UNIVERSITY GRANTS COMMISSION

COMMERCE **CODE: 08**

UNIT – 5: BUSINESS STATISTICS AND RESEARCH METHODS

SYLLABUS

Sub Unit – 1: MEASURES OF CENTRAL TENDENCY
Sub Unit – 2: MEASURES OF DISERSION
Sub Unit – 3: MEASURES OF SKEWNESS
Sub Unit – 4: CORRELATION AND REGRESSION OF TWO VARIABLES
Sub Unit – 5: PROBABILITY

SL. NO	TOPICS
5.5.1	Approaches to Probability
5.5.2	Bayes' Theorem

Sub Unit –6: PROBABILITY DISTRIBUTION

SL. NO	TOPICS
5.6.1	Binomial Distribution
5.6.2	Poisson Distribution
5.6.3	Normal Distribution

Sub Unit – 7: RESEARCH

SL. NO	TOPICS
5.7.1	Concept and Types
5.7.2	Research Designs

Sub Unit – 5.8: DATA

SL. NO	TOPICS
5.8.1	Collection of Data
5.8.2	Classification of Data

Sub Unit – 5.9: SAMPLING AND ESTIMATION

SL. NO	TOPICS
5.9.1	Concept of Sampling and Estimation
5.9.2	Methods of Sampling – Probability and Non-probability Methods
5.9.3	Sampling Distribution
5.9.4	Central Limit Theorem
5.9.5	Standard Error
5.9.6	Statistical Estimation

Sub Unit – 5.10 HYPOTHESIS TESTING

SL. NO	TOPICS
5.10.1	z – test
5.10.2	t – test
5.10.3	ANOVA
5.10.4	Chi-square test
5.10.5	Mann-Whitney test (U- test)
5.10.6	Kruskal-Wallis test (H-test)
5.10.7	Rank correlation test

Sub Unit – 11: REPORT WRITING

SECTION – 1: UNITS AT A GLANCE

Sub Unit – 1: MEASURES OF CENTRAL TENDENCY

Definition: Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. Along with the variability (dispersion) of a dataset, central tendency is a branch of descriptive statistics.

Measures of Central Tendency: Generally, the central tendency of a dataset can be described using the following measures:

- **Mean (Average)**
- **Median**
- **Mode**

Arithmetic Mean: The arithmetic mean of a set of observed data is defined as being equal to the sum of the numerical values of each and every observation divided by the total number of observations.

Geometric Mean: A special type of average where we multiply the numbers together and then take a square root (for two numbers), cube root (for three numbers) etc.

In mathematics, the **geometric mean** is a mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the *n*th root of the product of *n* numbers.

Harmonic Mean: In mathematics, the **harmonic mean** (sometimes called the **subcontrary mean**) is one of several kinds of average, and in particular, one of the Pythagorean means. Typically, it is appropriate for situations when the average of rates is desired.

The harmonic mean can be expressed as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations.

Median: In statistics and probability theory, the **median** is the value separating the higher half from the lower half of a data sample, a population or a probability distribution. For a data set, it may be thought of as the "middle" value.

Mode: The mode is the number that appears most frequently in a set. A set of numbers may have one mode, more than one mode, or no mode at all.

Mean Median Mode Relation With Frequency Distribution

- **Frequency Distribution with Symmetrical Frequency Curve**

If a frequency distribution graph is having a symmetrical frequency curve, the mean, median, and mode will be equal.

- **For Positively Skewed Frequency Distribution**

In case of a positively skewed frequency distribution, the mean is always greater than median and the median is always greater than the mode.

- **For Negatively Skewed Frequency Distribution**

In case of a negatively skewed frequency distribution, the mean is always lesser than median and the median is always lesser than the mode.

Sub Unit – 2: MEASURES OF DISERSION

Dispersion (also called **variability**, **scatter**, or **spread**) is the extent to which a distribution is stretched or squeezed.[1] Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range. Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions.

Absolute Measure of Dispersion: The types of absolute measures of dispersion are:

1. Range
2. Variance
3. Standard Deviation
4. Quartiles and Quartile Deviation
5. Mean and Mean Deviation

Relative Measure of Dispersion: Common relative dispersion methods include:

1. Coefficient of Range
2. Coefficient of Variation
3. Coefficient of Standard Deviation
4. Coefficient of Quartile Deviation
5. Coefficient of Mean Deviation

Coefficient of Dispersion: The coefficients of dispersion are calculated along with the measure of dispersion when two series are compared which differ widely in their averages. The dispersion coefficient is also used when two series with different measurement unit are compared. It is denoted as C.D.

Sub Unit – 3: MEASURES OF SKEWNESS

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined.

Absolute skewness: This is obtained by finding the difference between any two measures of dispersion viz: Mean, Median and Mode.

Co-efficient of skewness: This is obtained by dividing the Skewness by any measure of dispersion.

Karl Pearson's Skewness: Prof. Karl pearson says that to study the skewness of a series, the difference between the Mean and Mode only should be found out. This is because, Mean is an average which is affected very much by the extreme values of a series and Mode is an average which is least affected by the extreme values of a series. Thus, according to him,

$Sk_{(p)}$ = Mean –Mode

Bowley's Skewness: According to Prof. A.L. Bowley, the presence, or absence of skewness will be determined on the basis of the distance of the quartiles from the Median. Thus, his skewness is given by

$$SK(B) = (Q_3 - M) - (M - Q_1)$$

Or $= Q_3 + Q_1 - 2M$

Sub Unit – 4: CORRELATION AND REGRESSION OF TWO VARIABLES

Correlation, in the finance and investment industries, is a statistic that measures the degree to which two securities move in relation to each other. Correlations are used in advanced portfolio management, computed as the correlation coefficient, which has a value that must fall between -1.0 and +1.0.

Types of Correlation:

In a bivariate distribution, the correlation may be:

1. Positive, Negative and Zero Correlation; and
2. Linear or Curvilinear (Non-linear).

Methods of Computing Co-Efficient of Correlation: In case of ungrouped data of bivariate distribution, the following three methods are used to compute the value of co-efficient of correlation:

1. Scatter diagram method.
2. Pearson's Product Moment Co-efficient of Correlation.
3. Spearman's Rank Order Co-efficient of Correlation.

Spearman's Rank Correlation Coefficient: There are some situations in Education and Psychology where the objects or individuals may be ranked and arranged in order of merit or proficiency on two variables and when these 2 sets of ranks covary or have agreement between them, we measure the degrees of relationship by rank correlation.

Again, there are problems in which the relationship among the measurements made is non-linear, and cannot be described by the product-moment r.

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which a researcher finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.

Classification of regression analysis: The regression analysis can be classified on the following bases: -

(i) Change in Proportion; and

(ii) Number of variables

Basis of Change in Proportions: On the basis of proportions the regression can be classified into the

following categories: -

1. Linear regression and
2. Non-linear regression

Simple regression: When only two variables are studied to find the regression relationships, it is known as simple regression analysis. Of these variables, one is treated as an independent variable while the other as dependent one.

2. Partial Regression: When more than two variables are studied in a functional relationship but
the relationship of only two variables is analyzed at a time, keeping other variables as constant, such a regression analysis is called partial regression.

3. Multiple Regression: When more than two variables are studied and their relationships are simultaneously worked out, it is a case of multiple regression.

Regression Lines: A regression line is a line that best describes the behavior of a set of data. In other words, it's a line that best fits the trend of a given data.

Regression lines are very useful for forecasting procedures. The purpose of the line is to describe the interrelation of a dependent variable (Y variable) with one or many independent variables (X variable). By using the equation obtained from the regression line an analyst can forecast future behaviors of the dependent variable by inputting different values for the independent ones.

Methods of Drawing Regression Lines: The regression lines can be drawn by two methods as given below: -

1. Free Hand Curve Method
2. The method of Least Squares

Difference between Correlation and Regression: **Correlation** is described as the analysis which lets us know the association or the absence of the relationship between two variables 'x' and 'y'. On the other end, **Regression** analysis, predicts the value of the dependent variable based on the known value of the independent variable, assuming that average mathematical relationship between two or more variables.

Sub Unit – 5: PROBABILITY

Approaches to Probability: The three approaches to probability are:

- Classical approach
- Frequency-based (or empirical) approach
- Subjective approach

Bayes' Theorem: In statistics and probability theory, the Bayes' theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes' theorem describes the probability.

Total Probability Rule: The Total Probability Rule (also known as the law of total probability) is a fundamental rule in statistics relating to conditional and marginal of an event based on prior knowledge of the conditions that might be relevant to the event.

The theorem is named after English statistician Thomas Bayes, who discovered the formula in 1763. It is considered the foundation of the special statistical inference approach called the Bayes' inference.

Sub Unit –6: PROBABILITY DISTRIBUTION

What is Binomial Distribution? Binomial distribution is a common probability distribution that models the probability of obtaining one of two outcomes under a given number of parameters. It summarizes the number of trials when each trial has the same chance of attaining one specific outcome. The value of binomial is obtained by multiplying the number of independent trials by the successes.

Bernoulli distribution: The Bernoulli distribution is a special case of the binomial distribution, where $n = 1$. Symbolically, $X \sim B(1, p)$ has the same meaning as $X \sim \text{Bernoulli}(p)$. Conversely, any binomial distribution, $B(n, p)$, is the distribution of the sum of n Bernoulli trials, Bernoulli(p), each with the same probability p.

Poisson approximation: The binomial distribution converges towards the Poisson distribution as the number of trials goes to infinity while the product np remains fixed or at least p tends to zero. Therefore, the Poisson distribution with parameter $\lambda = np$ can be used as an approximation to $B(n, p)$ of the binomial distribution if n is sufficiently large and p is sufficiently small. According to two rules of thumb, this approximation is good if $n \geq 20$ and $p \leq 0.05$, or if $n \geq 100$ and $np \leq 10$.

Normal approximation: If n is large enough, then the skew of the distribution is not too great. In this case a reasonable approximation to $B(n, p)$ is given by the normal distribution and this basic approximation can be improved in a simple way by using a suitable continuity correction. The basic approximation generally improves as n increases (at least 20) and is better when p is not near to 0 or 1. Various rules of thumb may be used to decide whether n is large enough, and p is far enough from the extremes of zero or one.

Sub Unit – 7: RESEARCH

Concept of Research: According to the American sociologist Earl Robert Babbie, "Research is a systematic inquiry to describe, explain, predict, and control the observed phenomenon. Research involves inductive and deductive methods."

Types of Research

- **Basic Research:** A basic research definition is data collected to enhance knowledge. The main motivation is knowledge expansion. It is a non-commercial research that doesn't facilitate in creating or inventing anything.
- **Applied Research:** Applied research focuses on analyzing and solving real-life problems. This type refers to the study that helps solve practical problems using scientific methods.
- **Problem Oriented Research:** As the name suggests, problem-oriented research is conducted to understand the exact nature of a problem to find out relevant solutions. The term "problem" refers to multiple choices or issues when analyzing a situation.
- **Problem Solving Research**: This type of research is conducted by companies to understand and resolve their own problems. The problem-solving method uses applied research to find solutions to the existing problems.

- **Qualitative Research:** Qualitative research is a process that is about inquiry. It helps create in-depth understanding of problems or issues in their natural settings. This is a non-statistical method.
- **Quantitative Research:** Qualitative research is a structured way of collecting data and analyzing it to draw conclusions. Unlike qualitative methods, this method uses a computational and statistical process to collect and analyze data. Quantitative data is all about numbers.

Research design: It is the framework of research methods and techniques chosen by a researcher. The design allows researchers to hone in on research methods that are suitable for the subject matter and set up their studies up for success.

The design of a research topic explains the type of research (experimental, survey, correlational, semi-experimental, review) and also its sub-type (experimental design, research problem, descriptive case-study).

There are three main types of research design: Data collection, measurement, and analysis.

Sub Unit – 8: DATA

Data Collection: Data collection is defined as the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research.

