# Statistics

## Measures of location

### Measures of central tendency

An average or a central value of a statistical series in the value of the variable which describes the characteristics of the entire distribution.

The following are the five measure of central tendency:

(1) Arithmetic mean (2) Geometric mean

(3) Harmonic mean (4) Median (5) Mode

(1) **Arithmetic mean :** If $x_1, x_2, ....., x_n$ are *n* value of a variate *x*, then arithmetic mean (*A.M.)* is given by

$$A.M. = \frac{1}{n}[x_1 + x_2 + x_3 + ..... + x_n] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mean is the most suitable method of measures of central tendency.

**Short cut method :** If *A* be an assumed mean, $d = x - A$, the deviation of each value of the variable from the assumed mean, then, $A.M. = A + \frac{1}{n}\sum d$ or $A.M. = A + \frac{\sum f_i d_i}{n}$, where $d_i = x_i - A$

The direct method is as $A.M. = \frac{\sum fx}{\sum f} = \frac{1}{N}\sum fx$

Where *f* is the frequency of the variable *x*.

(2) **Geometric mean :** If $x_1, x_2, x_3, ....., x_n$ are *n* values of a variate *x,* none of them being zero ,then geometric mean (G.M.) is given by $G.M. = (x_1.x_2.x_3 ..... x_n)^{1/n} \Rightarrow \log(G.M.) = \frac{1}{n}(\log x_1 + \log x_2 + ..... + \log x_n)$

In case of frequency distribution, G.M. of *n* values $x_1, x_2, ....., x_n$ of a variate *x* occurring with frequency $f_1, f_2, ....., f_n$ is given by $G.M. = \left(x_1^{f_1}.x_2^{f_2} .....x_n^{f_n}\right)^{1/N}$, where $N = f_1 + f_2 + ..... + f_n$ .

(3) **Harmonic mean :** The harmonic mean of $n$ items $x_1, x_2, ....., x_n$ is defined as $H.M. = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+.....+\frac{1}{x_n}}$

If the frequency distribution is $f_1, f_2, f_3, ....., f_n$ respectively then $H.M. = \frac{f_1+f_2+f_3+.....+f_n}{\left(\frac{f_1}{x_1}+\frac{f_2}{x_2}+.....+\frac{f_n}{x_n}\right)}$

(4) **Median :** Median is that value which divides the entire set of data into two equal parts.

(i) **Individual series :** First arrange the data in ascending or descending order. Then the median is

(a) $\frac{1}{2}(n+1)th$ item, if $n$ is odd. (b) Mean of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ item, if $n$ is even.

**Continuous frequency distribution:** Median = $l+\left[\frac{\frac{N}{2}-F}{f}\right]h$

Where $l$ = lower limit of the median class.

$f$ = Frequency of the median class.

$h$ = Size of the median class.

$F$ = Cumulative frequency of the class preceding the median class

$N = \sum_{i=1}^{n} f_i$

(5) **Mode :** The mode or model value of a distribution is that value of the variable for which the frequency is maximum. For continuous series, mode in calculated as, $\text{Mode} = l_1+\left[\frac{f_1-f_0}{2f_1-f_0-f_2}\right]\times i$

Where, $l_1$ = The lower limit of the model class.

$f_1$ = The frequency of the model class.

$f_0$ = The frequency of the class preceding the model class.

$f_2$ = The frequency of the class succeeding the model class.

$i$ = The size of the model class.

**Note** : ❑ Mean – mode = 3 (Mean – median)

$\Rightarrow$ Mode = 3 median – 2 mean.

## Dispersion

The degree to which numerical data tend to spread about an average value is called the dispersion of the data.

The four measure of dispersion are :

(1) **Range :** It is the difference between the values of extreme items in a series. $\text{Range} = X_{\max} - X_{\min}$

The coefficient of range (scatter) = $\frac{x_{\max.} - x_{\min.}}{x_{\max.} + x_{\min.}}$.

Range is not the measure of central tendency. Range is widely used in statistical series relating to quality control in production.

(2) **Quartile deviation or semi inter quartile range :** It is one – half of the difference between the third quartile and first quartile *i.e.,* $Q.D. = \frac{Q_3 - Q_1}{2}$ and coefficient of quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$.

