STATISTICS AND QUANTITATIVE METHODS FOR BUSINESS

UNIT 1

INTRODUCTION TO STATISTICS

STATISTICS-ORIGIN

- The Word statistics have been derived from Latin word “Status” or the Italian word “Statista”, meaning of these words is “Political State” or a Government.
- In the past, the statistics was used by rulers. The application of statistics was very limited but rulers and kings needed information about lands, agriculture, commerce, population of their states to assess their military potential, their wealth, taxation and other aspects of government.

DEFINING STATISTICS

1. Statistics can be defined as the collection presentation and interpretation of numerical data- Croxton and Cowden.
2. Statistics are numerical statement of facts in any department of enquiry placed interrelation to each other- Bowley.
3. Statistics are the classified facts representing the conditions of the people in a State specially those facts which can be stated in numbers or any tabular or classified arrangement. -Webster
4. The science of Statistics is essentially a branch of applied mathematics and can be regarded as a mathematics applied to observation data- R.A fisher.
5. Statistics is the science which deals with the collection, classification and tabulation of numerical facts as the basis of explanation, description and comparison of phenomenon – Lovitt.

FUNCTIONS OF STATISTICS

1. To present facts in a definite form
2. To simplify and represent complex data
3. To use it as a technique for making comparisons
4. To enlarge individual experience
5. To provide guidance in the formulation of policies
6. To enable measurement of the magnitude of a phenomenon

SCOPE OF STATISTICS

1.Social Sciences

-Man Power Planning

-Crime Rates

-Income & Wealth Analysis of Society

-In studying Pricing, Production, Consumption, Investments & Profits etc.

2.Planning

-Agriculture

-Industry

-Textiles

-Education etc.

For ex. Five Year Plans in India.

SCOPE OF STATISTICS

3. Mathematics

-Extensive use of Differentiation, Algebra, Trigonometry, Matrices etc in modern business analysis.

-Statistics now treated as Applied Mathematics.

4. Economics

- Family Budgeting

-Applied in solving economic problems related to production, consumption, distribution of products as per income & wealth related patterns, wages, prices, profits & individual savings, investments, unemployment & poverty etc.

SCOPE OF STATISTICS

5. Business Management

- Trend Analysis
- Market Research & Analysis
- Product Life Cycle

i) Marketing

- Marketing Policy Decisions depend on forecasting, demand analysis, time & motion studies, inventory control, investments & analysis of consumer data for production & sales.

SCOPE OF STATISTICS

ii) Production

- Designs
  - Methods of Production
  - Technology Selection
  - Quality Control Mechanisms
  - Product Mix
  - Quantities
  - Time Schedules for Manufacturing & Distribution

SCOPE OF STATISTICS

iii) Finance

-Correlation Analysis of profits & dividends, assets & liabilities

-Analysis of income & expenditure

- Financial forecasts, break-even analysis, investment & risk analysis

iv) Sales

-Demand Analysis

-Sales Forecasts

v) Personnel

- Wage plans, Incentive plans, Cost of living, Labor turnover ratio, Employment trends, Accidental Rates, Performance Appraisals etc.

SCOPE OF STATISTICS

vi) Accounting & Auditing

-Analysis of Income, Expenditure, Investment, Profits and Optimization of Production etc

- Forecasting costs of production & price

vii) Other Areas

-Insurance, Astronomy, Medical Sciences, Psychology, Education, war field, etc.

LIMITATIONS OF STATISTICS

- Does not study individual items, deals with aggregates.
- Statistical laws are not exact.
- Not suitable for the study of qualitative phenomenon.
- Statistical methods are only means and not end for solving problems.
- Statistics is liable to be misused.

TERMINOLOGIES

- **Descriptive Statistics**: collection, presentation, and description of sample data.
- **Inferential Statistics**: making decisions and drawing conclusions about populations.
- **Population**: A collection, or set, of individuals or objects or events whose properties are to be analyzed. Two kinds of populations: *finite* or *infinite*.
- **Sample**: A subset of the population.
- **Variable**: A characteristic about each individual element of a population or sample.

TERMINOLOGIES

- **Data (singular)**: The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.
- **Data (plural)**: The set of values collected for the variable from each of the elements belonging to the sample.
- **Experiment**: A planned activity whose results yield a set of data.
- **Parameter**: A numerical value summarizing all the data of an entire population.
- **Statistic**: A numerical value summarizing the sample data.

EXAMPLE: A college Principal is interested in learning about the average age of faculty. Identify the basic terms in this situation.

The *population* is the age of all faculty members at the college.

