FY BCOM

Introduction of Statistics

The word 'Statistics' has been derived from the Latin word 'status', Italian word 'statistia', German word 'statistik', each of which mean a political state. It was first used by professor Achenwell in 1749 to refer to the subject-matter as a whole.
By statistics we mean numerical or quantitative data or information.

Statistics in the Plural Senses (as a subject) means the collection of the following four scientific steps.

1. Collection of Data: Under this stage the relevant data are collected from various sources, primary or secondary.

2. Presentation (Classification and Tabulation) of Data: The data collected to be understood should be presented in suitable form.

3. Analysis: The next step after presentation is that of analysis. There are various method used for analyzing the data.

4. Interpretation of Data

The valid conclusions on the basis of date analysed is drawn.

Limitation or Demerits of Statistics

1. Statistics does not deal with Individuals

Individual values of the observations has no specific importance. Statistics is the study of mass data and deals with aggregates of facts.

2. Statistics does not study Qualitative Data

Statistics is the study of only of those facts which are capable of being stated in number or quantity.

3. Statistics gives Result only on an Average

Statistics method are not exact. The result obtained are true only on an average in the long run.

4. The results can be biased: The data collection may sometime be biased which will make the whole investigation useless.

Importance or Uses of Statistics

It is impossible to think any sphere of human activities where statistics does not creep in statistics has covered all branches of science and commerce.

1. Business must be planned properly and planning to be fruitful must be based on the right analysis of complex statistical data.
2. In Economics, Statistics is used in GDP, Stock market, Demand, Supply etc. calculations.
3. It is used in Medicine to compare effect of Medicine.
4. It is used in Sports, Psychology, Sociology etc. fields also.

Function of Statistics

1. Statistics presents the facts in definite form.
2. Statistics simplifies complex data.
3. It provides a techniques of comparison.
4. Statistics studies the relationship
5. It help in formulating policies
6. It help in forecasting

Population:

It is the entire collection of observations (person, animal, plant or things which is actually studied by a researches) from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.

Example :

1. If we are studying the weight of adult women, the population is the set of weights of all women in the world.
2. If we are studying the grade point average of students of Mumbai University, the population is the set of GPA's of all students of Mumbai University.

Sample:

Sometimes the population is too large to study in its entirety. Sometimes collecting data from too large population becomes time-consuming and expensive. To save time and money, generally a part of population is selected for study. A sample is a part (a group of units) of population which is representative of the actual population. By studying the sample it is hoped that valid conclusions are drawn about the larger group

Example :

The population for a study of infant health might be all children born in India in one particular year. The sample might be all babies born on 7th May in that year.

Data can be classified based on the characteristics they have. It is of two types:

1. Variates

A characteristic which varies from one individual to another and can be expressed in numerical terms is called **variate**.

Example : Prices of a given commodity, wages of workers, heights and weights of students in a class etc.

2. Attributes

A characteristic which varies from one individual to another but can't be expressed in numerical terms is called **attribute**.

Example : Colour of ball (red, blue, green etc.), religion of human etc.

Quantitative variables can be further classified as discrete and continuous –

.

A **parameter** is a numerical value that states something about the entire population being studied. A parameter is usually unknown value which is fixed.
Example : Population mean, population standard deviation etc.

Since parameter is unknown, it has to be estimated. **Statistic** is used to estimate parameter. A statistic is a quantity that is calculated from a sample of data. It is used to give information about unknown values in the corresponding population. For example, the sample mean is used to estimate the parameter population mean. Statistic is also called Estimator.

Collection of Data

Researchers or investigators need to collect data from respondents. There are two types of data.

Primary Data

Primary data are data which is collected directly by investigator using such methods as:

Direct Interview Method: A face to face contact is made with the informants (persons from whom the information is to be obtained) under this **method of collecting data**. The interviewer asks them questions pertaining to the survey and collects the desired information.

Questionnaires: Questionnaires are survey instruments that can contain short closed-ended questions (multiple choice) or broad open-ended questions. Questionnaires are used to collect data from a large group of subjects on a specific topic. Currently, many questionnaires are developed and administered online.

Census and sample survey

In a **census**, data about all individual units (e.g. people or households) are collected in the population. In a **survey**, data are only collected for a sub-part of the population; this part is called a **sample**. These data are then used to estimate the characteristics of the whole population. In this case, it has to be ensured that the sample is representative of the population in question. For example, the proportion of people below the age of 18 or the proportion of women and men in the selected sample of households has to reflect the reality in the total population.

