Statistics

Unit - 1

1) Dispersion of a distribution is the amount of Scatteredness the Individual values from a measure of central tendency.

2) Range is the most simple and obvious measure of dispersion.

3) The quartile deviation (or) semi inter quartile range is defined by $Q.D = \frac{1}{2}(Q_3 - Q_1)$

4) The mean deviation of a frequency distribution from any average A is defined by $M.D = \frac{\Sigma f_i |x_i - A|}{N}$ where $N = \Sigma f_i$

5) The Standard deviation σ of a frequency distribution is defined by $\sigma = \left[\frac{\Sigma f_i (x_i - \bar{x})^2}{N}\right]^{1/2}$

6) The Square of the Standard deviation of a frequency distribution is called the Varience of the frequency distribution.

7) Varience $= \sigma^2$.

8) The root mean Square deviation of a frequency distribution is defined to be $S = \left[\frac{\Sigma f_i (x_i - A)^2}{N}\right]^{1/2}$

9) A is any arbitrary origin and S^2 is called the mean Square deviation.

10) Co-efficient of Variation of a frequency distribution is defined to be $C.V = \frac{\sigma}{\bar{x}} \times 100$.

11) The Standard deviation is the least possible root mean Square deviation

12) The Standard deviation σ is independent of change of origin and is dependent on change of Scale.

13) When we effect a change in origin as well as in Scale σ^2 is multiplied by the Square of the Scale introduced.

14) Variance of combined set, $\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1+n_2} + \frac{n_1 n_2}{(n_1+n_2)^2}(\bar{x}_1 - \bar{x}_2)^2$.

15) The r^{th} moment about any point A, denoted by μ_r' of a frequency distribution (f_i/x_i) is defined by $\mu_r' = \frac{\Sigma f_i (x_i - A)^r}{N}$.

16) The r^{th} moment about the arithmetic mean $\bar{x}$ of a frequency distribution is given by $\mu_r = \frac{\Sigma f_i (x_i - \bar{x})^r}{N}$

17) μ_r is also called the r^{th} central moment.

18) $\mu_1 = \frac{\Sigma f_i (x_i - \bar{x})}{N} = 0$.

19) Karl pearson's β and γ coefficients are defined as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad ; \quad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$\gamma_1 = \sqrt{\beta_1} \quad ; \gamma_2 = \beta_2 - 3.$$

20) Thus Skewness means lack of Symmetry.

21) β_1 can be taken as a measure of Skewness.

22) The frequency distribution has positive Skewness if $\beta_1 > 0$ and negative Skewness if $\beta_1 < 0$.

23) Mean - Mode and Mean - Median may be taken as measures of Skewness.

24) These measures were suggested by Karl pearson.

25) Another measure of Skewness due to Bowley is based on the fact that for a positively Skewed distribution on the third quartile is farther from the median that the first quartile so that $Q_3 - \text{Median} > \text{Median} - Q_1$.

26) The measures of Skewness are the absolute measures of Skewness.

27) $\dfrac{\text{Mean} - \text{Mode}}{\sigma}$ and $\dfrac{3(\text{Mean} - \text{Median})}{\sigma}$ are called the karl pearson's co-efficients of Skewness.

28) Bowley's co-efficient of Skewness is given by $\dfrac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$

29) Kurtosis is the degree of peakedness of a distribution usually taken relative to a normal distribution.

30) Thus Kurtosis enable us to have an idea about the flatness or peakedness of a frequency curve. It is measured by the co-efficient β_2.

31) For a normal curve $\beta_2 = 3$ (or) $(\gamma_2 = 0)$ Messokurtic.

32) For a curve which is flater than the normal curve $\beta_2 < 3$ (or) $(\gamma_2 < 0)$ and Such a curve is known as platy kurtic.

33) For a curve which is more peaked than the normal curve $\beta_2 > 3$ (or) ($\gamma_2 > 0$) and such a curve is known as leptokurtic.

④

Unit-2

1) (x_i, y_i); $i = 1, 2, 3 \dots n$ is called a bivariate data.

2) There are two main problems involved in the relationship between x and y.