A *sample* is any subset of that population. For example, we might select 10 faculty members and determine their age.

The *variable* is the "age" of each faculty member.

One *data* would be the age of a specific faculty member.

The *data* would be the set of values in the sample.

The *experiment* would be the method used to select the ages forming the sample and determining the actual age of each faculty member in the sample.

The *parameter* of interest is the "average" age of all faculty at the college.

The *statistic* is the "average" age for all faculty in the sample.

TERMINOLOGIES

- **Frequency**: The frequency (f) of a particular observation is the number of times the observation occurs in the data.
- The *distribution* of a variable is the pattern of frequencies of the observation. Frequency distributions are portrayed as frequency tables, histograms or polygons.
- **Frequency distributions** are visual displays that organise and present frequency counts so that the information can be interpreted more easily.
- Frequency distributions can show absolute frequencies or relative frequencies, such as proportions or percentages.

TERMINOLOGIES

- When a variable can take some discrete or isolated values within its range of variation, then it is known as **discrete variable**. For example, children per family of a village, number of misprints per page of a book, number of telephone calls, number of students in a class etc. are discrete variables.
- Data on discrete variables is known as discrete data and its distribution is called **Discrete frequency distribution.**

TERMINOLOGIES

- A variable is said to be continuous if it can assume any numerical value within its range of variation. For example, daily temperatures of a place in a month, blood pressure of a group of people, weight, height, IQ of a student etc. are **continuous variables**.
- Data on continuous variables is known as continuous data and its distribution is **Continuous frequency distribution.**
- **Cumulative Frequency** of a class is the sum of the frequency of that class and the frequencies of all the preceding or succeeding classes which are listed in some sensible order (numerical order, alphabetical order, etc.)

TERMINOLOGIES

- **Class Interval:** The whole range of variable values is classified in some groups in the form of intervals. Each interval is called a class interval.
- **Class Frequency:** The number of observations in a class is termed as the frequency of the class or class frequency.
- **Class limits** are the two endpoints of a class interval which are used for the construction of a frequency distribution. The lowest value of the variable that can be included in a class interval is called the lower class limit of that class interval. The highest value of the variable that can be included in a class interval is called the upper class limit of that class interval. These are not the real limits or endpoints of a class interval. Hence, class limits are called apparent limits of a class

MEASURES OF CENTRAL TENDENCY

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.
- As such, measures of central tendency are sometimes called measures of central location.
- They are also classed as summary statistics.
- The mean (often called the average) is most likely the measure of central tendency that is the most familiar with, but there are others, such as the median, mode, geometric mean and harmonic mean.

REQUISITES FOR AN IDEAL MEASURE OF CENTRAL TENDENCY

1. It should be rigidly defined so as to avoid different people choosing different values for the same measure of central tendency.
2. It should be easily comprehensible and easy to calculate.
3. It should be based upon all observations.
4. It should be amenable for further mathematical treatment.
5. It should be affected as little as possible by the presence of extreme values.
6. It should be least affected by sampling fluctuation, i.e., an ideal measure of central tendency should not vary in its value too much from one sample to another when all the samples are taken from the same set of observations.

MEAN

- The mean (or average) is the most popular and well known measure of central tendency.
- **The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.**
- It can be used with both discrete and continuous data, although its use is most often with continuous data.

CALCULATION OF MEAN

- The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x_1, x_2, ..., x_n, the sample mean, usually denoted by (pronounced x bar), is:
- $\mu = \frac{x_1 + x_2 + \ldots.. x_n}{N} = \frac{\Sigma x}{N}$

CALCULATION OF MEAN

For discrete frequency distributions,

Arithmetic Mean $\mu = \frac{f_1x_1 + f_2x_2 + \ldots.f_nx_n}{N} = \frac{\sum fx}{\sum f}$

Where $N = f_1 + f_2 + \ldots + f_n$

n = no. of observations

X	Freq.	fx
x_1	f_1	f_1x_1
x_2	f_2	f_2x_2
x_3	f_3	f_3x_3
x_4	f_4	f_4x_4

CALCULATION OF MEAN

- **Continuous Frequency Distribution**

Direct Method:

$$\mu = \frac{f_1x_1 + f_2x_2 + \ldots..f_nx_n}{N} = \frac{\Sigma fx}{\Sigma f}$$

Where $N = f_1 + f_2 + \ldots + f_n$

x = mid value of a C.I

$$= \frac{(U.L + L.L)}{2}$$

Assumed Mean Method:

$$\mu = A + \frac{\Sigma fd}{N}$$

Where A = assumed mean

$N = \Sigma f$

$d = [x - A]$

x = mid - value

Step Deviation Method:

$$\mu = A + \frac{\Sigma fd \times i}{N}$$

where A = assumed mean

$N = \Sigma f$

$$d = \frac{x - A}{i}$$

x = mid – value

i = width of C.I = U.L – L.L

Properties of Arithmetic Mean

1. The sum of deviations of the items from the arithmetic mean is always zero i.e.

   $\sum(X-\bar{X}) = 0$.