Data Collection Methods:

1. Closed-ended Surveys and Online Quizzes
2. Open-Ended Surveys and Questionnaires
3. 1-on-1 Interviews
4. Focus groups
5. Direct observation

Steps for Effective Data Classification

- **Understand the Current Setup:** Taking a detailed look at the location of current data and all regulations that pertain to your organization is perhaps the best starting point for effectively classifying data.
- **Creating a Data Classification Policy:** Staying compliant with data protection principles in an organization is nearly impossible without proper policy. Creating a policy should be your top priority.
- **Prioritize and Organize Data:** Now that you have a policy and a picture of your current data, it's time to properly classify the data. Decide on the best way to tag your data based on its sensitivity and privacy.

Sub Unit – 9: SAMPLING AND ESTIMATION

Concept of Sampling: Sampling is the process of converting continuous signal to discrete form. Ex- Conversion of a sound wave into sequence of samples. A sample denotes a set of values at a point in time and/or space. A sampler is a subsystem that extracts samples from a continuous signal.

Concept of Estimation: The procedure of making judgment or decision about a population parameter is referred to as statistical estimation or simply estimation. Statistical estimation procedures provide estimates of population parameter with a desired degree of confidence.

Point Estimation: The objective of point estimation is to obtain a single number from the sample which will represent the unknown value of the population parameter.

Probability Sampling: A **probability sampling** scheme is one in which each unit in the population has a chance (greater than zero) of being selected in the sample, and this possibility can be accurately determined.

Simple Random Sampling: In a simple random sample ('SRS') of a given size, all such subsets of the frame are given an equal probability. Each component of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned. This minimizes bias and simplifies analysis of results.

Systematic sampling: Systematic sampling depend on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards.

Stratified Sampling: The sampling where the population embraces several distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.

Cluster Sampling: It is an example of 'two-stage sampling' or 'multistage sampling': in the first stage a sample of areas is chosen; in the second stage a sample of respondents within those areas is selected.

Multistage Sampling: Multistage sampling is a complex form of cluster sampling in which two or more levels of units are embedded one in the other. The first stage consists of constructing the clusters that will be used to sample from. In the second stage, a sample of primary units is randomly selected from each cluster (rather than using all units contained in all selected clusters). In following stages, in each of those selected clusters, additional samples of units are selected, and so on. All ultimate units (individuals, for instance) selected at the last step of this procedure are then surveyed.

Non-Probability Sampling: Non-probability sampling is any sampling technique where some elements of the population have no definite chance of selection, or where the probability of selection can't be correctly determined.

Non-probability sampling may be of the following types:

Quota Sampling

Convenience Sampling

What is a Sampling Distribution? A sampling distribution is a graph of a statistic for your sample data.

Central Limit Theorem: The **central limit theorem** states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough.

Standard Error: The Standard Error (SE) is very similar to standard deviation. Both are measures of spread. The higher the number, the more spread out your data is. To put it simply, the two terms are essentially equal — but there is one important difference. While the standard error uses **statistics** (sample data) standard deviations use **parameters** (population data). (What is the difference between a statistic and a parameter?).

Statistical Estimation: Statistical inference is the process of making judgment about a population based on sampling properties. An important aspect of statistical inference is using **estimates** to approximate the value of an unknown population parameter. Another type of inference involves choosing between two opposing views or statements about the population; this process is called **hypothesis testing**.

Confidence Intervals: Statisticians use a confidence interval to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

- A confidence level.
- A statistic.
- A margin of error.

Confidence Level: The probability part of a confidence interval is called a **confidence level**. The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

Margin of Error: In a confidence interval, the range of values above and below the sample statistic is called the **margin of error**.

Sub Unit – 10 HYPOTHESIS TESTING

A statistical hypothesis is an assumption about a population which may or may not be true. Hypothesis testing is a set of formal procedures used by statisticians to either accept or reject statistical hypotheses. Statistical hypotheses are of two types:

- **Null hypothesis, *H*0** - represents a hypothesis of chance basis.
- **Alternative hypothesis, *Ha*** - represents a hypothesis of observations which are influenced by some non-random cause.

Hypothesis Tests: Following formal process is used by statistician to determine whether to reject a null hypothesis, based on sample data. This process is called hypothesis testing and is consists of following **four steps:**

1. **State the hypotheses** - This step involves stating both null and alternative hypotheses. The hypotheses should be stated in such a way that they are mutually exclusive. If one is true then other must be false.

2. **Formulate an analysis plan** - The analysis plan is to describe how to use the sample data to evaluate the null hypothesis. The evaluation process focuses around a single test statistic.
3. **Analyze sample data** - Find the value of the test statistic (using properties like mean score, proportion, t statistic, z-score, etc.) stated in the analysis plan.
4. **Interpret results** - Apply the decisions stated in the analysis plan. If the value of the test statistic is very unlikely based on the null hypothesis, then reject the null hypothesis.

z – test: A ***Z*-test** is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Z-test tests the mean of a distribution in which we already know the population variance σ^2 .

t – test: The ***t*-test** is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the null hypothesis.

A *t*-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.

Correlated (or Paired) T-Test: The correlated t-test is performed when the samples typically consist of matched pairs of similar units, or when there are cases of repeated measures. For example, there may be instances of the same patients being tested repeatedly—before and after receiving a particular treatment. In such cases, each patient is being used as a control sample against themselves.

ANOVA: Analysis of variance (**ANOVA**) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample.

Chi-square test: The term "chi-squared test," also written as **χ^2 test**, refers to certain types of statistical hypothesis tests that are valid to perform when the test statistic is chi-squared distributed under the null hypothesis.

Mann-Whitney test (U- test): In statistics, the **Mann–Whitney *U* test** (also called the **Mann–Whitney–Wilcoxon** (**MWW**), **Wilcoxon rank-sum test**, or **Wilcoxon–Mann–Whitney test**) is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one population will be less than or greater than a randomly selected value from a second population.

This test can be used to investigate whether two *independent* samples were selected from populations having the same distribution. A similar nonparametric test used on *dependent* samples is the Wilcoxon signed-rank test.

Kruskal-Wallis test (H-test): The **Kruskal–Wallis test** by ranks, **Kruskal–Wallis *H* test** (named after William Kruskal and W. Allen Wallis), or **one-way ANOVA on ranks** is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes. It extends the Mann–Whitney *U* test, which is used for comparing only two

groups. The parametric equivalent of the Kruskal–Wallis test is the one-way analysis of variance (ANOVA).

Rank correlation test: A Spearman's Rank correlation test is a non-parametric measure of rank correlation. It is a statistical test used to determine the strength and direction of the association between two ranked variables.

Sub Unit – 11: REPORT WRITING

Report Writing: It is a formal style of writing elaborately on a topic. The tone of a report is always formal. The audience it is meant for is always thought out section. For example – report writing about a school event, report writing about a business case, etc.

Step in Report Writing: An effective report can be written going through the following steps-

1. Determine the objective of the report, i.e., identify the problem.
2. Collect the required material (facts) for the report.
3. Study and examine the facts gathered.
4. Plan the facts for the report.
5. Prepare an outline for the report, i.e., draft the report.
6. Edit the drafted report.
7. Distribute the draft report to the advisory team and ask for feedback and recommendations.

SECTION – 2: KEY STATEMENTS

Every candidates appearing for NET/SET examination should follow these key (main) points those can help them a better understanding regarding this unit very quickly.

Basic Key Statements: Central tendency (5.1.1), Mean (5.1.4), Median (5.1.4), Mode (5.1.4), Dispersion (5.2.1), Skewness (5.3), Co-efficient of skewness (5.3.3), Karl Pearson's Skewness (5.3.4), Correlation (5.4.1), Regression (5.4.6), Partial Regression (5.4.8), Simple regression (5.4.8), Free Hand Curve Method (5.4.10), Probability (5.5.1), Binomial Distribution (5.6.1), Poisson Distribution (5.6.2), Normal Distribution (5.6.3), Concept of Research (5.7.1), Research Designs (5.7.2), Data (5.8.1), Sampling (5.9.1), Estimation (5.9.1), Sampling Distribution (5.9.3), Statistical Hypothesis (5.10), Report Writing (5.11).

Standard Key Statements: Arithmetic Mean (5.1.4), Geometric Mean (5.1.4), Harmonic Mean (5.1.4), Standard deviation (5.2.1), Range (5.2.1), Median absolute deviation (5.2.1**),** Quartile Deviation (5.2.4), Mean Deviation (5.2.4), Bowley's Skewness (5.3.5), Coefficient of correlation (5.4.4), Regression Lines (5.4.9), Regression Coefficients (5.4.12), Bernoulli distribution (5.6.1), Basic Research (5.7.1), Applied Research (5.7.1), Survey Research (5.7.1), Descriptive Research (5.7.1), Correlational Research (5.7.1), Point Estimation (5.9.1), Interval Estimation (5.9.1), Simple Random Sampling (5.9.2), Systematic sampling (5.9.2), Non-Probability Sampling (5.9.2), Null hypothesis (5.10), Alternative hypothesis (5.10), Rank correlation test (5.10.7).

Advanced Key Statements: Weighted Harmonic Mean (5.1.4), Coefficient of variation (5.2.1), Variance (5.2.1), Coefficient of Quartile Deviation (5.2.4), Coefficient of Mean Deviation (5.2.4), Coefficient of Dispersion (5.2.5), *Linear regression* (5.4.8), Nonlinear Regression (5.4.8), Multiple Regression (5.4.8), Method of Least Squares (5.4.10), Bayes' Theorem (5.5.2), Qualitative Research (5.7.1), Quantitative Research **(5.7.1),** Stratified Sampling (5.9.2), Cluster Sampling (5.9.2), Multistage Sampling **(5.9.2),** Quota Sampling **(5.9.2),** Convenience Sampling (5.9.2), Central Limit Theorem (5.9.4), Standard Error (5.9.5), Statistical Estimation (5.9.6), Confidence Intervals (5.9.6), Confidence Level (5.9.6), Margin of Error (5.9.6), z – test (5.10.1), t – test (5.10.2), Correlated (or Paired) T-Test (5.10.2), ANOVA (5.10.3), Chi-square test (5.10.4), Mann-Whitney test (5.10.5), Kruskal-Wallis test (5.10.6).

[N.B. – Values in parenthesis are the reference number]

SECTION – 3: KEY FACTS AND FIGURES
Sub Unit - 1: MEASURES OF CENTRAL TENDENCY

5.1.1 Definition: Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. Along with the variability (dispersion) of a dataset, central tendency is a branch of descriptive statistics.

The central tendency is one of the most quintessential concepts in statistics. Although it does not provide information regarding the individual values in the dataset, it delivers a comprehensive summary of the whole dataset.

5.1.2 Objectives and Functions of Averages

- **Representative of the group:** An average represents all the features of a group; hence the results about the whole group can be deduced from it.
- **Brief description:** An average gives us simple and brief description of the main features of the whole data.
- **Helpful in comparison:** The measures of central tendency or averages reduce the data to a single value which is highly useful for making comparative studies. For example, comparing the per capita income of two countries, we can conclude that which country is richer.
- **Helpful in formulation of policies:** Averages help to develop a business in case of a firm or help the economy of a country to develop.
- **Base of other statistical Analysis:** Other statistical devices such as mean deviation, co-efficient of variation, co-relation, analysis of time series and index numbers are also based on the averages.

5.1.3 Characteristics or Essentials of a Good Average: The following are the main features of averages:

- **Simplicity:** The fundamental feature of the average is that it should be easy calculate and simple to follow.
- **Representation:** Average should represent the entire mass of data.
- **Rigidly Defined:** Averages should be rigidly defined. If it is so, instability in its value will be no more and would always be a definite figure.
- **Algebraic Treatment:** Averages are always capable of further algebraic treatment.
- **Clear and Stable Definition:** A good average should have a clear and stable definition.
- **Absolute Number:** A good average should be an absolute number.
- **Effect of fluctuations of Sampling:** A good average should not be affected by actuations of sampling. In other words, if different samples are taken from the production of rice, the mean of these samples should be equal.

- **Not affected by skewness:** A good average is one which is not affected by skewness in the distribution. Contrary to this, if it is affected by skewness, it cannot become a true representative.
- **Based on all values of a variable:** An average is said to be a true preventative only when it is based on all the values of a variable otherwise, it cannot considered a good average.
- **No Effect of Extreme values:** For a good average, it should not be unduly affected by extreme values. If it is so, it will not be a true representative.
- **Value can be found by Graphic Method:** A good average is one which can found by arithmetic as well as graphic method.
- **Possible to find control Tendency for open end class interval:** In many distributions ends are open. So, a good average is one which can be calculated even in en end class intervals.