Secondary Data

Secondary data are the Second hand information. The data which have already been collected and processed by some agency or persons and are not used for the first time are termed as secondary data. According to M. M. Blair, "Secondary data are those already in existence and which have been collected for some other purpose." Secondary data may be abstracted from existing records, published sources or unpublished sources.

The distinction between primary and secondary data is a matter of degree only.
1. The data which are primary in the hands of one become secondary for all others.
For example, the population census report is primary for the Registrar General of India and the information from the report are secondary for all of us.

2. Both the primary and secondary data have their respective merits and demerits. Primary data are original as they are collected from the source. So they are more accurate than the secondary data. But primary data involves more money, time and energy than the secondary data. In an enquiry, a proper choice between the two forms of information should be made.

TABULATION OF DATA: The systematic and logical presentation of numeric data in rows and columns to facilitate comparison and statistical analysis is called **tabulation of data.**

CLASSIFICATION OF DATA: It is the process of **grouping** data into different categories on the basis of nature, behaviour or common characteristics.

CLASS MARK: The class mark is also known as the **class mid-point.** It is a specific point in the centre of the class intervals of each class in a frequency distribution table. It is calculated as below:

Class mark = (upper class limit + lower class limit)/2

CLASS LENGTH: The width of the class interval is known as class length.

RELATIVE FREQUENCY: The relative frequency of each class is equal to the frequency of each class divided by the total frequency. The sum of all relative frequencies is equal to one.

MEASURES OF CENTRAL TENDENCY: A single value that attempts to describe a set of data by identifying the central position within the set of data is called measure of central tendency. The most common measures of central tendency are mean, median and mode.

MEAN: The average of the numbers is known as mean. It is denoted by $\bar{x}$ where

$\bar{x} = \Sigma x/n$

For discrete distribution ,

$\bar{x} = \Sigma fx/\Sigma f$ where, Σf = total frequencies = N

MEDIAN: Median is defined as the value of the middle observation when the observations are arranged in the order of their magnitude.

Thus M = the value of (N + 1)/2 th observation if the number of observations are odd.

If the number of observations are even, then

M = {(N/2)+(N+1/2)}/2 th observation

MEDIAN FOR CONTINUOUS DISTRIBUTION:

$M = l_1 + (N/2 - c)(l_2 - l_1)/f$

Where, l_1 = lower limit of the median class

L_2 = upper limit of the median class

f=frequency of the median class

c=cumulative frequency of the class preceding the median class

N= total frequency

Median class is the class in which the value is very near to the value of N/2 in the cumulative frequency table(equal or next number in c.f table).

MODE: It is the value of the greatest frequency.

Mode in a continuous distribution is given by:

$Z = l_1 + (f_1 - f_0)(l_2 - l_1)/(2f_1 - f_0 - f_2)$

Where, l_1= lower limit of the modal class

L_1 = upper limit of the modal class

f_1= frequency of the modal class

f_0= frequency of the class preceding the modal class

f_2= frequency of the class succeeding the modal class

CUMULATIVE FREQUENCY: It is the sum of the frequencies in each class with the cumulative frequency of the previous class. There are two types of cumulative frequencies: less than type and greater than type.

(i) It is obtained by adding successively the frequencies of all the previous classes including the class against which it is written. In this type the cumulative is started from the first frequency.

(ii) It is obtained by subtracting the first frequency from the total frequency and successively subtracting the frequencies of the class against which it is written.

MERITS OF MEAN:

1. It is rigidly defined and has a definite value.
2. It is based on all observations.
3. It is easy to calculate and easy to understand.

DEMERITS OF MEAN:

1. It is affected by extreme values.
2. It is a value which may not be present in the data.
3. Sometimes it gives absurd result like 4.4 children per family.

MERITS OF MEDIAN:

1. It is rigidly defined.
2. It is not affected by extreme values.
3. Even if the extreme values are not known, median can be calculated if the number of items are known.

DEMERITS OF MEDIAN:

1. It is not based on all observations.
2. It is affected by sampling fluctuations.

3. It is not capable of further algebraic treatment.

MERITS OF MODE:

1. It is easy to calculate and understand.
2. It is not affected much by sampling fluctuations.
3. It is not necessary to know all items. Only the point of maximum concentration is required.

DEMERITS OF MODE:

1. It is ill defined as it is not based on all observations.
2. It is not capable of further algebraic treatment.
3. It is not a good representative.

CHARACTERISTICS OF A GOOD MEASURE OF DISPERSION:

1. It should be rigidly defined.
2. It should be based on all observations.
3. It should be easy to calculate and understand.
4. It should be capable of further algebraic treatment.
5. It should not be affected much by sampling fluctuations.