Where, $Q_3$ is the third or upper quartile and $Q_1$ is the lower or first quartile.

(3) **Mean deviation :** It is the average of the modulus of the deviation of the observation in a series taken from mean or median or mode.

(i) **For ungrouped data :** In this case the mean deviation is given by the formula.

$$M.D = \frac{\Sigma|x - A|}{n} = \frac{\Sigma|d|}{n}$$

(ii) **Mean deviation for grouped data –** $M.D = \frac{\Sigma f|x - M|}{\Sigma f} = \frac{\Sigma f|d|}{n}$, (where $\Sigma f = n$)

(4) **Standard deviation :** The positive square root of the average of squared deviation of all observations taken from their mean is called standard deviation. It is denoted by $\sigma$ and $\sigma = \sqrt{\frac{\Sigma(x - M)^2}{n}}$.

For the frequency distribution, $\sigma = \sqrt{\frac{\Sigma f(x - M)^2}{\Sigma f}}$

Where, $x$ is variable, $M$ is arithmetic mean, and $n$ is total number frequency.

**Short Cut Method**

(1) $\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{\Sigma x^2}{n} - M^2}$ (2) $\sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left[\frac{\Sigma f(x)}{\Sigma f}\right]^2} = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - M^2}$

## Variance

The square of standard deviation in called the variance.

**Coefficient of standard deviation and variance** – The coefficient of standard deviation is the ratio of the *S.D.* to *A.M. i.e.*, $\frac{\sigma}{\bar{x}}$. Coefficient of variance = coefficient of $S.D \times 100 = \frac{\sigma}{\bar{x}} \times 100$ .

**Variance of the combined series :** If $n_1$; $n_2$ are the sizes, $\bar{x}_1; \bar{x}_2$ the means and $\sigma_1; \sigma_2$ the standard deviation of two series, then $\sigma^2 = \frac{1}{n_1 + n_2}\left[n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)\right]$

where, $d_1 = \bar{x}_1 - \bar{x}, d_2 = \bar{x}_2 - \bar{x},$ and $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

## Skewness.

The distribution is skewed if,

(i) Mean ≠ median ≠ mode

(ii) Quartiles are not equidistant from the median and

(iii) The frequency curve is streched more to one side than to the other.

A distribution is positively skewed if the value of mean is maximum and that of mode is least – the median in between the two. In a negatively skewed distribution the value of mode is maximum and that of mean is least – the median lies in between the two.

(1) **Measures of Skewness :** (i) **Absolute measures of skewness** : Various measures of skewness are

(a) $S_K = M - M_d$ (b) $S_k = M - M_o$ (c) $S_k = Q_3 + Q_1 - 2M_d$

Where, $M_d$ = median , $M_o$ = mode, $M$ = mean

Absolute measures of skewness are not useful to compare two series, therefore relative measure of dispersion are used, as their pure numbers.

(2) **Relative measures of Skewness :**

(i) **The Karl pearson's coefficient of skewness :** $S_k = \frac{M - M_o}{\sigma} = 3\frac{(M - M_d)}{\sigma}, -3 \le S_k \le 3$

(ii) **Bowley's coefficient of skewness :** $S_k = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$

Bowley's coefficient of skewness lies between –1 and 1.

(iii) **Kelly's coefficient of skewness :** $S_k = \frac{P_{10} + P_{90} - 2M_d}{P_{90} - P_{10}} = \frac{D_1 + D_9 - 2M_d}{D_9 - D_1}$

## Some important results.

(1) For a symmetrical distribution, the following area relationship holds good

$\overline{X} \pm \sigma$ covers 68.27% items

$\overline{X} \pm 2\sigma$ covers 95.45% items

$\overline{X} \pm 3\sigma$ covers 99.74% items

(2) **Relationship between different measures of dispersion.**

(i) $Q.D = \frac{2}{3}\sigma$ (Approx.)

(ii) $M.D = \frac{4}{5}\sigma$ (Approx.)

(iii) $S.D = \frac{3}{2}Q.D.$ (Approx.)

(iv) $Q.D. = \frac{5}{6}M.D.$ (Approx.)

(v) $M.D. = \frac{6}{5}Q.D.$ (Approx).