5.1.4 Measures of Central Tendency: Generally, the central tendency of a dataset can be described using the following measures:

- **Mean (Average):** Represents the sum of all values in a dataset divided by the total number of the values.
- **Median:** The middle value in a dataset that is arranged in ascending order (from the smallest value to the largest value). If a dataset contains an even number of values, the median of the dataset is the mean of the two middle values.
- **Mode:** Defines the most frequently occurring value in a dataset. In some cases, a dataset may contain multiple modes while some datasets may not have any mode at all.
- Even though the measures above are the most commonly used to define central tendency, there are some other central tendency measures, including, but not limited to, geometric mean, harmonic mean, midrange, and geometric median.
- The selection of central tendency as a measure depends on the properties of a dataset. For instance, mode is the only central tendency measure of categorical data while a median works best with ordinal data.
- Although mean is regarded as the best measure of central tendency for quantitative data, it is not always the case. For example, mean may not work well with quantitative datasets that contain extremely large or extremely small values. The extreme values may distort the mean. Thus, you may consider other options of central tendency.
- The measures of central tendency can be found using a formula or definition. Also, they can be identified using a frequency distribution graph. Note that for the datasets that follow a normal distribution, the mean, median, and mode are located on the same spot on the graph.

- **Arithmetic Mean:** The arithmetic mean of a set of observed data is defined as being equal to the sum of the numerical values of each and every observation divided by the total number of observations.

Properties of Arithmetic Mean:

Property 1: If all the observations assumed by a variable are constants, say "k", then arithmetic mean is also "k".

For example, if the height of every student in a group of 10 students is 170 cm, the mean height is, of course 170 cm.

Property 2: The algebraic sum of deviations of a set of observations from their arithmetic mean is zero. That is, for unclassified data, $\sum(x - \bar{x}) = 0$.

And for a grouped frequency distribution, $\sum f(x - \bar{x}) = 0$.

For example, if a variable "x" assumes five observations, say 10, 20, 30, 40, 50, then $\bar{x} = 30$.

The deviations of the observations from arithmetic mean $(x - \bar{x})$ are -20, -10, 0, 10, 20.

Now, $\sum(x - \bar{x}) = (-20) + (-10) + 0 + 10 + 20 = 0$

Property 3: Arithmetic mean is affected due to a change of origin and/or scale which implies that if the original variable "x" is changed to another variable "y" effecting a change of origin, say "a" and scale, say "b", of "x". That is $y = a + bx$.

Property 4: If there are two groups containing n_1 and n_2 observations $\bar{x}_1$ and $\bar{x}_2$ are the respective arithmetic means, then the combined arithmetic mean is given by $\bar{x} = (n_1\bar{x}_1 + n_2\bar{x}_2) / (n_1 + n_2)$. This property could be extended to more than two groups and we may write it as $\bar{x} = \sum n\bar{x} / \sum n$

Here, $\sum n\bar{x} = n_1\bar{x}_1 + n_2\bar{x}_2 +$.....

$\sum n = n_1 + n_2 +$.....

Some Other Properties of Arithmetic Mean

1) It is rigidly defined.
2) It is based on all the observations.
3) It is easy to comprehend.
4) It is simple to calculate.
5) It is least affected by the presence of extreme observations.
6) It is amenable to mathematical treatment or properties.

- **Geometric Mean**

Introduction: A special type of average where we multiply the numbers together and then take a square root (for two numbers), cube root (for three numbers) etc.

In mathematics, the **geometric mean** is a mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the *n*th root of the product of *n* numbers,.

A geometric mean is often used when comparing different items—finding a single "figure of merit" for these items—when each item has multiple properties that have different numeric ranges.[3] For example, the geometric mean can give a meaningful value to compare two companies which are each rated at 0 to 5 for their environmental

sustainability, and are rated at 0 to 100 for their financial viability. If an arithmetic mean were used instead of a geometric mean, the financial viability would have greater weight because its numeric range is larger. That is, a small percentage change in the financial rating (e.g. going from 80 to 90) makes a much larger difference in the arithmetic mean than a large percentage change in environmental sustainability (e.g. going from 2 to 5). The use of a geometric mean normalizes the differently-ranged values, meaning a given percentage change in any of the properties has the same effect on the geometric mean. So, a 20% change in environmental sustainability from 4 to 4.8 has the same effect on the geometric mean as a 20% change in financial viability from 60 to 72.

The geometric mean applies only to positive numbers.[4] It is also often used for a set of numbers whose values are meant to be multiplied together or are exponential in nature, such as data on the growth of the human population or interest rates of a financial investment.

The geometric mean is also one of the three classical Pythagorean means, together with the aforementioned arithmetic mean and the harmonic mean. For all positive data sets containing at least one pair of unequal values, the harmonic mean is always the least of the three means, while the arithmetic mean is always the greatest of the three and the geometric mean is always in between (see Inequality of arithmetic and geometric means.)

- **Harmonic Mean**

In mathematics, the **harmonic mean** (sometimes called the **subcontrary mean**) is one of several kinds of average, and in particular, one of the Pythagorean means. Typically, it is appropriate for situations when the average of rates is desired.

The harmonic mean can be expressed as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations.

Harmonic mean is a type of average that is calculated by dividing the number of values in the data series by the sum of reciprocals (1/x_i) of each value in the data series. A harmonic mean is one of the three Pythagorean means (the other two are arithmetic mean and geometric mean). The harmonic mean always shows the lowest value among the Pythagorean means.

The harmonic mean is often used to calculate the average of the ratios or rates. It is the most appropriate measure for ratios and rates because it equalizes the weights of each data point. For instance, the arithmetic mean places a high weight to large data points, while geometric mean gives a lower weight to the smaller data points.

In finance, the harmonic mean is used to determine the average for financial multiples such as price-to-earnings (P/E) ratio. The financial multiples should not be averaged using the arithmetic mean because it is biased toward larger values. One of the most

common problems in finance that uses the harmonic mean is the calculation of the ratio of a portfolio that consists of several securities.

Formula for Harmonic Mean

The general formula for calculating a harmonic mean is:

Harmonic mean = n / (∑1/x_i)

Where:

- n – the number of the values in a dataset
- x_i – the point in a dataset

The weighted harmonic mean can be calculated using the following formula:

Weighted Harmonic Mean = (∑w_i) / (∑w_i/x_i)

Where:

- w_i – the weight of the data point
- x_i – the point in a dataset

Relationship among the Average:

In any distribution where the original items differ in size, then

either the values of A.M > G.M > H.M (or)

H.M < G.M < A.M in case all items are identical then A.M. = G.M = H.M

Median: In statistics and probability theory, the **median** is the value separating the higher half from the lower half of a data sample, a population or a probability distribution. For a data set, it may be thought of as the "middle" value. For example, the basic advantage of the median in describing data compared to the mean (often simply described as the "average") is that it is not skewed so much by a small proportion of extremely large or small values, and so it may give a better idea of a "typical" value. For example, in understanding statistics like household income or assets, which vary greatly, the mean may be skewed by a small number of extremely high or low values. Median income, for example, may be a better way to suggest what a "typical" income is. Because of this, the median is of central importance in robust statistics, as it is the most resistant statistic, having a breakdown point of 50%: so long as no more than half the data are contaminated, the median will not give an arbitrarily large or small result.

Use of Median: The median can be used as a measure of location when one attaches reduced importance to extreme values, typically because a distribution is skewed, extreme values are not known, or outliers are untrustworthy, i.e., may be measurement/transcription errors.

As a median is based on the middle data in a set, it is not necessary to know the value of extreme results in order to calculate it. For example, in a psychology test investigating the time needed to solve a problem, if a small number of people failed to solve the problem at all in the given time a median can still be calculated.

The median is simple to understand and easy to calculate, while also a robust approximation to the mean, the median is a popular summary statistic in descriptive statistics. In this context, there are several choices for a measure of variability: the range, the interquartile range, the mean absolute deviation, and the median absolute deviation.

For practical purposes, different measures of location and dispersion are often compared on the basis of how well the corresponding population values can be estimated from a sample of data. The median, estimated using the sample median, has good properties in this regard. While it is not usually optimal if a given population distribution is assumed, its properties are always reasonably good. For example, a comparison of the efficiency of candidate estimators shows that the sample mean is more statistically efficient when — and only when — data is uncontaminated by data from heavy-tailed distributions or from mixtures of distributions.[*citation needed*] Even then, the median has a 64% efficiency compared to the minimum-variance mean (for large normal samples), which is to say the variance of the median will be ~50% greater than the variance of the mean.

Mode: The mode is the number that appears most frequently in a set. A set of numbers may have one mode, more than one mode, or no mode at all. Other popular measures of central tendency include the mean, or the average (mean) of a set, and the median, the middle value in a set.

Understanding the Mode

In statistics, data are distributed in various ways. The most often cited distribution is the classic normal (bell-curve) distribution. In this, and some other distributions, the mean (average) value falls at the mid-point, which is also the peak frequency of observed values. For such a distribution, this value is also the mode—the most frequently occurring value in the data.

In other distributions, the most frequent value may differ from the modal value. For instance, the average frequency of people born with six fingers is around 0.2%, but the mode is zero since the most common outcome is five fingers.

Key Takeaways

- In statistics, the mode is the most commonly observed value in a set of data.
- For the normal distribution, the mode is also the same value as the mean and median.
- In many cases, the modal value will differ from the average value in the data.

Examples of the Mode

For example, in the following list of numbers, 16 is the mode since it appears more times in the set than any other number:

- 3, 3, 6, 9, **16, 16, 16**, 27, 27, 37, 48

A set of numbers can have more than one mode (this is known as *bimodal* if there are two modes) if there are multiple numbers that occur with equal frequency, and more times than the others in the set.

- **3, 3, 3**, 9, **16, 16, 16**, 27, 37, 48

In the above example, both the number 3 and the number 16 are modes as they each occur three times and no other number occurs more often.

If no number in a set of numbers occurs more than once, that set has no mode:

- 3, 6, 9, 16, 27, 37, 48

A set of numbers with two modes is **bimodal**, a set of numbers with three modes is **trimodal**, and a set of numbers with four or more nodes is **multimodal**.

Advantages and Disadvantages of the Mode

Advantages:

- The mode is easy to understand and calculate.
- The mode is not affected by extreme values.
- The mode is easy to identify in un-grouped data and discrete frequency distribution.
- The mode is useful for qualitative data.
- The mode can be computed in an open-ended frequency table.
- The mode can be located graphically.

Disadvantages:

- The mode is not well defined.
- The mode is not based on all values.
- The mode is stable for large values and will not be well defined if the data consist of a small number of values.
- The mode is not capable of further mathematical treatment.
- Sometimes data have one mode, more than one mode, or no mode at all.
- Empirical Relationship between Mean, Median and Mode
- In case of a moderately skewed distribution, the difference between mean and mode is almost equal to three times the difference between the mean and median. Thus, the empirical mean median mode relation is given as:

Mean – Mode = 3 (Mean – Median)

Mean Median Mode Relation With Frequency Distribution

- **Frequency Distribution with Symmetrical Frequency Curve**

If a frequency distribution graph is having a symmetrical frequency curve, the mean, median, and mode will be equal.

- **For Positively Skewed Frequency Distribution**

In case of a positively skewed frequency distribution, the mean is always greater than median and the median is always greater than the mode.

- **For Negatively Skewed Frequency Distribution**

In case of a negatively skewed frequency distribution, the mean is always lesser than median and the median is always lesser than the mode.

Sub Unit - 2: MEASURES OF DISERSION

5.2.1 Introduction: In statistics, **dispersion** (also called **variability**, **scatter**, or **spread**) is the extent to which a distribution is stretched or squeezed.[1] Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range.

Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions.

A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse.