FOUR MEASURES OF DISPERSION:

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation

RANGE:

Range is defined as a difference between the highest and the lowest values taken by the variable x, i.e., if x_n is the maximum value and x_1 is the minimum value of x , then **range** is defined as

Range = $x_n - x_1$

Co-efficient of Range = $(x_n - x_1)/(x_n + x_1)$

MERITS OF RANGE:

1. It is easy to understand.
2. It is easy to calculate.

DEMERITS OF RANGE:

1. It is not based on all observations.
2. It does not have sampling stability. A single observation may change the value of range.
3. As the amount of data increases , range becomes less satisfactory.

QUARTILE DEVIATION OR SEMI- INTERQUARTILE RANGE:

It is the mid-point of the range between two quartiles. Quartile Deviation is defined as

$Q.D = (Q_3 - Q_1)/2$

Where **Q_1 = 1st quartile and Q_3 = 3rd quartile.**

Co-efficient of $Q.D = (Q_3 - Q_1)/ (Q_3 + Q_1)$

Interquartile range is defined as **$(Q_3 - Q_1)$.**

MERITS OF QUARTILE DEVIATION:

1. It is easy to calculate and understand.
2. It is not affected by extreme values.

DEMERITS OF QUARTILE DEVIATION:

1. It is not based on all observations.
2. It is not capable of further algebraic treatment.
3. It is affected by sampling fluctuations.

MEAN DEVIATION:

It is the deviations from mean or median when calculated considering their absolute values and are averaged. But median is the best to use because mean deviation from the median is less than that from any other value.

If **d** denotes the deviation from median, mean deviation is defined as:

$$\delta = (\Sigma \mid x - M \mid)/N = (\Sigma \mid d \mid)/N$$ where **M = median**

CO-EFFICIENT OF MEAN DEVIATION FROM MEDIAN: δ/M

$$\delta = (\Sigma \mid x - \bar{x} \mid)/ \Sigma f$$ where $\bar{x}$ **= mean**

CO-EFFICIENT OF MEAN DEVIATION FROM MEAN = δ/$\bar{X}$

If **d** denotes the deviation from mode, mean deviation is defined as:

$$\delta = (\Sigma \mid x - z \mid)/\Sigma f$$ where **z = mode**

CO-EFFICIENT OF MEAN DEVIATION FROM MODE = δ/Z

MERITS OF MEAN DEVIATION:

1. It is based on all observations.
2. It is easy to understand and also easy to calculate.
3. It is not affected by extreme values.

DEMERITS OF MEAN DEVIATION:

1. Mean deviation ignores algebraic signs, hence it is not capable of further algebraic treatment.
2. It is not very accurate measure of dispersion.

STANDARD DEVIATION:

Here, the deviations from arithmetic mean are squared, averaged and the square root of the resulting quantity is taken. So it is also known as ' **root mean-square deviation'.**

If a variable x takes values x_1, x_2,……, x_n, then the **standard deviation** is denoted by:

$$\sigma = \sqrt{[\ \Sigma(x - \bar{x})^2/N}$$

MERITS OF STANDARD DEVIATION:

1. It is rigidly defined and has a definite value.
2. It is based on all observations.
3. It is not affected much by sampling fluctuations.

DEMERITS OF STANDARD DEVIATION:

1. It is not easy to calculate.
2. It is not easy to understand.
3. It gives more weight to extreme items.

BIVARIATE LINEAR CORRELATION

If there exists a relation between a pair of variables (say x and y) such that change in one variable is accompanied by change in the other, then we say that variables x and y are correlated.

The correlation may be classified according to the following criteria:

(i) The number of variables: If correlation between only two variables is considered, it is bivariate or simple correlation. If the number of variables is more than two then the relation between one of the variables and a group of remaining variables is multiple correlation, e.g., price, demand and supply.

(ii) The Degree of Correlation: If two variables have an exact relation , i.e., the changes in one variable are proportional to the changes in the other , then there is linear correlation. Such relationship can be expressed with the help of an equation of a straight line of the type y = a + bx.

(iii) Types of Correlation : If both the variables increase or decrease together, then there is positive correlation between them, e.g., height and weight, income and expenditure, etc. On the other hand, if the changes in the variables are in opposite directions means that increase in one of them is reflected by decrease in the other and vice versa, then there is negative correlation , e.g., demand and price, number of strikes and industrial production, etc.