# Correlation and Regression

## Some definitions.

(1) **Univariate distribution :** These are the distributions in which there is only one variable such as the heights of the students of a class.

(2) **Bivariate distirbution :** Distribution involving two discrete variable is called a bivariate distribution. For example, the heights and the weights of the students of a class in a school.

(3) **Bivariate frequency distribution :** Let *x* and *y* be two suppose *x* takes the values $x_1, x_2, ....., x_n$ and *y* takes the values $y_1, y_2, ....., y_n$, then we record our observations in the form of ordered pairs $(x_1, y_1)$, where $1 \le i \le n, 1 \le j \le n$ If a certain pair occurs $f_{ij}$ times, we say that its frequency is $f_{ij}$.

The function which assigns the frequencies $f_{ij}'s$ to the pairs $(x_i, y_j)$ is known as a bivariate frequency distribution.

## Covariance.

Let $(x_i, x_i); i = 1,2, ....., n$ be a bivariate distribution , where $x_1, x_2, ....., x_n$ are the values of variable *X* and $y_1, y_2 ..... y_n$ those of *Y*. Then the covariance Cov (*X,Y*) between *X* and *Y* is given by

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) \quad \text{or} \quad \mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\ \overline{Y}$$

where, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are means of variables $X$ and $Y$ respectively.

Covariance is not affected by the change of origin, but it is affected by the change of scale.

## Correlation

The relationship between two variables such that a change in one variable results in a positive or negative change in the other variable is known as correlation.

(1) **Type of correlation :** (i) **Perfect correlation :** If the two variable vary in such a manner that their ratio is always constant, then the correlation is said to be perfect.

(ii) **Positive or direct correlation :** If an increase or decrease in one variable corresponds to an increase or decrease in the other, the correlation is said to be positive.

(iii) **Negative or indirect correlation :** If an increase or decrease in one variable corresponds to a decrease or increase in the other, the correlation is said to be negative.

(2) **Karl Pearson's coefficient of correlation :** The correlation coefficient $r(X,Y)$, between two variable $X$ and $Y$ is given by, $r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ or $\frac{Cov(X,Y)}{\sigma_x \sigma_y}$, $r(X,Y) = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i\right)^2}}$

$$r = \frac{\Sigma(x-\overline{x})(y-\overline{y})}{\sqrt{\Sigma(x-\overline{x})^2}\sqrt{\Sigma(y-\overline{y})^2}} = \frac{\Sigma dxdy}{\sqrt{\Sigma dx^2}\sqrt{\Sigma dy^2}}$$

(3) **Modified Formula :** $r = \frac{\sum dxdy - \frac{\sum dx . \sum dy}{n}}{\sqrt{\left\{\sum dx^2 - \frac{\left(\sum dx\right)^2}{n}\right\}\left\{\sum dy^2 - \frac{\left(\sum dy\right)^2}{n}\right\}}}$ ,where $dx = x - \overline{x}; dy = y - \overline{y}$

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

(4) **Rank Correlation :** $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$

Where $\sum d^2$ = sum of the squares of the difference of two ranks and $n$ is the number of pairs of observations.

(5) **Important points about Correlation Coefficient ($r$)**

(i) $r$ lies between – 1 and + 1

(ii) The correlation is

(a) Perfect and positive if $r = +1$ (b) Perfect and negative if $r = -1$

(c) Not correlated if $r = 0$ (d) Positive if $r > 0$

(e) Negative if $r < 0$

(iii) It is independent of the change of origin and scale.

(iv) It is a pure number and hence unitless.

(v) If $x$ and $y$ are independent, then $r = 0$.

## Regression analysis

In a statistical relationship, if the value of one variable is known, we can estimate the value of the another variable through a procedure known as regression analysis.