Most measures of dispersion have the same units as the quantity being measured. In other words, if the measurements are in metres or seconds, so is the measure of dispersion. Examples of dispersion measures include:

- Standard deviation
- Interquartile range (IQR)
- Range
- Mean absolute difference (also known as Gini mean absolute difference)
- Median absolute deviation (MAD)
- Average absolute deviation (or simply called average deviation)
- Distance standard deviation

Other measures of dispersion are **dimensionless**. In other words, they have no units even if the variable itself has units. These include:

- Coefficient of variation
- Quartile coefficient of dispersion
- Relative mean difference, equal to twice the Gini coefficient
- Entropy: While the entropy of a discrete variable is location-invariant and scale-independent, and therefore not a measure of dispersion in the above sense, the entropy of a continuous variable is location invariant and additive in scale: If Hz is the entropy of continuous variable z and $y=ax+b$, then $Hy=Hx+log(a)$.

There are other measures of dispersion:

- Variance (the square of the standard deviation) – location-invariant but not linear in scale.
- Variance-to-mean ratio – mostly used for count data when the term coefficient of dispersion is used and when this ratio is dimensionless, as count data are themselves dimensionless, not otherwise.

5.2.2 Importance of Dispersion:

- **Measures of dispersion supplement the information given by the measures of central tendency:** Measures of dispersion are also called averages of the 'second order i,e., second time averaging the deviations from a measure of central tendency. It affords an estimate of the phenomena to which the given (original) data relate. This will increase the accuracy of statistical analysis and interpretation and we can be in a position to draw more dependable inferences.
- **Measures of dispersion make possible comparison between different groups:** If the original data is expressed in different units, comparisons will not be possible. But with the help of relative measures of dispersion, all such comparisons can be easily made. Accurate and dependable comparison between the variability of two series will lead to dependable and accurate results.
- **Measures of dispersion are very important in many social problems:** Social problems of different area of the country can be compared with different areas and then these social evils can be removed by taking effective steps.
- **Measures of dispersion serve as a useful check on drawing wrong conclusions from the comparison of averages or measures of central tendency:** The arithmetic mean may be the same of two different groups but it will not reveal about the prosperity of one group and backwardness of other. This type of internal make-up can be known by the study of dispersion.

Thus with the help of the study of dispersion we will not conclude that both the groups are similar. We may find that one group is prosperous and the other is backward by knowing the amount of variability around the measures of central tendency.

5.2.3 Properties or Features of a good Measure of Dispersion:

1. In should be capable of treating it by Algebraic or Statistical techniques.
2. It should be easy to calculate i.e. by simple methods.
3. It should be easy to understand i.e. Even a layman must understand about its message or what it demonstrates.
4. It must not be affected by different samples or fluctuation of sampling. Every sample should give same type of information.
5. The quality and quantity of each term must affect it. As in median, last value may be 15 or 15000 in the series 3, 5, 7, 9, 15 (15000); does not effect at all.

5.2.4. Types of Measures of Dispersion:

- **Absolute and Relative Measures:**
  Absolute measures of Dispersion are expressed in same units in which original data is presented but these measures cannot be used to compare the variations between the two series. Relative measures are not expressed in units but it is a pure number. It is the ratio of absolute dispersion to an appropriate average such as co-efficient of Standard Deviation or Co-efficient of Mean Deviation.
- **Absolute Measure of Dispersion**
  An absolute measure of dispersion contains the same unit as the original data set. Absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, standard deviation, quartile deviation, etc.

The types of absolute measures of dispersion are:

- **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6
- **Variance:** Deduct the mean from each data in the set then squaring each of them and adding each square and finally dividing them by the total no of values in the data set is the variance. Variance $(\sigma^2)=\sum(X-\mu)^2/N$
- **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. = $\sqrt{\sigma}$.
- **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
- **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation.

Relative Measure of Dispersion:

The relative measures of depression are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

- **Coefficient of Range:** It is defined as the relative measure of the distribution based on the range of any given data set, which is the difference between the maximum and minimum value in the given set. It is also known as range coefficient. In the case of grouped data, the range is the difference between the upper boundary of the highest class and the lower boundary of the lowest class. It is also calculated by using the difference between the mid points of the highest class and the lowest class.

- **Coefficient of Variation:** In probability theory and statistics, the **coefficient of variation** (**CV**), also known as **relative standard deviation** (**RSD**), is a standardized measure of dispersion of a probability distribution or frequency distribution.
- **Coefficient of Standard Deviation**
- **Coefficient of Quartile Deviation:** The Quartile Deviation is a simple way to estimate the spread of a distribution about a measure of its central tendency (usually the mean). So, it gives you an idea about the range within which the central 50% of your sample data lies. Consequently, based on the quartile deviation, the Coefficient of Quartile Deviation can be defined, which makes it easy to compare the spread of two or more different distributions.
- **Coefficient of Mean Deviation:** It is calculated to compare the data of two series. The coefficient of mean deviation is calcvilated by dividing mean deviation by the average. If deviations are taken from mean, we divide it by mean, if the deviations are taken from median, then it is divided by mode and if the "deviations are taken from median, then we divide mean deviation by median.

5.2.5 Coefficient of Dispersion: The coefficients of dispersion are calculated along with the measure of dispersion when two series are compared which differ widely in their averages. The dispersion coefficient is also used when two series with different measurement unit are compared. It is denoted as C.D.

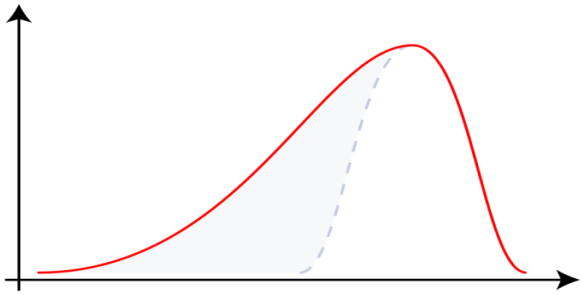
Sub Unit - 3: MEASURES OF SKEWNESS

5.3.1 Introduction: In probability theory and statistics, **skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined.

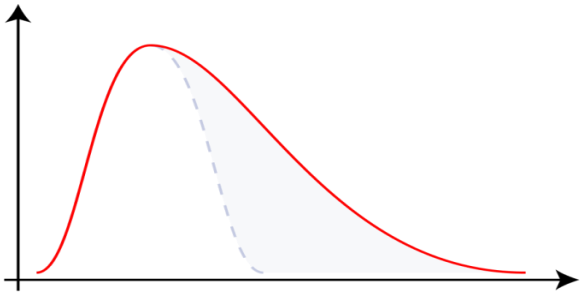
For a unimodal distribution, negative skew commonly indicates that the *tail* is on the left side of the distribution, and positive skew indicates that the tail is on the right. In cases where one tail is long but the other tail is fat, skewness does not obey a simple rule. For example, a zero value means that the tails on both sides of the mean balance out overall; this is the case for a symmetric distribution, but can also be true for an asymmetric distribution where one tail is long and thin, and the other is short but fat.

Consider the two distributions in the figure just below. Within each graph, the values on the right side of the distribution taper differently from the values on the left side. These tapering sides are called *tails*, and they provide a visual means to determine which of the two kinds of skewness a distribution has:

1. ***Negative skew*:** The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be *left-skewed*, *left-tailed*, or *skewed to the left*, despite the fact that the curve itself appears to be skewed or leaning to the right; *left* instead refers to the left tail being drawn out and, often, the mean being skewed to the left of a typical center of the data. A left-skewed distribution usually appears as a *right-leaning* curve.
2. ***Positive skew*:** The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be *right-skewed*, *right-tailed*, or *skewed to the right*, despite the fact that the curve itself appears to be skewed or leaning to the left; *right* instead refers to the right tail being drawn out and, often, the mean being skewed to the right of a typical center of the data. A right-skewed distribution usually appears as a *left-leaning* curve.



Negative Skew



Positive Skew

5.3.2 Absolute skewness: This is obtained by finding the difference between any two measures of dispersion viz: Mean, Median and Mode. Thus, Skewness or

$$Sk = |X - M| \text{ or } |X - Z| \text{ or } |M - Z|$$

Any positive value obtained by any of the above formulae is marked as the extent of the positive skewness. Any negative value obtained by any of the above formulae is marked as the extent of the negative skewness of the distribution. If the result produced is zero, it signifies the absence of skewness in the distribution.

5.3.3 Co-efficient of skewness: This is obtained by dividing the Skewness by any measure of dispersion.

5.3.4 Karl Pearson's Skewness: Prof. Karl pearson says that to study the skewness of a series, the difference between the Mean and Mode only should be found out. This is because, Mean is an average which is affected very much by the extreme values of a series and Mode is an average which is least affected by the extreme values of a series. Thus, according to him,

$Sk_{(p)}$ = Mean –Mode

When, Mode is ill defined i.e. when it has different values, Prof. Pearson proposes to find out the skewness by the following formula:

$Sk_{(p)} = 3$ (Mean –Median)

This formula is based on the empirical relationship between Mean, Median and Mode which is as follows:

Absolute Mode= 3 Median – 2 Mean

For finding the coefficient of skewness, Prof. Pearson advocates that only standard diviation should be taken as the divisor of the absolute skewness. This is because, standard deviation is the only measure of dispersion which possesses many algebraic properties, and other measures of dispersion are not capable of algebraic treatments.

5.3.5 Bowley's Skewness: According to Prof. A.L. Bowley, the presence, or absence of skewness will be determined on the basis of the distance of the quartiles from the Median. Thus, his skewness is given by

$$SK(B) = (Q_3 - M) - (M - Q_1)$$

$$\text{Or} \quad = Q_3 + Q_1 - 2M$$

If the above equation results in zero, it will indicate the absence of skewness or symmetricity of the distribution. On the other hand, if the said equation results in some positive, or negative figure, the same will be marked as the extent of the positive, or negative skewness of the series respectively.

5.3.6 Kelley's Skewness: Prof. Kelley says that the above measure of skewness as propounded by Prof. Bowely ignores the first 25% of the data by measuring the skewness on the basis of the distance of the two extreme Quartiles form Median. To minimize the magnitude of such avoidance of the data, he advocates that the skewness should be measured on the basis of the distance of the two extreme Deciles, or Percentiles from the Median. Thus, according to him, the skewness is given by:

$SK\ (K) = (D_9 - M) - (M - D_1)$

$= D_9 + D_1 - 2M$

$= (P_{90} - M) - (M - P_{10})$

$= P_{90} + P_{10} - 2M$

Sub Unit - 4: CORRELATION AND REGRESSION OF TWO VARIABLES

CORRELATION:

5.4.1 Introduction: Correlation, in the finance and investment industries, is a statistic that measures the degree to which two securities move in relation to each other. Correlations are used in advanced portfolio management, computed as the correlation coefficient, which has a value that must fall between -1.0 and +1.0.

A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one security moves, either up or down, the other security moves in lockstep, in the same direction. A perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no relationship at all.

Co-efficient of correlation is a numerical index that tells us to what extent the two variables are related and to what extent the variations in one variable changes with the variations in the other. The co-efficient of correlation is always symbolized either by r or ρ (Rho).

The notion 'r' is known as product moment correlation co-efficient or Karl Pearson's Coefficient of Correlation. The symbol 'ρ' (Rho) is known as Rank Difference Correlation coefficient or spearman's Rank Correlation Coefficient.

The size of '*r*' indicates the amount (or degree or extent) of correlation-ship between two variables. If the correlation is positive the value of '*r*' is + ve and if the correlation is negative the value of V is negative. Thus, the signs of the coefficient indicate the kind of relationship. The value of V varies from +1 to -1.

Correlation can vary in between perfect positive correlation and perfect negative correlation. The top of the scale will indicate perfect positive correlation and it will begin from +1 and then it will pass through zero, indicating entire absence of correlation.

The bottom of the scale will end at -1 and it will indicate perfect negative correlation. Thus numerical measurement of the correlation is provided by the scale which runs from +1 to -1.

[NB—The coefficient of correlation is a number and not a percentage. It is generally rounded up to two decimal places].

(i) Finding characteristics of psychological and educational tests (reliability, validity, item analysis, etc.).

(ii) Testing whether certain data is consistent with hypothesis.

(iii) Predicting one variable on the basis of the knowledge of the other(s).

(iv) Building psychological and educational models and theories.

(v) Grouping variables/measures for parsimonious interpretation of data.

(vi) Carrying multivariate statistical tests (Hoteling's T^2; MANOVA, MANCOVA, Discriminant analysis, Factor Analysis).

(vii) Isolating influence of variables.