INTERPRETATION OF CORRELATION:

For bivariate data, measure of association between two variables is obtained by computing correlation coefficient which lies between

-1 and +1 can be interpreted only when two variables are meaningfully associated or related. Other it is termed as Spurious or Nonsense correlation. For eg: birth rate and production of T.V. sets in a country because these two have no relation between one another. The rise or fall in the birth rate has no relation with the production of T.V. sets.

In case the correlation coefficient is +1 or -1 it is called perfect correlation. In case the correlation coefficient is reaching +1 or -1 then there is high degree of correlation. When the correlation coefficient is reaching zero from positive and negative side of number line but not exactly zero then there is low degree of correlation.

The correlation coefficient lies between +1 and -1.

Thus, the value of the coefficient indicates the extent and nature of correlation between two variables. It is based on all the observations and hence it is affected by slight change in any observation.

METHODS OF DETERMINING CORRELATION

There are three important methods to ascertain correlation for a bivariate data:

(i) Scatter Diagram
(ii) Karl Pearson's Product Moment Coefficient
(iii) Spearman's Rank Correlation Coefficient

Scatter Diagram: The pairs of values of X and Y are represented by dot or points plotted on a graph paper. The graph is called a Scatter Diagram.

Karl Pearson's Product Moment Coefficient : The coefficient gives numerical measure of nature and extent of correlation. It is a pure number independent of the units of measurement of x and y . it always lies between -1 and +1. It is independent of change of origin and scale. It is defined as

$$r = [\Sigma (x - \bar{x})(y - \bar{y})]/n\sigma_x\sigma_y \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(i)$$

Where σ_x, σ_y are the standard deviation of x and y and n is the number of pairs of x and y.

We have,

$$Cov(x, y) = [\Sigma (x - \bar{x})(y - \bar{y})]/n\ldots\ldots\ldots\ldots\ldots\ldots(ii)$$

Hence r can also be defined as

$$r = Cov(x,y)/ \sigma_x\sigma_y$$

By substituting the values of σ_x and σ_y in (i), we get,

$$r=[1/n\ \Sigma (x - \bar{x})(y - \bar{y})]/\surd 1/n\{\Sigma(x - \bar{x})^2\}\surd 1/n\{\Sigma(y - \bar{y})^2\}$$

$= [\Sigma xy - (\Sigma x \Sigma y)/n]/\sqrt{[\Sigma x^2 - (\Sigma x)^2/n]}\sqrt{[\Sigma y^2 - (\Sigma y)^2/n]}$……..(iii)

Sometimes values of x and y are large and to simplify the calculations, we define new variables u and v as

$u=x - x_0$ and $v=y - y_0$ where x_0, y_0 are constants and $r_{xy} = r_{uv}$

$= \{\Sigma uv - [\Sigma u \Sigma v]/n\}/\sqrt{[\Sigma u^2 - (\Sigma u)^2/n]}\sqrt{[\Sigma v^2 - (\Sigma v)^2/n]}$……..(iv)

Spearman's Rank Correlation Coefficient:

Sometimes there are certain characteristics which are qualitative in nature and they cannot be measured numerically , eg., beauty, intelligence, skill etc. We can rank the individuals according to these characteristics in ascending or descending order, these ranks provide the data to calculate Spearman's Rank Correlation Coefficient which is derived from Karl Pearson's coefficient.

The formula for Rank Correlation Coefficient is

$R = 1 - 6\Sigma d^2/n(n^2 - 1)$

Where d represents difference between ranks i.e., $d = R_1 - R_2$, where R_1, R_2 are ranks assigned for the characteristics, n = number of pairs.

Repeated Ranks:

If two or more observations have the same value then common rank by considering the average can be given to all repeated ranks.

Here a correction factor is to be added to Σd^2 while calculating the rank correlation coefficient.

$C.F. = m(m^2 - 1)/12$

Where m is the number of times a rank is repeated.

When C.F. is added to the formula , Spearman's Rank Correlation Coefficient is given by

$R = 1- 6[\Sigma d^2 + \Sigma m(m^2 – 1)/12]/n(n^2 – 1)$

BIVARIATE LINEAR REGRESSION

Regression Analysis is a method of predicting or estimating one variable knowing the value of the other variable. Estimation is required in different fields in everyday life. A businessman wants to know the effect of increase in advertising expenditure on sales or a doctor wishes to observe the effect of a new drug on patients.

We observe , different pairs of variables related to each other like saving depends upon income, cost of production depends upon the number of units produced, the production depends on the number of workers present on a particular day etc. The relationship between two variables can be established with the help of any measure of correlation. When it is observed that two variables are highly correlated , it leads to interdependence of the variables. We can study the cause and effect relationship between them and then we can apply the regression analysis. The analysis helps in finding a mathematical model of the relationship.