3) The first is to find a measure of the degree of association or correlation between the values of and those of

4) The second problem in to find the most suitable form of an for determining the probable value of one variable corresponding to a given value of the other. This is the problem of regression.

5) Consider a set of bivariate data (x_i, y_i); $i = 1, 2, \dots n$. If there is a change in one variable corresponding to a change in the other variable we say that the variables are correlated.

6) If the two variables deviate in the same direction the correlation is said to be direct (or) positive.

7) If they always deviate in the opposite direction the correlation is said to be inverse (or) negative.

8) If the change in one variable corresponds to a proportainal change in the other variable then the correlation is said to be perfect.

9) Karl pearson's coefficient of correlation between the variables x and y is defined by $r_{xy} = \dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x \sigma_y}$

10) The covariance between x and y is defined by

$$\text{cov}(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n} \quad \text{Hence,} \quad r_{xy} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$$

11) If $r=1$ the correlation is perfect and positive.

12) If $r=-1$ the correlation is perfect and negative.

13) If $r=0$ the variables are uncorrelated.

14) If the variables x and y are uncorrelated then $\text{cov}(x,y)=0$.

15) The co-efficient of correlation between the ranks x_i and y_i is called the rank correlation coefficients and is denoted by ρ.

16) $\rho = 1 - \dfrac{6\,\Sigma(x-y)^2}{n(n^2-1)}$.

17) This is known as Spearman's formula for rank correlation coefficient.

18) If there is a functional relationship between the two variables x_i and y_i the points in the scatter diagram will cluster around some curve called the curve of regression.

19) If the curve is a straight line it is called a line of regression between the two variables.

20) If we fit a straight line by the principle of least squares to the points of the scatter diagram in such a way that the sum of the squares of the distance parallel to they y-axis from the points to the line is minimized we obtain a line of best fit for the data and it is called the regression line of y on x.

21) Similarly, we can define the regression line of x on y.

22) The equation of the regression line of y on x is given by $y - \bar{y} = \gamma \frac{\sigma_y}{\sigma_x}(x - \bar{x})$.

23) The equation of the regression line of x on y is

$$x - \bar{x} = \gamma \frac{\sigma_x}{\sigma_y}(y - \bar{y}).$$

24) $(\bar{x}, \bar{y})$ is the point of intersection of the two regression lines.

25) The Slope of the regression line of y on x is called the regression co-efficient of y on x and it is denoted by b_{yx}. Hence $b_{yx} = \gamma \frac{\sigma_y}{\sigma_x}$.

26) The regression co-efficient of x on y is given by $b_{xy} = \gamma \frac{\sigma_x}{\sigma_y}$.

27) correlation co-efficient is the geometric mean between the regression co-efficients (ie) $\gamma = \pm\sqrt{b_{xy} \cdot b_{yx}}$.

28) If one of the regression coefficients is greater than unity the other one is less the unity.

29) Arithmetic mean of the regression coefficient is greater than or equal to the correlation coefficient.

30) The obtuse angle between the regression lines is given by,

$$\tan^{-1}\left[\left(\frac{\gamma^2-1}{\gamma}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right]$$

31) If $\gamma = 0$ then $\tan\theta = \infty$. Hence $\theta = \pi/2$. Thus if the two variables are uncorrelated then the lines of regression are perpendicular to each other.

32) If $r = \pm 1$ then $\tan\theta = 0$. Hence $\theta = 0$ (or) π. The two lines of regressions are parallel.

33) For any fixed i we have $\sum_{j=1}^{m} f_{ij} = g_i$ = the Sum of all the cell frequencies of the i^{th} column.

34) For any fixed j we have $\sum_{i=1}^{n} f_{ij} = f_j$ = the sum of all the frequencies of the j^{th} row.

35) If the total frequency of all the mn cells is N then,

$$N = \sum_{i=1}^{n} g_i = \sum_{j=1}^{m} f_j \quad \text{and} \quad N = \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} = \sum_{j=1}^{m} \sum_{i=1}^{n} f_{ij}.$$

36) The correlation coefficient between x and y is given by,

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