5.4.2 Need for Correlation:

(i) Finding characteristics of psychological and educational tests (reliability, validity, item analysis, etc.).

(ii) Testing whether certain data is consistent with hypothesis.

(iii) Predicting one variable on the basis of the knowledge of the other(s).

(iv) Building psychological and educational models and theories.

5.4.3 Types of Correlation:

In a bivariate distribution, the correlation may be:

1. Positive, Negative and Zero Correlation; and
2. Linear or Curvilinear (Non-linear).

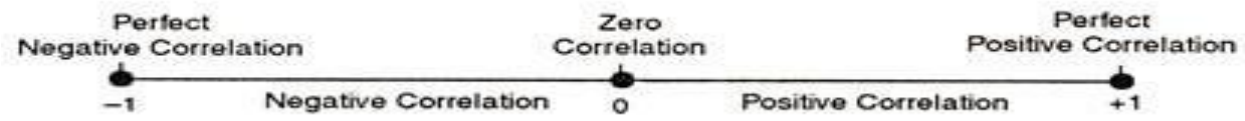
1. Positive, Negative or Zero Correlation: When the increase in one variable (X) is followed by a corresponding increase in the other variable (Y); the correlation is said to be positive correlation. The positive correlations range from 0 to +1; the upper limit i.e. +1 is the perfect positive coefficient of correlation.

The perfect positive correlation specifies that, for every unit increase in one variable, there is proportional increase in the other. For example "Heat" and "Temperature" have a perfect positive correlation.

If, on the other hand, the increase in one variable (X) results in a corresponding decreases in the other variable (Y), the correlation is said to be negative correlation.

The negative correlation ranges from 0 to – 1; the lower limit giving the perfect negative correlation. The perfect negative correlation indicates that for every unit increase in one variable, there is proportional unit decrease in the other.

Zero correlation means no relationship between the two variables X and Y; i.e. the change in one variable (X) is not associated with the change in the other variable (Y). For example, body weight and intelligence, shoe size and monthly salary; etc. The zero correlation is the mid-point of the range – 1 to + 1.



2. Linear or Curvilinear Correlation: Linear correlation is the ratio of change between the two variables either in the same direction or opposite direction and the graphical representation of the one variable with respect to other variable is straight line.

Consider another situation. First, with increase of one variable, the second variable increases proportionately upto some point; after that with an increase in the first variable the second variable starts decreasing.

The graphical representation of the two variables will be a curved line. Such a relationship between the two variables is termed as the curvilinear correlation.

5.4.4 Methods of Computing Co-Efficient of Correlation: In case of ungrouped data of bivariate distribution, the following three methods are used to compute the value of co-efficient of correlation:

1. Scatter diagram method.
2. Pearson's Product Moment Co-efficient of Correlation.
3. Spearman's Rank Order Co-efficient of Correlation.

1. Scatter Diagram Method: Scatter diagram or dot diagram is a graphic device for drawing certain conclusions about the correlation between two variables.

In preparing a scatter diagram, the observed pairs of observations are plotted by dots on a graph paper in a two dimensional space by taking the measurements on variable X along the horizontal axis and that on variable Y along the vertical axis.

The placement of these dots on the graph reveals the change in the variable as to whether they change in the same or in the opposite directions. It is a very easy, simple but rough method of computing correlation.

The frequencies or points are plotted on a graph by taking convenient scales for the two series. The plotted points will tend to concentrate in a band of greater or smaller width according to its degree. 'The line of best fit' is drawn with a free hand and its direction indicates the nature of correlation. Scatter diagrams, as an example, showing various degrees of correlation are shown in Fig. 5.1 and Fig. 5.2.

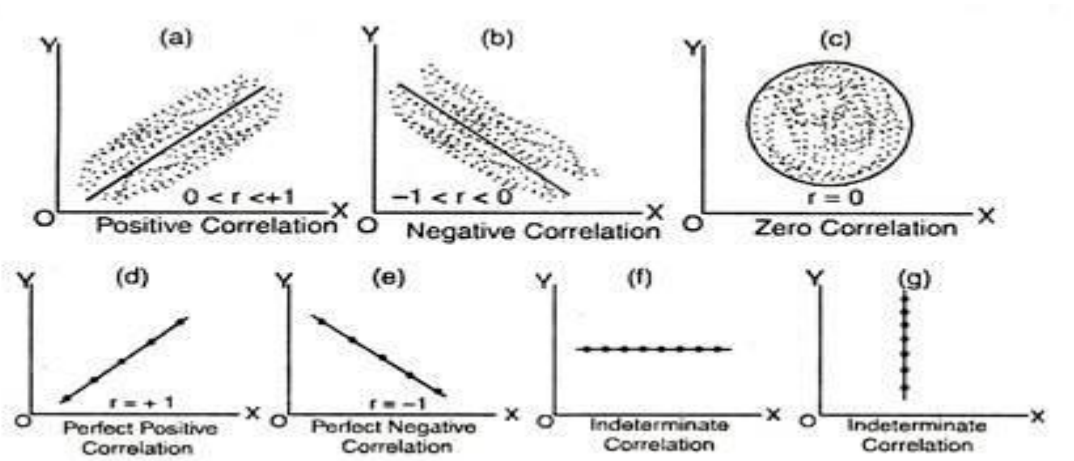


Fig. 5.1 Scatter Diagrams Showing Varying Degree of Relationship between X and Y.

Fig. 5.2 Scatter Diagram illustrating Linear and Curvilinear relationships.

If the line goes upward and this upward movement is from left to right it will show positive correlation. Similarly, if the lines move downward and its direction is from left to right, it will show negative correlation.

The degree of slope will indicate the degree of correlation. If the plotted points are scattered widely it will show absence of correlation. This method simply describes the 'fact' that correlation is positive or negative.

2. Pearson's Product Moment Co-efficient of Correlation: The coefficient of correlation, r, is often called the "Pearson r" after Professor Karl Pearson who developed the product-moment method, following the earlier work of Gallon and Bravais.

Coefficient of correlation as ratio: The product-moment coefficient of correlation may be thought of essentially as that ratio which expresses the extent to which changes in one variable are accompanied by—or dependent upon-changes in a second variable.

As an illustration, consider the following simple example which gives the paired heights and weights of five college students:

Students :	A	B	C	D	E
Height in inches :	72	69	66	70	68
Weight in Lbs. :	170	165	150	180	185

(1) Studetns	(2) HT in inches X	(3) WT in Lbs Y	(4) x	(5) y	(6) xy	(7) $\frac{x}{\sigma_x}$	(8) $\frac{y}{\sigma_y}$	(9) $\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}$
A	72	170	3	0	0	1.34	.00	.00
B	69	165	0	− 5	0	.00	− .37	.00
C	66	150	− 3	− 20	60	−1.34	− 1.46	1.96
D	70	180	1	10	10	.45	.73	.33
E	68	185	− 1	15	− 15	− .45	1.10	− .49
					$\Sigma xy = 55$			$\Sigma\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right) = 1.80$

$M_X = 69$ in. $\sigma_x = 2.24$ in.

$M_Y = 170$ lbs. $\sigma_y = 13.69$ lbs. Correlation $= \frac{\Sigma\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right)}{N} = \frac{1.80}{5} = .36$

The mean height is 69 inches, the mean weight 170 pounds, and the o is 2.24 inches and o is 13.69 pounds, respectively. In the column (4) the deviation (x) of each student's height from the mean height, and in column (5) the deviation, (y) of each student's weight from the mean weight are given. The product of these paired deviations (xy) in column (6) is a measure of the agreement between individual heights and weights. The larger the sum of xy column the higher the degree of correspondence. In above example the value of ∑xy/N is 55/5 or 11. Where perfect agreement, i.e. r = ± 1.00, the value of ∑*xy*/N exceeds maximum limit.

> ***The sum of the deviations from the mean (raised to some power) and divided by N is called a "moment". When corresponding deviations in x and y are multiplied together, summed, and divided by N*** $\left(\text{to give } \frac{\Sigma xy}{N}\right)$ ***the term "product-moment" is used.***

Thus, ∑*xy*/N would not yield a suitable measure of relationship between x and y. The reason is that such an average is not a stable measure, as it is not independent of the units in which height and weight have been expressed.

In consequence, this ratio will vary if centimeters and kilograms are employed instead of inches and pounds. One way to avoid the trouble-some matter of differences in units is to express each deviation as a σ score or standard score or Z score, i.e. to divide each x and y by its own σ.

Each x and y deviation is then expressed as a ratio, and is a pure number, independent of the test units. The sum of the products of the σ scores column (9) divided by N yields a ratio which is a stable expression of relationship. This ratio is the "product-moment" coefficient of correlation. In our example, its value of .36 indicates a fairly high positive correlation between height and weight in this small sample.

The student should note that our ratio or coefficient is simply the average product of the σ scores of corresponding X and Y measures i.e.

$r_{xy} = \frac{\Sigma Z_x Z_y}{N}$.

Thus, the quotient is $r = \frac{\Sigma\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right)}{N}$ ...(26)

When this ratio is written $\frac{\Sigma xy}{N \sigma_x \sigma_y}$ it becomes the well-known expression for r, the product-moment coefficient of correlation.

Correlation Coefficient (r_{xy}) :

$$r_{xy} = \frac{\text{Covariance (X, Y)}}{\sqrt{\text{Var. (X)} \cdot \text{Var (Y)}}} = \frac{S_{XY}}{S_X S_Y}$$

$$= \frac{\Sigma xy/n}{\sqrt{\frac{\Sigma x^2}{n} \times \frac{\Sigma y^2}{n}}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{\Sigma xy}{N S_X S_Y} = \frac{\Sigma Z_x Z_y}{n}$$

In raw score form $r_{xy} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$

Nature of r_{xy}:

(i) r_{xy} is a product moment r $\left(r_{xy} = \frac{\Sigma Z_x Z_y}{n}\right)$.
(ii) r_{xy} is a ratio, = r_{xy}.
(iii) r_{xy} can be + ve or – ve bound by limits – 1.00 to + 1.00.
(iv) r_{xy} may be regarded as an arithmetic mean (r_{xy} is the mean of standard score products).
(v) r_{xy} is not affected by any linear transformation of scores on either X or Y or both.
(vi) When variables are in the standard score form, r gives a measure of the average amount of change in one variable associated with the change of one unit the other variable.
(vii) $r_{xy} = \sqrt{b_{yx}\, b_{xy}}$ where b_{yx} = regression coefficient of Y on X, b_{xy} = regression coefficient of X on Y. r_{xy} = square root of the slopes of the regression lines.
(viii) r_{xy} is not influenced by the magnitude of means (scores are always relative).
(ix) r_{xy} cannot be computed if one of the variables has no variance S^2x or $S^2Y = 0$
(x) r_{xy} of 60 implies the same magnitude of relationship as $r_{xy} = – .60$. The sign tells about the direction of relationship, and the magnitude about the strength of the relationship.
(xi) df for r_{xy} is N – 2, which is used for testing significance of r_{xy}. Testing significance of r is testing significance of regression. Regression line involves slope and intercept, hence 2 *df* is lost. So when N = 2, r_{xy} is either + 1.00 or – 1.00 as there is no freedom for sampling variation in the numerical value of r.

A. Computation of r_{xy} (Ungrouped Data):

Here, using the formula for computation of r depends on "where from the deviations are taken". In different situations deviations can be taken either from actual mean or from zero or from A.M. Type of Formula conveniently applied for the calculation of coefficient correlation depends upon mean value (either in fraction or whole).

(i) The Formula of r when Deviations are taken from Means of the Two Distributions X and Y.

$$r_{xy} = \frac{\Sigma xy}{N \sigma_x \sigma_y} \quad \ldots (27)$$

where r_{xy} = Correlation between X and Y

x = deviation of any X score from the mean in the test X

y = deviation of corresponding Y score from the mean in test Y.

$\sum xy$ = Sum of all the products of deviations (X and Y)

σ_x and σ_y = Standard deviations of the distribution of X and Y score.

If we write $\sqrt{\frac{\Sigma x^2}{N}}$ for σ_x, and $\sqrt{\frac{\Sigma y^2}{N}}$ for σ_y, the N's cancel and formula becomes

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \quad \ldots (28)$$

in which x and y are deviations from the actual means and $\sum x^2$ and $\sum y^2$ are the sums of squared deviations in x and y taken from the two means.

