac{n}{n}\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\right]$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x}$$

$$= \frac{\lambda^x}{x!}\lim_{n\to\infty}\left[\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\right]$$

$$e^{-\lambda} \times 1 = e^{-\lambda}\frac{\lambda^x}{x!}$$

Thus the probability generating function o Poisson distribution is given by:

$$P(X;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}(x = 0,1,2,\cdots)$$

35. Discuss the skew ness and kurtosis of Poiss distribution.

Ans. For Poisson distribution $\mu_2 = \lambda$

$\mu_3 = \lambda$ and $\mu_4 = 3\lambda^2 + \lambda$

So $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\lambda^2}{\lambda^3} = \frac{1}{\lambda}$ and $\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{\lambda}}$

which is positive because λ is always a positiv quantity. Thus Poisson distribution is always a positivel skewed distribution.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\lambda^2+\lambda}{\lambda^2} = 3 + \frac{1}{\lambda} \text{ and } \gamma_2 = \beta_2 - 3 = \frac{1}{\lambda}$$

This is also always positive. Hence Poisso distribution is always a Leptokurtic distribution.

36. Derive a relation between mean and varianc of Poisson distribution.

Ans. Mean of Poisson distribution is given by:

$$E(X) = \sum_{x=0}^{\infty} x \times P(X;\lambda) = \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda}\lambda\left[\frac{\lambda^0}{0!}+\frac{\lambda^1}{1!}+\frac{\lambda^2}{2!}+\cdots\right]$$

$$= e^{-\lambda}\lambda e^{\lambda} = \lambda$$

For Poisson distribution,

$$E(X^2) = \sum_{x=0}^{\infty} x^2\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\left[x(x-1)+x\right]\frac{\lambda^x}{x!}$$

$$= e^{-\lambda}\sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!} + e^{-\lambda}\sum_{x=0}^{\infty} x\frac{\lambda^x}{x!}$$

$$= e^{-\lambda}\lambda^2\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} + e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}$$

$$= e^{-\lambda}\lambda^2\left[\frac{\lambda^0}{0!}+\frac{\lambda^1}{1!}+\frac{\lambda^2}{2!}+\cdots\right] + e^{-\lambda}\lambda\left[\frac{\lambda^0}{0!}+\frac{\lambda^1}{1!}+\frac{\lambda^2}{2!}+\cdots\right]$$

$= e^{-\lambda}\lambda^2 e^{\lambda} + e^{-\lambda}\lambda e^{\lambda} = \lambda^2 + \lambda$

Hence the variance of Poison distribution is given by: $E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

Thus for Poisson distribution, Mean = Variance.

37. Write a note on Binomial distribution and mention its properties.

Ans. The probability generating function of a Binomial distribution with parameters n and p is given by $B(x; n, p) = {}^nC_x p^x q^{n-x}$ where $p + q = 1$ and $x = 0, 1, 2, \ldots, n$. Here n = number of independent Bernoullian trials, p = probability of success at each trial. q = probability of failure at each trial = 1–p. The probability generating function gives the probability of getting x successes and n-x failures in n independent Bernoullian trials with constant probability of success at each trial p and constant probability of failure at each trial q.

The properties of Binomial distribution are as follows:

(i) The sum of all probabilities of a binomial distribution is 1.

(ii) Binomial distribution has two parameters n and p.

(iii) Mean of Binomial distribution is np.

(iv) Variance of Binomial distribution is npq and the standard deviation of Binomial distribution is $\sqrt{npq}$.

(v) For a Binomial distribution $\mu_3 = npq(q - p)$,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq}, \quad \gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}}$$

So if $p = q = ½$, the distribution is symmetrical. If $p < ½$, the distribution is positively skewed and if $p > ½$, the distribution is negatively skewed.

(vi) For a Binomial distribution

$\mu_4 = npq\,[1 + 3(n-2)\,pq]$,

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1-6pq}{npq}, \quad \gamma_2 = \beta_2 - 3 = \frac{1-6pq}{npq}$$

So if $pq = \frac{1}{6}$, the distribution is mesokurtic.

If $pq < \frac{1}{6}$, the distribution is leptokurtic and

if $pq > \frac{1}{6}$, the distribution is platykurtic.

(vii) If $(n+1)p$ is an integer the distribution is bimodal having modes at $(n+1)p$ and $(n+1)p - 1$. If $(n+1)p$ is not an integer, the distribution is unimodal having mode at the integral part of $(n+1)p$.