2. The Sum of the squared deviations of the items from A.M. is minimum, which is less than the sum of the squared deviations of the items from any other values.
3. The product of the arithmetic mean and the number of items gives the total of all items.

$$\bar{x} = \frac{\sum x_i}{N} \Rightarrow \bar{x} \cdot N = \sum x_i$$

Properties of Arithmetic Mean

4. Mean of Combined Series: If and are the arithmetic mean of two samples of sizes n_1 and n_2 respectively then, the arithmetic mean of the distribution combining the two can be calculated as

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

MERITS AND DEMERITS OF MEAN

MERITS:

- Is rigidly defined and has a definite value.
- Is based on all the observations.
- Is capable of algebraic treatments for further data analysis & interpretation.
- Easy to calculate & simple to understand.
- For a large no. of observations, A.M provides a good basis of comparison.

DEMERITS:

- Cannot be calculated even if a single observation is missing.
- Cannot be obtained just by inspection as in case of median & mode.
- May give absurd results.
- Cannot be used for qualitative data.

Weighted Average/Mean

- A weighted average is an average estimated with due weight or importance given to all the observations. The terms ‘weight’ stands for the relative importance of the different observations.
- Formula:

$$Xw = \frac{\sum WX}{\sum W}$$

MEDIAN

- **The median is the *middle value* in distribution when the values are arranged in ascending or descending order.**

- Its value is the value of the middle item irrespective of all other values.

CALCULATION OF MEDIAN

Individual Series

N = no. of observations or items in the series

- Arrange all the items in ascending or descending order of magnitude.

Case I

Discrete Data/ Ungrouped Data:

Median = Value at $\frac{(N+1)}{2}$ th position in the arranged series.

CALCULATION OF MEDIAN

Case 2:

Continuous Frequency Distribution/Grouped Data:

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

Where:

l = lower class boundary of the median class

h = Size of the median class interval

f = Frequency corresponding to the median class

N = Total number of observations i.e. sum of the frequencies

c = Cumulative frequency preceding median class.

MERITS AND DEMERITS OF MEDIAN

MERITS:

- Is rigidly defined.
- Can be easily calculated.
- Not affected by extreme values.
- Can be located merely by inspection.

DEMERITS:

- May not represent the entire series in many cases.
- Not suitable for further algebraic treatment.
- More likely to be affected by sampling fluctuations.

Positional Measures

- Positional measures are those that are estimated by dividing a series into a equal number of parts.
- Important amongst these are quartiles, deciles and percentiles.
- Quartiles are divides the total frequency into four equal parts
- Deciles divide the total frequency in 10 equal parts
- Percentiles divide the total frequency in 100 equal parts.

MODE

- The value occurring the largest no. of times in a series. That is the value having the maximum frequency.
- Is calculated for discrete and continuous frequency distributions only.
- For ungrouped data the modal value is identified by inspection

- Mode for Grouped data:

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h$$

where l = lower limit of the modal class,

h = size of the class interval (assuming all class sizes to be equal),

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.

- Empirical Relationship between Mean, Median & Mode:

Mode = 3 (Median) – 2 (Mean)

MERITS AND DEMERITS OF MODE

MERITS:

- Readily comprehensible and easy to calculate.
- Mode can be located in some cases merely by inspection.
- It is not affected by extreme values.
- Mode can be located easily from any type of frequency distribution.

DEMERITS:

- It is not based upon all the observations.
- It is not useful for further mathematical treatment.
- Compared to mean, mode is affected to a greater extent by fluctuations of sampling.
- Distributions with bi-modal or multimodal values.

MEASURES OF DISPERSION

- The measures of central tendency are not adequate to describe data.
- Two data sets can have the same mean but they can be entirely different.
- Thus to describe data, one needs to know the extent of variability. This is given by the measures of dispersion.
- Range, inter-quartile range, and standard deviation are the three commonly used measures of dispersion.

RANGE

- The simplest of our methods for measuring dispersion is *range*.
- Range is the difference between the largest value and the smallest value in the data set.
- Formula: Range = L-S
- Coefficient of Range $= \frac{L-S}{L+S}$
- While being simple to compute, the range is often unreliable as a measure of dispersion since it is based on only two values in the set.