This formula is preferred:

i. When mean values of both the variables are not in fraction.

ii. When to find out correlation between short, ungrouped series (say, twenty- five cases or so).

iii. When deviations are to be taken from actual means of the two distributions.

The steps necessary are illustrated in Table 5.1. They are enumerated here:

Step 1: List in parallel columns the paired X and Y scores, making sure that corresponding scores are together.

Step 2: Determine the two means M_x and M_y. In table 5.1, these are 7.5 and 8.0, respectively.

Step 3: Determine for every pair of scores the two deviations x and y. Check them by finding algebraic sums, which should be zero.

Step 4: Square all the deviations, and list in two columns. This is for the purpose of computing σ_x and σ_y.

Step 5: Sum the squares of the deviations to obtain $\sum x^2$ and $\sum y^2$ Find xy product and sum these for $\sum xy$.

Step 6: From these values compute σ_x and σ_y.

Table 5.1 Computation of *r* when deviations are taken from Means

X	Y	x	y	x^2	y^2	xy
13	11	+ 5.5	+ 3	30.25	9	+ 16.5
12	14	+ 4.5	+ 6	20.25	36	+ 27.0
10	11	+ 2.5	+ 3	6.25	9	+ 7.5
10	7	+ 2.5	− 1	6.25	1	− 2.5
8	9	+ 0.5	+ 1	0.25	1	+ 0.5
6	11	− 1.5	+3	2.25	9	− 4.5
6	3	− 1.5	− 5	2.25	25	+ 7.5
5	7	−2.5	−1	6.25	1	+ 2.5
3	6	−4.5	−2	20.25	4	+ 9.0
2	1	−5.5	−7	30.25	49	+ 38.5
$\Sigma X = 75$ $M_X = 7.5$	$\Sigma Y = 80$ $M_Y = 8.0$	$\Sigma x = 0.0$	$\Sigma y = 0.0$	$\Sigma x^2 = 124.50$	$\Sigma y^2 = 144$	$\Sigma xy = 102.0$

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{124.50}{10}} = \sqrt{12.450} = 3.528$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{144}{10}} = \sqrt{14.4} = 3.795$$

Applying Formula (27) $r_{xy} = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{102.0}{(10)\,(3.528)\,(3.795)} = \frac{102.0}{133.90} = +.76$

An alternative and shorter solution: There is an alternative and shorter route that omits the computation of σ_x and σ_y, should they not be needed for any other purpose.

Applying Formula (28):

$$r_{xy} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)\,(\Sigma y^2)}} \qquad \text{(Alternative formula for a Pearson } r\text{)}$$

$$= \frac{102.0}{\sqrt{(124.5)\,(144)}} = \frac{102.0}{\sqrt{17,928.0}} = \frac{102.0}{133.90} = +.76$$

(ii) The Calculation of r_{xy} from Original scores or Raw scores: It is an another procedure with ungrouped data, which does not require the use of deviations. It deals entirely with original scores. The formula may look forbidding but is really easy to apply.

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)\,(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2]\,[N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad \ldots(29)$$

This formula is preferred:

i. When to compute r from direct raw scores.

ii. Original scores ft. when data are small ungrouped.

iii. When mean values are in fractions.

iv. When good calculating machine is available.

X and Y are original scores in variables X and Y. Other symbols tell what is done with them.

We follow the steps that are illustrated in Table 5.2:

Table 5.2 Computation of r from Original Scores

X	Y	X^2	Y^2	XY
13	7	169	49	91
12	11	144	121	132
10	3	100	9	30
8	7	64	49	56
7	2	49	4	14
6	12	36	144	72
6	6	36	36	36
4	2	16	4	8
3	9	9	81	27
1	6	1	36	6
$\Sigma X = 70$	$\Sigma Y = 65$	$\Sigma X^2 = 624$	$\Sigma Y^2 = 533$	$\Sigma XY = 472$

Step 1: Square all X and Y measurements.
Step 2: Find the XY product for every pair of scores.
Step 3: Sum the X's, the Y's, the X^2, the Y^2, and the XY.
Step 4: Apply formula (29):

$$\gamma_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{(10 \times 472) - (70 \times 65)}{\sqrt{(10 \times 624 - 4{,}900)(10 \times 533 - 4{,}225)}} = \frac{170}{\sqrt{1{,}340 \times 1{,}105}}$$

$$= \frac{170}{\sqrt{1{,}480{,}700}} = \frac{170}{1216.84} = +.14$$

(ii) Computation of r_{xy} when deviations are taken from Assumed Mean: Formula (28) is useful in calculating r directly from two ungrouped series of scores, but it has the disadvantages as it requires "long method" of calculating means and σ's. The deviations x and y when taken from actual means are usually decimals and the multiplication and squaring of these values is often a tedious task.

For this reason—even when working with short ungrouped series—it is often easier to assume means, calculate deviations from these A.M.'s and apply the formula (30).

$$r_{xy} = \frac{\frac{\Sigma x'y'}{N} - C_x C_y}{\sigma'_x \sigma'_y} \qquad \ldots(30)$$

This formula is preferred:

i. When actual means are usually decimals and the multiplication and squaring of these values is often a tedious task.

ii. When deviations are taken from A.M.'s.

iii. When we are to avoid fractions.

The steps in computing r may be outlined as follows:

Step 1: Find the mean of Test 1 (X) and the mean of Test 2 (Y). The means as shown in Table 5.3 $M_X = 62.5$ and $M_Y = 30.4$ respectively.

Step 2: Choose A.M.'s of both X and Y i.e. $A.M._X$ as 60.0 and $A.M._Y$ as 30.0.

Step 3: Find the deviation of each score on Test 1 from its A.M., 60.0, and enter it in column x'. Next find the deviation of each score in Test 2 from its A.M., 30.0, and enter it in column y'.

Step 4: Square all of the x' and all of they' and enter these squares in column x'^2 and y'^2, respectively. Total these columns to obtain $\sum x'^2$ and $\sum y'^2$.

Step 5: Multiply x' and y', and enter these products (with due regard for sign) in the x'y' column. Total x'y' column, taking account of signs, to get $\sum x'y'$.

Step 6: The corrections, C_x and C_y, are found by subtracting AM_X from M_x and AM_y from M_y. Then, C_x found as 2.5 (62.5 – 60.0) and C_y as .4 (30.4 – 30.0).

Step 7: Substitute for $\sum x'y'$, 334, for $\sum x'^2$, 670 and for $\sum y'^2$, 285 in formula (30), as shown in Table 5.3, and solve for r_{xy}.

Table 5.3 Computation of r_{xy} when deviations are from A.M.

Subject	X	Y	x'	y'	x'^2	y'^2	$x'y'$
A	50	22	– 10	– 8	100	64	80
B	54	25	– 6	– 5	36	25	30
C	56	34	– 4	4	16	16	– 16
D	59	28	– 1	– 2	1	4	2
E	60	26	0	– 4	0	16	0
F	62	30	2	0	4	0	0
G	61	32	1	2	1	4	2
H	65	30	5	0	25	0	0
I	67	28	7	– 2	49	4	– 14
J	71	34	11	4	121	16	44
K	71	36	11	6	121	36	66
L	74	40	14	10	196	100	140
	$\Sigma X = 750$	$\Sigma Y = 365$			$\Sigma x'^2 = 670$	$\Sigma y'^2 = 285$	$\Sigma x'y' = 334$

$AM_X = 60.0$ $AM_Y = 30.0$ $\sigma_{x'} = 7.04$

$M_X = 62.5$ $M_Y = 30.4$ $\sigma_{y'} = 4.86$

$C_x = 2.5$ $C_y = .4$

$C_x^2 = 6.25$ $C_y^2 = .16$

Applying formula (30)

$$r_{xy} = \frac{\frac{\Sigma x' y'}{N} - C_x C_y}{\sigma'_x \sigma'_y}$$

$$= \frac{\frac{334}{12} - 1.00}{7.04 \times 4.86}$$

$$= \frac{27.83 - 1.00}{34.21}$$

$$\sigma_{x'} = \sqrt{\frac{\Sigma x^2}{N} - Cx^2} = \sqrt{\frac{670}{12} - 6.25}$$

$$= \sqrt{55.85 - 6.25} = \sqrt{49.58} = 7.04$$

$$\sigma'_y = \sqrt{\frac{\Sigma y'^2}{N} - C_y^2} = \frac{26.83}{34.21}$$

$$r = \cdot 78$$

$$= \sqrt{\frac{285}{12} - .16} = \sqrt{23.75 - .16} = \sqrt{23.59} = 4.86$$

Properties of *r*:

1. The value of the coefficient of correlation r remains unchanged when a constant is added to one or both variables: In order to observe the effect on the coefficient correlation r when a constant is added to one or both the variables, we consider an example.

Now, we add a score of 10 to each score in X and 20 to each score of Y and represent these scores by X' and Y' respectively.

The calculations for computing r for original and new pairs of observations are given in Table 5.4:

Table 5.4 Effect on *r* when a constant is added to variables

Sl. No.	Original Scores					New Scores				
	X	Y	X^2	Y^2	XY	X'	Y'	X'^2	Y'^2	X'Y'
1	1	4	1	16	4	11	24	121	576	264
2	3	3	9	9	9	13	23	169	529	299
3	5	7	25	49	35	15	27	225	729	405
4	6	8	36	64	48	16	28	256	784	448
N =4	ΣX =15	ΣY =22	ΣX^2 =71	ΣY^2 =138	ΣXY =96	ΣX' =55	ΣY' =102	$\Sigma X'^2$ =771	$\Sigma Y'^2$ =2618	ΣX'Y' =1416

By using formula (29), the coefficient of correlation of original score will be:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{4 \times 96 - 15 \times 22}{\sqrt{[4 \times 71 - (15)^2][4 \times 138 - (22)^2]}} = \frac{384 - 330}{\sqrt{[284 - 225][552 - 484]}}$$

$$= \frac{54}{63.34} = +0.85$$

The same formula for new scores can be written as:

$$r = \frac{N\Sigma X'Y' - (\Sigma X')(\Sigma Y')}{\sqrt{[N\Sigma X'^2 - (\Sigma X')^2][N\Sigma Y'^2 - (\Sigma Y')^2]}}$$

$$= \frac{4 \times 1416 - 55 \times 102}{[4 \times 771 - (55)^2][4 \times 2618 - (102)^2]}$$

$$= \frac{5664 - 5610}{\sqrt{[3084 - 3025][10472 - 10404]}} = \frac{54}{\sqrt{59 \times 68}} = \frac{54}{63.34} = +0.85$$

Thus, we observe that the value of the coefficient of correlation r remains unchanged when a constant is added to one or both variables.

2. The value of the coefficient of correlation r remains unchanged when a constant is subtracted from one or both variables: Students can examine this by taking an example. When each score of one or both variables are subtracted by a constant the value of coefficient of correlation r also remains unchanged.

3. The value of the coefficient of correlation r remains unaltered when one or both sets of variate values are multiplied by some constant: In order to observe the effect of multiplying the variables by some constant on the value of r, we arbitrarily multiply that original scores of first and second sets in the previous example by 10 and 20 respectively.

The r between X' and Y' may then be calculated as under:

Table 5.5 Effect on *r* when variables are multiplied by some constant

	New Scores				
S.No.	X'	Y'	X'^{2}	Y'^{2}	X'Y'
1	10	80	100	6400	800
2	30	60	900	3600	1800
3	50	140	2500	19600	7000
4	60	160	3600	25600	9600
N=4	$\Sigma X'=150$	$\Sigma Y'=440$	$\Sigma X'^2=7100$	$\Sigma Y'^2=55200$	$\Sigma X'Y'=19200$

The correlation of coefficient between X' and Y' will be:

$$r = \frac{N\Sigma X'Y' - (\Sigma X')(\Sigma Y')}{\sqrt{[N\Sigma X'^2 - (\Sigma X')^2][N\Sigma Y'^2 - (\Sigma Y')^2]}}$$

Putting values from table,

$$r = \frac{4 \times 19200 - (150)(440)}{\sqrt{[4 \times 7100 - (150)^2][4 \times 55200 - (440)^2]}}$$

$$= \frac{10800}{\sqrt{5900 \times 27200}} = \frac{108}{126.68} = +0.85$$

Thus, we observe that the value of the coefficient of correlation r remains unchanged when a constant is multiplied with one or both sets of variate values.