Following are the methods of regression analysis:

(i) Scatter Diagram: After plotting the given set of values as points on a graph paper, we can study the nature of the diagram. Then a straight line can be drawn by inspection which seems to be the best fit for the given set of points. Some points will lie on the line and the others will be near the line. While drawing the line, care has to be taken about the number of points and below the line which should be approximately same.

This method provides a very rough estimate of the dependent variable, though it is easy to understand and simple to apply. Also , personal bias may be introduced while

drawing the estimating line and so estimated value obtained may noy be reliable.

(ii) Regression Lines: We can assume x as independent variable and y as dependent variable, to get regression equation of y on x, to be used to estimate y when x is known. Similarly, we can assume y as independent variable and x as dependent variable, to find regression equation of x on y , which can help in estimating x when y is known.

We use Least Square Method to determine the equations.

Least Square Method: A linear equation in two variables x and y represents a straight line. We try to minimize the sum of the squares of the distances of points from the line. For regression line of y on x, the distances parallel to y axis are considered and for regression line of x on y, the distances parallel to x axis are measured.

Let regression equation of y on x be $y = a + bx$ where a and b are constants to be determined. The slope of the line is represented by b and a represents the y-intercept of the line. To calculate the values of a and b, consider a set of n pairs of variables x and y. Different summations like Σx, Σy, Σx^2 and Σxy can be calculated.

The regression equation is

$y = a + bx$……………………………………….(i)

Taking sum over all n observations

$\Sigma y = na + b\Sigma x$……………………………………(ii)

Multiplying (i) by x and then taking summation, we get

$\Sigma xy = a\Sigma x + b\Sigma x^2$ ……………………………….(iii)

Equations (ii) and (iii) are called Normal Equations.

The values of the summations can be substituted and normal equations can be solved simultaneously to obtain a and b. After finding a and b , the regression equation of y on x represented by (i) can be formed which can be used to estimate y when x is known.

Similarly, let the regression of x on y be $x = a_1 + b_1y_1$ where b_1 represents the reciprocal of the slope of the line and a_1 is the x-intercept of the line.

To find the values of a_1 and b_1 consider n pairs of values of variables x, y to get summations Σx, Σy, Σxy and Σy^2.

The regression equation of x on y is

$x = a_1 + b_1y$ ……..(iv)

Taking summation over all n pairs

$\Sigma x = na_1 + b_1 \Sigma y$ ……..(v)

Multiplying (iv) by y and then taking sum we get,

$\Sigma xy = a_1 \Sigma y + b1\Sigma y^2$ …….(vi)

Equations (v) and (vi) are Normal Equations which are to be solved simultaneously to get the values of a_1 and b_1.

After obtaining a_1 and b_1, the regression equation of x on y expressed as (iv) can be formed which can be used to estimate x when y is known.

Thus using least square method, we can get both the regression equations, which can be used for future estimation.

Some Properties of Regression Equations:

We have seen that two regression equations and correspondingly two regression lines can be drawn, one where x is independent and y is dependent(y on x) and the other is where y is independent and x is dependent (x on y). It is obvious that the regression coefficients b, b_1 represent the slope of the regression lines.

If there is a perfect positive or negative correlation between the variables, then the two regression lines coincide.

If there is high degree of correlation than the angle between the two interesting regression lines is small. The angle becomes large as the correlation decreases and the lines become perpendicular to each other if there is no correlation between the variables.

Thus, less angle means less slope of the line leading to high degree of correlation and large angle means more slope leading to less degree of correlation. So in general, more the slope means less correlation and vice versa.

The point $(\bar{x}, \bar{y})$ satisfies both the regression equations as it lies on both the lines so that it is the point of intersection of the two lines. This can be helpful to us whenever the regression equations are known and the mean values of x and y are to be obtained. In this case, the two regression equations can be solved simultaneously and the common solution represents $\bar{x}$ and $\bar{y}$.

Regression coefficients b and b_1 can be expressed as

$b = r\sigma_y/\sigma_x$ and $b_1 = r\sigma_x/\sigma_y$

Hence, $b * b_1 = r\sigma_y/\sigma_x * r\sigma_x/\sigma_y = r^2$

$\Rightarrow r=\pm\sqrt{b * b_1}$

Note that r is positive if b and b_1 both are positive and r is negative if b and b_1 both are negative.

Thus, r, the correlation coefficient , is the geometric mean of the regression coefficients b and b_1.

This property can be used to obtain r, the correlation coefficient from the regression equations.