(1) **Regression lines :**

(i) **Regression Line of $y$ on $x$:** If value of $x$ is known, then value of $y$ can be found as

$$y - \bar{y} = \frac{Cov(x,y)}{\sigma_x^2}(x - \bar{x}) \text{ or } y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

(ii) **Regression Line of $x$ on $y$:** It estimates $x$ for the given value of $y$ as

$$x - \bar{x} = \frac{Cov(x,y)}{\sigma_y^2}(y - \bar{y}) \text{ or } x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

(2) **Regression Coefficient :** (i) of $y$ on $x$ is $b_{yx} = \frac{r\sigma_y}{\sigma_x} = \frac{\text{cov}(x,y)}{\sigma_x^2}$ (ii) of $x$ on $y$ is $b_{xy} = \frac{r\sigma_x}{\sigma_y} = \frac{Cov(x,y)}{\sigma_y^2}$

(3) **Important points about Regression Coefficients $b_{xy}$ and $b_{yx}$ :**

(i) $r = \sqrt{b_{yx}.b_{xy}}$ *i.e.* the coefficient of correlation is the geometric mean of the coefficient of regression.

(ii) If $b_{yx} > 1$, then $b_{xy} < 1$ *i.e.* if one of the regression coefficient is greater than unity, the other will be less than unity.

(iii) If the correlation between the variable is not perfect, then the regression lines intersect at $(\bar{x}, \bar{y})$.

(iv) $b_{yx}$ is called the slope of regression line *y* on *x* and $\frac{1}{b_{xy}}$ is called the slope of regression line *x* on *y*.

(v) $b_{yx} + b_{xy} > 2\sqrt{b_{yx} b_{xy}}$ or $b_{yx} + b_{xy} > 2r$ *i.e.* the arithmetic mean of the regression coefficient is greater than the correlation coefficient.

(vi) Regression coefficients are independent of change of origin but not of scale.

(vii) The product of lines of regression's gradients is given by $\frac{\sigma_y^2}{\sigma_x^2}$.

(viii) If the angle between lines of regression is $\theta$, then $\tan\theta = \left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$.

(ix) If both the lines of regression coincide, then correlation will be perfected linear.

(x) If both $b_{yx}$ and $b_{xy}$ are positive, the $r$ will be positive and if both $b_{yx}$ and $b_{xy}$ are negative, the *r* will be negative.

(4) **Important points on Regression Lines :**

(i) If $r = 0$, then tan$\theta$ is not defined *i.e.* $\theta = \frac{\pi}{2}$. Thus the regression lines are perpendicular

(ii) If $r = +1$ or $-1$, then tan $\theta$ = 0 *i.e.* $\theta$ = 0. Thus the regression lines are coincident.

(iii) If regression lines are $y = ax + b$ and $x = cy + d$, then $\bar{x} = \frac{bc + d}{1 - ac}$ and $\bar{y} = \frac{ad + b}{1 - ac}$

## Standard error and probable error.

(1) **Standard error of Prediction :** The deviation of the predicted value form the observed value is known as the standard error prediction and is defined as $S_y = \sqrt{\left\{\frac{\sum (y - y_p)^2}{n}\right\}}$

Where *y* is actual value of $y_p$ is predicted value.

In relation to coefficient of correlation, it is given by

(i) Standard error of estimate of *x* is $S_x = \sigma_x \sqrt{1 - r^2}$ (ii) Standard error of estimate of *y* is $S_y = \sigma_y \sqrt{1 - r^2}$.

(2) **Relation between probable error and standard error :** If *r* is the correlation coefficient in a sample of *n* pairs of observations, then its standard error S.E.$(r) = \frac{1-r^2}{\sqrt{n}}$

and probable error P.E. $(r)$ = 0.6745 (S.E.) = 0.6745 $\left(\frac{1-r^2}{\sqrt{n}}\right)$. The probable error or the standard error are used for interpreting the coefficient of correlation.

(i) If $r < P.E.(r)$, there is no evidence of correlation.

(iii) If $r > 6P.E.(r)$, the existence of correlation is certain.

The square of the coefficient of correlation for a bivariate distribution is known as the "coefficient of determination."

## Some important points.

(1) If *byx*, *bxy* and $r \geq 0$ then $\frac{1}{2}(bxy + byx) \geq r$ and if *bxy*, *byx* and $r \leq 0$ then $\frac{1}{2}(bxy + byx) \leq r$.

(2) Correlation measures the relationship between variables while regression measures only the cause and effect of relationship between the variables.

(3) Standard deviation $\leq$ Range *i.e.* variance $\leq (\text{Range})^2$.