38. Derive the standard deviation of binomial distribution.

Ans. Variance of Binomial distribution is given by: $V(X) = E(X^2) - [E(X)]^2$

$$E(X) = \sum_{x=0}^{n} x.{}^nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x.\frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!(n-x)!} p^x q^{n-x}$$

$$= np\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= np(q+p)^{n-1} = np \ (\because p+q=1)$$

$$E(X^2) = \sum_{x=0}^{n} x^2.{}^nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1)+x].\frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1)\frac{n!}{x!(n-x)!} p^x q^{n-x} + \sum_{x=0}^{n} x\frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} q^{n-x} + \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= n(n-1)p^2(q+p)^{n-2} + np(q+p)^{n-1}$$

$$= n^2p^2 - np^2 + np \ (\because p+q=1)$$

Hence $V(X) = E(X^2) - [E(X)]^2$

$$= n^2p^2 - np^2 + np - n^2p^2 = np(1-p) = npq$$

So standard deviation of Binomial distribution

$$= \sqrt{V(X)} = \sqrt{npq}$$

39. Derive the mean of binomial distribution.

Ans. The mean of binomial distribution is given by:

$$E(X) = \sum_{x=0}^{n} x.{}^nC_x p^x q^{n-x} = \sum_{x=0}^{n} x.\frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!(n-x)!} p^x q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= np(q+p)^{n-1} = np \ (\because p+q=1)$$

Thus mean of Binomial distribution is np.

40. Derive the relation between mean and variance of binomial distribution.

Ans. The mean of binomial distribution is given by:

$$E(X) = \sum_{x=0}^{n} x.{}^nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x. \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!(n-x)!} p^x q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= np(q+p)^{n-1} = np \ (\because p+q=1)$$

Thus mean of Binomial distribution is np.

Variance of Binomial distribution is given by:

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = \sum_{x=0}^{n} x^2.{}^nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1)+x]. \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$+ \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} q^{n-x}$$

$$+ \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= n(n-1)p^2(q+p)^{n-2} + np(q+p)^{n-1}$$

$$= n^2p^2 - np^2 + np \ (\because p+q=1)$$

Hence $V(X) = E(X^2) - [E(X)]^2$

$$= n^2p^2 - np^2 + np - n^2p^2 = np(1-p) = npq$$

Thus $\frac{Variance}{Mean} = \frac{npq}{np} = q < 1$, Hence for a binomial distribution, Mean > Variance.

41. Discuss the skew ness and kurtosis of binomial distribution.

Ans. For a Binomial distribution, $\mu_2 = npq$

$$\mu_3 = npq(q-p), \ \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq},$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}}$$

So if $p = q = ½$, the distribution is symmetrical. If $p < ½$, the distribution is positively skewed and if $p > ½$, the distribution is negatively skewed.

For a Binomial distribution

$$\mu_4 = npq\,[1 + 3(n-2)\,pq],$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1-6pq}{npq},$$

$$\gamma_2 = \beta_2 - 3 = \frac{1-6pq}{npq}$$

So if $pq = \frac{1}{6}$, the distribution is mesokurtic. If $pq < \frac{1}{6}$, the distribution is leptokurtic and if $pq > \frac{1}{6}$, the distribution is platykurtic.

42. Derive the mode of binomial distribution.

Ans. Mode of a probability distribution is that value of the random variable which has the maximum probability. Let the mode of a Binomial distribution with parameter n and p be r. Then r is a positive integer between 0 and n such that $P(X = r)$ is the maximum.

Hence $P(X = r) \geq P(X = r-1) \cdots (i)$

and $P(X = r) \geq P(X = r+1) \cdots (ii)$

From (i):

$$P(X=r) \geq P(X=r-1) \Rightarrow \frac{P(X=r)}{P(X=r-1)} \geq 1$$

$$\Rightarrow \frac{{}^nC_r p^r q^{n-r}}{{}^nC_{r-1} p^{r-1} q^{n-r+1}} \geq 1 \Rightarrow \frac{\frac{n!}{r!(n-r)!}}{\frac{n!}{(r-1)!(n-r+1)!}} \times \frac{p}{q} \geq 1$$

$$\Rightarrow \frac{n-r+1}{r} \times \frac{p}{q} \geq 1 \Rightarrow np - rp + p \geq r(1-p)$$