Quartile Deviation

- It depends on the lower quartile $Q1$ and the upper quartile $Q3$. The difference $Q3-Q1$ is called the inter quartile range. The difference $Q3-Q1$ divided by 2 is called semi-inter quartile range or the quartile deviation.
- Formulas: (i) **Quartile deviation** $= \dfrac{Q_3 - Q_1}{2}$

$$(ii)\ \text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Mean Deviation

- Referred to as average deviation, it is defined as the sum of the deviations(ignoring signs) from an average divided by the number of items in a distribution.
- The mean deviation can be calculated from mean, median or mode.
- Theoretically median is the best average of choice because sum of deviations from median is minimum.

Formula for calculating Mean Deviation

- For Individual Series:

  Mean Deviation = $\sum |D| / N$

  where, $|D| = |X - M|$ and M=Mean/Median

- For grouped data:

  Mean Deviation = $\sum f|D| / N$

  where, $|D| = |X - M|$ and M=Mean/Median

- Coefficient of Mean Deviation

  = Mean Deviation / M

Where M = Median

STANDARD DEVIATION

- **Standard deviation** is the square root of the average of squared deviations of the items from their mean. Symbolically it is represented by σ.
- The variance is a measure in squared units and has little meaning with respect to the data.
- Thus, the standard deviation is a measure of variability expressed in the same units as the data.
- Formula:

  Standard Deviation, $\sigma = \sqrt{[\ \Sigma\ d^2 / N\ ]}$

  where $d = (X - \mu)$ and μ = Mean

  N = No. of Observations

– Coefficient of Standard Deviation = σ / μ

Formula for calculating Standard Deviation

- For Discrete data:

Population Data

$$Population\ variance\ (\sigma^2) = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2$$

$$Population\ Standard\ Deviation\ (\sigma) = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$

- For grouped data: Without calculating D Values:

Population Data

$$Population\ variance\ (\sigma^2) = \frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2$$

$$Population\ Standard\ Deviation\ (\sigma) = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

- For Discrete Series:

$$\text{Standard deviation} = \sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

$$where\ d = x - A, \qquad A = assumed\ mean,$$
$$n = total\ number\ of\ observation.$$

- For Continuous frequency distribution:

$$\text{Standard deviation} = \sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times i$$

$$\text{Where} i = \text{common class interval},$$
$$d = \frac{x - A}{i},$$

$A =$ is assumed mean

$f =$ f is the respective frequency.

VARIANCE

- The variance is a measure based on the deviations of individual scores from the mean.
- The variance is based on squared deviations of scores about the mean.
- The result is the average of the sum of the squared deviations and it is called the variance.

Coefficient of Variation

- The coefficient of variation (CV) is a measure of relative variability.
- It is the ratio of the standard deviation to the mean.
- Formula:

$$CV\ (\%) = \left(\frac{Standard\ deviation}{Mean}\right) \times 100$$

Moments

- The concept of moments has crept into the statistical literature from mechanics. In mechanics, this concept refers to the turning or the rotating effect of a force whereas it is used to describe the peculiarities of a frequency distribution in statistics.
- Moments also help in measuring the scatteredness, asymmetry and peakedness of a curve for a particular distribution.
- Moments about mean are generally used in statistics.

Calculating Moments

- First order, second order, third order and fourth order moments are calculated using the following formulas:

  - First Moment $$\mu_1 = \frac{\sum f(x-\bar{x})}{N}$$
  - Second Moment $$\mu_2 = \frac{\sum f(x-\bar{x})^2}{N}$$
  - Third Moment $$\mu_3 = \frac{\sum f(x-\bar{x})^3}{N}$$
  - Fourth Moment $$\mu_4 = \frac{\sum f(x-\bar{x})^4}{N}$$

STATISTICALLY DESCRIBING DISTRIBUTIONS

- **Symmetric** - Distributions that have the same shape on both sides of the centre are called symmetric.
- A symmetric distribution with only one peak is referred to as a **normal distribution**.

SKEWNESS

- **Skewness** - Refers to the degree of asymmetry in a distribution. Asymmetry often reflects extreme scores in a distribution.
  - **Positively skewed** - A distribution is positively skewed when is has a tail extending out to the right (larger numbers) When a distribution is positively skewed, the mean is greater than the median reflecting the fact that the mean is sensitive to each score in the distribution and is subject to large shifts when the sample is small and contains extreme scores.
  - **Negatively skewed** - A negatively skewed distribution has an extended tail pointing to the left (smaller numbers) and reflects bunching of numbers in the upper part of the distribution with fewer scores at the lower end of the measurement scale.