4. The value of r will remain unchanged even when one or both sets of variate values are divided by some constant:

Students can examine this by taking an example.

B. Coefficient of Correlation in Grouped Data: When the number of pairs of measurements (N) on two variables X and Y are large, even moderate in size, and when no calculating machine is available, the customary procedure is to group data in both X and Y and to form a scatter diagram or correlation diagram which is also called two-way frequency distribution or bivariate frequency distribution.

The choice of size of class interval and limits of intervals follows much the same rules as were given previously. To clarify the idea, we consider a bivariate data concerned with the scores earned by a class of 20 students in Physics and Mathematics examination.

Preparing a Scatter diagram: In setting up a double grouping of data, a table is prepared with columns and rows. Here, we classify each pair of variates simultaneously in the two classes, one representing score in Physics (X) and the other in Mathematics (Y) as shown in Table 5.6.

Student No.	Scores in Physics (X)	Scores in Mathematics (Y)	Student No.	Scores in Physics (X)	Scores in Mathematics (Y)
1	32	25	11	31	10
2	34	41	12	42	25
3	48	53	13	57	44
4	35	12	14	48	32
5	52	26	15	63	42
6	45	28	16	53	45
7	57	51	17	48	31
8	62	54	18	43	23
9	67	50	19	71	28
10	73	48	20	52	22

The scores of 20 students in both Physics (X) and Mathematics (Y) are shown in Table below:

Table 5.6 Bivariate Frequency Table

Y–Variate (Scores in Mathematics)

X–Variate (Scores in Physics) C.i's	10-19	20-29	30-39	40-49	50-59	f_x
70-79		1 /		1 /		2
60-69				1 /	2 //	3
50-59		2 //		2 //	1 /	5
40-49		3 ///	2 //		1 /	6
30–39	2 //	1 /		1 /		4
f_y	2	7	2	5	4	N = 20

We can easily prepare a bivariate frequency distribution table by putting tallies for each pair of scores. The construction of a scattergram is quite simple. We have to prepare a table as shown in the diagram above.

Along the left hand margin the class intervals of X-distribution are laid off from bottom to top (in ascending order). Along the top of the diagram the c.i's of Y-distribution are laid off from left to right (in ascending order).

Each pair of scores (both in X and Y) is represented through a tally in the respective cell. No. 1 student has secured 32 in Physics (X) and 25 in Mathematics (Y). His score of 32 in (X) places him in the last row and 25 in (Y) places him in the second column. So, for the pair of scores (32, 25) a tally will be marked in the second column of 5th row.

In a similar way, in case of No. 2 student, for scores (34, 41), we shall put a tally in the 4th column of the 5th row. Likewise, 20 tallies will be put in the respective rows and columns. (The rows will represent the X-scores and the columns will represent the Y-scores).

Along the right-hand margin the f_x column, the number of cases in each c.i., of X-distribution are tabulated and along the bottom of the diagram in the f_y row the number of cases in each c.i., of Y-distribution are tabulated.

The total of f_x column is 20 and the total of f_y row is also 20. It is in fact a bi-variate distribution because it represents the joint distribution of two variables. The scattergram is then a "correlation table."

Calculation of r from a correlation table: The following outline of the steps to be followed in calculating r will be best understood if the student will constantly refer to Table 5.7 as he reads through each step:

Step 1: Construct a scattergram for the two variables to be correlated, and from it draw up a correlation table.

Step 2: Count the frequencies of each c.i. of distribution – X and write it in the f_x column. Count the frequencies for each c.i. of distribution – Y and fill up the f_y row.

Step 3: Assume a mean for the X-distribution and mark off the c.i. in double lines. In the given correlation table, let us assume the mean at the c.i., 40 – 49 and put double lines as shown in the table. The deviations above the line of A.M. will be (+ ve) and the deviations below it will be (- ve).

The deviation against the line of A.M., i.e., against the c.i. where we assumed the mean is marked 0 (zero) and above it the *d*'s are noted as +1, +2. 13 and below it *d* is noted to be – 1. Now dx column is filled up. Then multiply f_x. and *dx* of each row to get *fdx*. Multiply *dx* and *fdx* of each row to get fdx^2.

[Note: While computing the S.D. in the assumed mean method we were assuming a mean, marking the d's and computing *fd* and fd^2. Here also same procedure is followed.]

Step 4: Adopt the same procedure as in step 3 and compute *dy*, *fdy* and fdy^2. For the distribution-Y, let us assume the mean in the c.i. 20-29 and put double lines to mark off the column as shown in the table. The deviations to the left of this column will be negative and right be positive.

Thus, d for the column where mean is assumed is marked 0 (zero) and the d to its left is marked – 1 and *d*'s to its right are marked +1, +2 and +3. Now *dy* column is filled up. Multiply the values of *fy* and *dy* of each column to get *fdy*. Multiply the values of *dy* and *fdy* to each column to get fdy^2.

Step 5: As this phase is an important one, we are to mark carefully for the computation of *dy* for different c.i.'s of distribution X and *dx* for different c.i.'s of distribution -Y.

dy for different c. i. 's of distribution-X: In the first row, 1*f* is under the column, 20-29 whose *dy* is 0 (Look to the bottom. The *dy* entry of this row is 0). Again 1*f* is under the column, 40- 49 whose *dy* is + 2. So *dy* for the first row = (1 x 0) + (1 x 2) = + 2.

In the second row we find that:

1 *f* is under the column, 40-49 whose *dy* is + 2 and

2 *f*s are under the column, 50-59 whose *dy*'s are + 3 each.

So *dy* for 2nd row = (1 x 2) + (2 X 3) = 8.

In the third row,

2 *f*s are under the column, 20-29 whose *dy*'s are 0 each,

2 *f*s are under the column, 40-49 whose *dy*'s are +2 each, and 1 *f* is under the column, 50-59 whose *dy* is +3.

So dy for the 3rd row = (2 x 0) + (2 x 2) + (1 X 3) = 7.

In the 4th row,

3 *f*s are under the column, 20-29 whose *dy*'s are 0 each,

2 *f*s are under the column, 30-39 whose *dy*'s are +1 each, and 1 *f* is under the column, 50-59 whose *dy* is + 3,

So *dy* for the 4th row = (3 X 0) + (2 X 1) + (1 x 3) = 5.

Likewise in the 5th row

dy for the 5th row = (2 x – 1) + (1 x 0) + (1 x 2) = 0

dx for different c.i. ,'v of distribution – Y :
In the first column,
2 *f*s are against the row, 30-39 whose *dx* is – 1.
So *dx* of the 1st column = (2 x – 1) = – 2
In the second column,
1 *f* is against the c.i., 70-79 whose *dx* is +3,
2 *f*s are against the c.i., 50-59 whose *dx*'s are +1 each,
3 *f*s are against the c.i., 40-49 whose *dx*'s are 0 each,
1 *f* is against the c.i., 30-39 whose *dx* is – 1.
So *dx* for the 2nd column = (1 x 3) + (2 X 1) + (3 X 0) + (1 x – 1) = 4. In the third column,
dx for the 3rd column = 2×0 = 0
In the fourth column,
dx for the 4th column = (1 x 3) + (1 x 2) + (2 x 1) + (1 x – 1) = 6.
In the fifth column,
dx for the 5th column = (2 x 2) + (1 x 1) + (1 X 0) = 5.

Step 6: Now, calculate *dx.dy* each row of distribution – X by multiplying the *dx* entries of each row by *dy* entries of each row. Then calculate *dx.dy* for each column of distribution – Y by multiplying *dy* entries of each column by the *dx* entries of each column.

Step 7: Now, take the algebraic sum of the values of the columns fdx, fdx^2, *dy* and *dx.dy* (for distribution – X). Take the algebraic sum of the values of the rows *fdy*, fdy^2, *dx* and *dx.dy* (for distribution – Y)

Step 8: ∑.*dx.dy* of X-distribution = ∑*dx.dy* of Y-distribution
∑*fdx* = total of *dx* row (i.e. ∑*dx*)
∑*fdy* = total of *dy* column (i.e. ∑*dy*)

Step 9: The values of the symbols as found
$\sum fdx = 13$, $\sum fd^2x = 39$
$\sum fdy = 22$, $\sum fd^2y = 60$
$\sum dx.dy = 29$ and N = 20.

Table 5.7
Computation of Coefficient of Correlation

Y–Variables –Scores on Mathematics

X–Variable–Scores on Physics

c.i's	10 –19	20 –29	30 –39	40 –49	50 –59	f_x	d_x	fd_x	fd_x^2	d_y	$d_x.d_y$
70–79		1		1		2	+3	6	18	2	6
60–69				1	2	3	+2	6	12	8	16
50–59		2		2	1	5	+1	5	5	7	7
40–49		3	2		1	6	0	0	0	5	0
30–39	2	1		1		4	–1	–4	4	0	0
f_y	2	7	2	5	4	20		Σfd_x=13	Σfdx^2=39	Σd_y=22	$\Sigma d_x d_y$=29
d_y	–1	0	+1	+2	+3						
fd_y	–2	0	2	10	12	Σfd_y=22					
fd_y^2	2	0	2	20	36	Σfd_y^2=60					
d_x	–2	4	0	6	5	Σd_x=13					
$d_x.d_y$	2	0	0	12	15	$\Sigma d_x d_y$=29					

Check

Check

Check

In order to compute coefficient of correlation in a correlation table following formula can be applied:

$$r = \frac{\frac{\Sigma dx.dy}{N} - C_x.C_y}{\sigma_x \cdot \sigma_y} \qquad ...(31)$$

$$C = \frac{\Sigma fd}{N}.$$

Thus, for $\frac{\Sigma fdx}{N}$ we can write C_x and for $\frac{\Sigma fdy}{N}$ we can write C_y.

Again while calculating S.D. by assumed mean method we know that

$$\sigma = i \times \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

We may mark that in the denominator of formula (31) we apply the formula for a_x and a_y with the exception of no i's. We may note here that C_x, C_y, σ_x, σ_v are all expressed in units of class intervals (i.e., in unit of i). Thus, while computing σ_x and σ_y, no i's are used. This is desirable because all the product deviations i.e., $\sum dx.dy$'s are in interval units.

Thus, we compute:

$$C_x = \frac{\Sigma fdx}{N} = \frac{13}{20} = .65,\ C_y = \frac{\Sigma fdy}{N} = \frac{22}{20} = 1.1$$

$$\sigma_x = \sqrt{\frac{\Sigma fd^2x}{N} - \left(\frac{\Sigma fdx}{N}\right)^2} = \sqrt{\frac{39}{20} - \left(\frac{13}{20}\right)^2} = \sqrt{1.95 - (.65)^2}$$

$$= \sqrt{1.95 - .4225} = \sqrt{1.5275} = 1.24.$$

So σ_x (ignoring 'i') = 1.24

$$\sigma_y = \sqrt{\frac{\Sigma fd^2y}{N} - \left(\frac{\Sigma fdy}{N}\right)^2} = \sqrt{\frac{60}{20} - \left(\frac{22}{20}\right)^2} = \sqrt{3 - (1.1)^2}$$

$$= \sqrt{3 - 1.21} = \sqrt{1.79} = 1.34$$

So σ_y (ignoring 'i') = 1.34

Applying formula (31)

$$r = \frac{\frac{\Sigma dx.dy}{N} - C_x \cdot C_y}{\sigma_x \cdot \sigma_y} = \frac{\frac{29}{20} - .65 \times 1.1}{1.24 \times 1.34} = \frac{.735}{1.66} = .44$$

Interpretation of the Coefficient of Correlation: Merely computation of correlation does not have any significance until and unless we determine how large must the coefficient be in order to be significant, and what does correlation tell us about the data? What do we mean by the obtained value of coefficient of correlation?