$$\Rightarrow np - rp + p \geq r - rp \Rightarrow (n+1)p \geq r$$

$$\Rightarrow r \leq (n+1)p \cdots (iii)$$

From (ii):

$$P(X=r) \geq P(X=r+1) \Rightarrow \frac{P(X=r)}{P(X=r+1)} \geq 1$$

$$\Rightarrow \frac{{}^nC_r p^r q^{n-r}}{{}^nC_{r+1} p^{r+1} q^{n-r-1}} \times \frac{q}{p} \geq 1$$

$$\Rightarrow \frac{\frac{n!}{r!(n-r)!}}{\frac{n!}{(r+1)!(n-r-1)!}} \times \frac{q}{p} \geq 1$$

$$\Rightarrow \frac{r+1}{n-r} \times \frac{q}{p} \geq 1 \Rightarrow (r+1)(1-p) \geq (n-r)p$$

$$\Rightarrow r+1-rp-p \geq np-rp \Rightarrow r \geq (n+1)p-1$$

$$\Rightarrow (n+1)p-1 \leq r \cdots (iv)$$

Combining (iii) and (iv) it is obtained that the mode of binomial distribution is that integer r which satisfies the condition: $(n+1)p-1 \leq r \leq (n+1)p$

Thus if $(n+1)p$ is an integer the distribution is bimodal having modes at $(n+1)p$ and $(n+1)p-1$.

But if $(n+1)p$ is not an integer, the distribution is unimodal having mode at the integral part of $(n+1)p$.

43. Prove that binomial distribution is a discrete probability distribution.

Ans. The random variable used in binomial distribution is number of successes which can take only positive integral values. So binomial distribution is a discrete distribution.

The probability generating function of a Binomial distribution with parameters n and p is given by $B(x; n, p) = {}^nC_x p^x q^{n-x}$ where $p+q=1$ and $x = 0, 1, 2, \ldots, n$.

Sum of all binomial probabilities

$$= \sum_{x=0}^{n} {}^nC_x p^x q^{n-x}$$

$$= {}^nC_0 p^0 q^{n-0} + {}^nC_1 p^1 q^{n-1} + \cdots + {}^nC_n p^n q^{n-n}$$

$$= (q+p)^n = 1$$

So binomial distribution is a probability distribution.

Hence binomial distribution is a discrete probability distribution.

44. Describe the important properties of a normal distribution.

Ans. The properties of normal distribution are:

(i) Its probability density function is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (-\infty < x < \infty)$$

(ii) Its probability distribution function is given by:

$$F(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (-\infty < t < \infty)$$

(iii) The mean of normal distribution is μ.

(iv) Its standard deviation is σ.

(v) All odd ordered moments of normal distribution are Zeros.

(vi) The even ordered moments of normal distribution are given by the formula:

$$\mu_{2r} = \frac{(2r)!}{2^r r!} \sigma^{2r}$$

So $\mu_2 = \sigma^2$ *and* $\mu_4 = 3\sigma^4$

(vii) Since $\mu_3 = 0, \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$ and $\gamma_1 = \sqrt{\beta_1} = 0$.

Thus Normal distribution is a symmetrical distribution.

(viii) For normal distribution, $\mu_2 = \sigma^2$ and

$$\mu_4 = 3\sigma^4, \text{so} \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 \text{ and } \gamma_2 = \beta_2 - 3 = 0$$

So Normal distribution is a mesokurtic distribution.

(ix) The mean deviation of normal distribution is

$\frac{4}{5}\sigma$ (approximately)

(x) The quartile deviation of normal distribution is $\frac{2}{3}\sigma$ (approximately)

(xi) The two points of inflexion of normal distribution are $\mu \pm \sigma$.

(xii) **Area property:**

$$P(\mu-\sigma \leq X \leq \mu+\sigma) = P(-1 \leq Z \leq 1) = 0.6826$$

$$P(\mu-2\sigma \leq X \leq \mu+2\sigma) = P(-2 \leq Z \leq 2) = 0.9544$$

$$P(\mu-3\sigma \leq X \leq \mu+3\sigma) = P(-3 \leq Z \leq 3) = 0.9973$$

$$P(\mu-1.96\sigma \leq X \leq \mu+1.96\sigma)$$

$$= P(-1.96 \leq Z \leq 1.96) = 0.95$$

$$P(\mu-2.58\sigma \leq X \leq \mu+2.58\sigma)$$

$$= P(-2.58 \leq Z \leq 2.58) = 0.99$$