Formulas to Calculate Skewness

$$\text{Pearson's coefficient of skewness} = \frac{\text{Mean - Mode}}{\text{Standard deviation}}$$

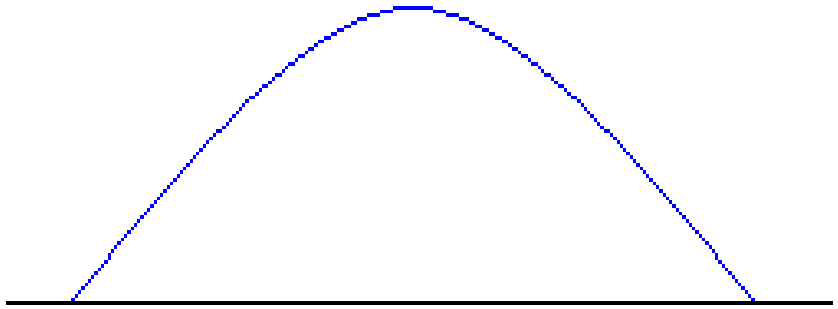
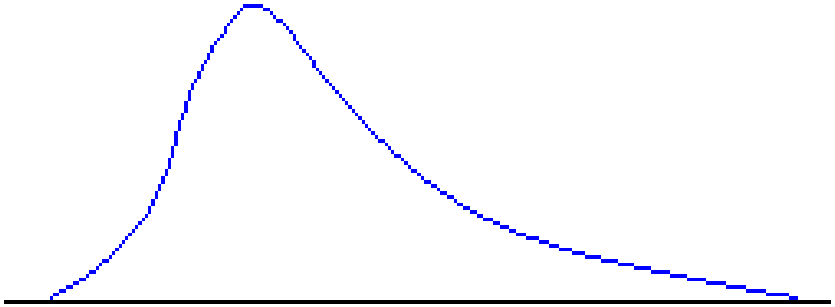
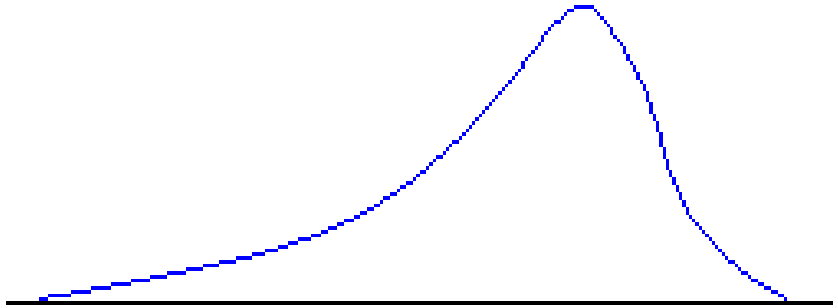
$$\text{Pearson's coefficient of skewness} = \frac{3(\text{Mean - Median})}{\text{Standard deviation}}$$

$$\text{Kelley's coefficient of skewness} = \frac{D_9 + D_1 - 2M_e}{D_9 - D_1}$$

$$\text{Bowley's coefficient of skewness} = \frac{(Q_3 + Q_1) - 2M_e}{Q_3 - Q_1}$$

$$\text{Moment based coefficient of skewness} = \frac{\mu_3^2}{\mu_2^3}$$

Terms that Describe Distributions

Term	Features	Example
"Symmetric"	left side is mirror image of right side	
"Positively skewed"	right tail is longer then the left	
"Negatively skewed"	left tail is longer than the right	

KURTOSIS

- **Kurtosis** - Like skewness, kurtosis has a specific mathematical definition, but generally it refers to how scores are concentrated in the center of the distribution, the upper and lower tails (ends), and the shoulders (between the center and tails) of a distribution.
  - **Mesokurtic** - A **normal distribution** is called mesokurtic. The tails of a mesokurtic distribution are neither too thin or too thick, and there are neither too many or too few scores in the center of the distribution.
  - **Platykurtic** - Starting with a mesokurtic distribution and moving scores from both the center and tails into the shoulders, the distribution flattens out and is referred to as platykurtic.
  - **Leptokurtic** - If you move scores from shoulders of a mesokurtic distribution into the center and tails of a distribution, the result is a peaked distribution with thick tails. This shape is referred to as leptokurtic.

Formula

- Co-efficient of Skewness $$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$
- Kurtosis $$\beta_2 = \frac{\mu_4}{\mu_2^{\ 2}}$$

Symmetrical, Bell-Shaped