Size of Correlation	Interpretation
± 1	Perfect Positive/Negative Correlation
± .90 to ± .99	Very High Positive/Negative Correlation
± .70 to ± .90	High Positive/Negative Correlation
± .50 to ± .70	Moderate Positive/Negative Correlation
± .30 to ± .50	Low Positive/Negative Correlation
± .10 to ± .30	Very Low Positive/Negative Correlation
± .00 to ± .10	Markedly Low and Negligible Positive/ Negative Correlation

Misinterpretation of the Coefficient of Correlation: Sometimes, we misinterpret the value of coefficient of correlation and establish the cause and effect relationship, i.e. one variable causing the variation in the other variable. Actually we cannot interpret in this way unless we have sound logical base.

Correlation coefficient gives us, a quantitative determination of the degree of relationship between two variables X and Y, not information as to the nature of association between the two variables. Causation implies an invariable sequence— A always leads to B, whereas correlation is simply a measure of mutual association between two variables.

Factors influencing the size of the Correlation Coefficient:

We should also be aware of the following factors which influence the size of the coefficient of correlation and can lead to misinterpretation:

1. The size of "r" is very much dependent upon the variability of measured values in the correlated sample. The greater the variability, the higher will be the correlation, everything else being equal.

2. The size of 'r' is altered, when an investigator selects an extreme group of subjects in order to compare these groups with respect to certain behavior. "r" obtained from the combined data of extreme groups would be larger than the "r" obtained from a random sample of the same group.

3. Addition or dropping the extreme cases from the group can lead to change on the size of "r". Addition of the extreme case may increase the size of correlation, while dropping the extreme cases will lower the value of "r".

Uses of Product moment r:

Correlation is one of the most widely used analytic procedures in the field of Educational and Psychological Measurement and Evaluation. It is useful in:

i. Describing the degree of correspondence (or relationship) between two variables.

ii. Prediction of one variable—the dependent variable on the basis of independent variable.

iii. Validating a test; e.g., a group intelligence test.

iv. Determining the degree of objectivity of a test.

v. Educational and vocational guidance and in decision-making.

vi. Determining the reliability and validity of the test.

vii. Determining the role of various correlates to a certain ability.

viii. Factor analysis technique for determining the factor loading of the underlying variables in human abilities.

Assumptions of Product moment r:

1. Normal distribution: The variables from which we want to calculate the correlation should be normally distributed. The assumption can be laid from random sampling.

2. Linearity: The product-moment correlation can be shown in straight line which is known as linear correlation.

3. Continuous series: Measurement of variables on continuous series.

4. Homoscedasticity: It must satisfy the condition of homoscedasticity (equal variability).

5.4.5 *Spearman's Rank Correlation Coefficient:* There are some situations in Education and Psychology where the objects or individuals may be ranked and arranged in order of merit or proficiency on two variables and when these 2 sets of ranks covary or have agreement between them, we measure the degrees of relationship by rank correlation.

Again, there are problems in which the relationship among the measurements made is non-linear, and cannot be described by the product-moment r.

For example, the evaluation of a group of students on the basis of leadership ability, the ordering of women in a beauty contest, students ranked in order of preference or the pictures may be ranked according to their aesthetic values. Employees may be rank-ordered by supervisors on job performance.

School children may be ranked by teachers on social adjustment. In such cases objects or individuals may be ranked and arranged in order of merit or proficiency on two variables. Spearman has developed a formula called Rank Correlation Coefficient to measure the extent or degree of correlation between 2 sets of ranks.

This coefficient of correlation is denoted by Greek letter ρ (called Rho) and is given as:

$$\rho = 1 - \frac{6 \times \Sigma D^2}{N(N^2 - 1)} \qquad ...(32)$$

where, ρ = rho = Spearman's Rank Correlation Coefficient

D = Difference between paired ranks (in each case)

N = Total number of items/individuals ranked.

Characteristics of Rho (ρ):

1. In Rank Correlation Coefficient the observations or measurements of the bivariate variable is based on the ordinal scale in the form of ranks.

2. The size of the coefficient is directly affected by the size of the rank differences.

(a) If the ranks are the same for both tests, each rank difference will be zero and ultimately D^2 will be zero. This means that the correlation is perfect; i.e. 1.00.

(b) If the rank differences are very large, and the fraction is greater than one, then the correlation will be negative.

Assumptions of Rho (ρ):

i. N is small or the data are badly skewed.

ii. They are free, or independent, of some characteristics of the population distribution.

iii. In many situations Ranking methods are used, where quantitative measurements are not available.

iv. Though quantitative measurements are available, ranks are substituted to reduce arithmetical labour.

v. Such tests are described as non-parametric.

vi. In such cases the data are comprised of sets of ordinal numbers, 1st, 2nd, 3rd….Nth. These are replaced by the cardinal numbers 1, 2, 3,……, N for purposes of calculation. The substitution of cardinal numbers for ordinal numbers always assumes equality of intervals.

I. Calculating ρ from Test Scores:

Example 1:

The following data give the scores of 5 students in Mathematics and General Science respectively: Compute the correlation between the two series of test scores by Rank Difference Method.

Table 5.8 Computation of ρ (Rho)

Student	Scores in Math.	Scores in Gen. Sci.	Rank in Test 1 R_1	Rank in Test 2 R_2	Diff. in ranks D R_1-R_2	D^2
A	8	10	2	1	1	1
B	7	8	3	2	1	1
C	9	7	1	3	−2	4
D	5	4	4	5	−1	1
E	1	5	5	4	1	1
N=5					ΣD=0	$\Sigma D^2=8$

$$\rho = 1 - \frac{6 \times \Sigma D^2}{N(N^2-1)} = 1 - \frac{6 \times 8}{5(5^2-1)} = 1 - \frac{48}{120} = 1 - .40 = +.60.$$

The value of coefficient of correlation between scores in Mathematics and General Science is positive and moderate.

Steps of Calculation of Spearman's Co-efficient of Correlation:

Step 1: List the students, names or their serial numbers in column 1.

Step 2: In column 2 and 3 write scores of each student or individual in test I and II.

Step 3: Take one set of score of column 2 and assign a rank of 1 to the highest score, which is 9, a rank of 2 to the next highest score which is 8 and so on, till the lowest score get a rank equal to N; which is 5.

Step 4: Take the II set of scores of column 3, and assign the rank 1 to highest score. In the second set the highest score is 10; hence obtain rank 1. The next highest score of B student is 8; hence his rank is 2. The rank of student C is 3, the rank of E is 4, and the rank of D is 5.

Step 5:

Calculate the difference of ranks of each student (column 6).

Step 6:

Check the sum of the differences recorded in column 6. It is always zero.

Step 7:

Each difference of ranks of column 6 is squared and recorded in column 7. Get the sum $\sum D^2$.

Step 8:

Put the value of N and $2D^2$ in the formula of Spearman's co-efficient of correlation.

2. Calculating from Ranked Data:

Example 2:

In a speech contest Prof. Mehrotra and Prof. Shukla, judged 10 pupils. Their judgements were in ranks, which are presented below. Determine the extent to which their judgements were in agreement.

Table 5.9 Computation of ρ (Rho)

Pupil	Prof. Mehrotra's Ranks (R_1)	Prof. Shukla's Ranks (R_2)	Difference $D = (R_1 - R_2)$	D_2
A	1	1	0	0
B	3	2	1	1
C	4	5	–1	1
D	7	9	–2	4
E	6	6	0	0
F	9	8	1	1
G	8	10	–2	4
H	10	7	3	9
1	2	4	–2	4
J	5	3	2	4
N=10			$\Sigma D = 0$	$\Sigma D^2 = 28$

$$\rho = 1 - \frac{6 \times \Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 28}{10(10^2 - 1)} = 1 - \frac{6 \times 28}{990} = 1 - .17 = +.83$$

The value of co-efficient of correlation is + .83. This shows a high degree of agreement between the two judges.

3. Calculating ρ (Rho) for tied Ranks:

Example 3:

The following data give the scores of 10 students on two trials of test with a gap of 2 weeks in Trial I and Trial II.

Compute the correlation between the scores of two trials by rank difference method:

Table 5.10 Computation of ρ (Rho)

Student	Trial-1 (X)	Trial-2 (Y)	Rank on Trial I R_1	Rank on Trial II R_2	Diff. D	D^2
A	10	16	6.5	5.5	1.0	1.00
B	15	16	3	5.5	−2.5	6.25
C	11	24	5	1.5	3.5	12.25
D	14	18	4	4	0	0
E	16	22	2	3	−1.0	1.00
F	20	24	1	1.5	− 0.5	0.25
G	10	14	6.5	7.5	−1.0	1.00
H	8	10	9	10	− 1.0	1.00
I	7	12	10	9	1.0	1.00
J	9	14	8	7.5	0.5	0.25
N=10					ΣD =00	ΣD^2 =24.00

$$\rho = 1 - \frac{6 \times \Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 24}{10(10^2 - 1)} = 1 - \frac{6 \times 24}{10 \times 99} = 1 - .145$$

$\rho = +.855.$

The correlation between Trial I and II is positive and very high. Look carefully at the scores obtained by the 10 students on Trial I and II of the test.

Do you find any special feature in the scores obtained by the 10 students? Probably, your answer will be “yes”.

In the above table in column 2 and 3 you will find that more than one students are getting the same scores. In column 2 students A and G are getting the same score viz. 10. In column 3, the students A and B, C and F and G and J are also getting the same scores, which are 16, 24 and 14 respectively.

Definitely these pairs will have the same ranks; known as Tied Ranks. The procedure of assigning the ranks to the repeated scores is somewhat different from the non-repeated scores. Look at column 4. Student A and G have similar scores of 10 each and they possess 6th and 7th rank in the group. Instead of assigning the 6th and 7th rank, the average of the two rank i.e. 6.5 (6 + 7/2 = 13/2) has been assigned to each of them.

The same procedure has been followed in respect of scores on Trial II. In this case, ties occur at three places. Students C and F have the same score and hence obtain the average rank of (1 + 2/2 = 1.5). Student A and B have rank position 5 and 6; hence are assigned 5.5 (5 + 6/2) rank each. Similarly student G and J have been assigned 7.5 (7 + 8/2) rank each.

If the values are repeated more than twice, the same procedure can be followed to assign the ranks:

For example:

if three students get a score of 10, at 5th, 6th and 7th ranks, each one of them will be assigned a rank of 5 + 6 + 7/3= 6.

The rest of the steps of procedure followed for calculation of ρ (rho) are the same as explained earlier.

Interpretation:

The value of ρ can also be interpreted in the same way as Karl Pearson's Coefficient of Correlation. It varies between -1 and + 1. The value + 1 stands for a perfect positive agreement or relationship between two sets of ranks while ρ = – 1 implies a perfect negative relationship. In case of no relationship or agreement between ranks, the value of ρ = 0.

Advantages of Rank Difference Method:

1. The Spearman's Rank Order Coefficient of Correlation computation is quicker and easier than (r)
2. Easy to interpret p.

Limitations:

1. Computed by the Pearson's Product Moment Method.
2. It is an acceptable method if data are available only in ordinal form or number of paired variable is more than 5 and not greater than 30 with minimum or a few ties in ranks.
3. When the interval data are converted into rank-ordered data the information about the size of the score differences is lost; e.g. in the Table 5.10, if D in Trial II gets scores from 18 up to 21, his rank remains only 4.
4. If the number of cases is more, giving ranks to them becomes a tedious job.

REGRESSION:

5.4.6 Introduction: **Regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which a researcher finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For example, the method of ordinary least squares computes the unique line (or hyperplane) that minimizes the sum of squared distances between the true data and that line (or hyperplane). For specific mathematical reasons (see linear regression), this allows the researcher to estimate the conditional expectation (or population average value) of the dependent variable when the independent variables take on a given set of values. Less common forms of regression use slightly different procedures to estimate alternative location parameters (e.g., quantile regression or Necessary Condition Analysis[1]) or estimate the conditional expectation across a broader collection of non-linear models (e.g., nonparametric regression).

Regression analysis is primarily used for two conceptually distinct purposes. First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Second, in some situations regression analysis can be used to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset. To use regressions for prediction or to infer causal relationships, respectively, a researcher must carefully justify why existing relationships have predictive power for a new context or why a relationship between two variables has a causal interpretation. The latter is especially important when researchers hope to estimate causal relationships using observational data.

5.4.7 Objectives of Regression: The aim of regression analysis is to examine the relationships between one set of variables (the dependent variable(s) aka outcome, target, etc) and another set (independent variables, predictors, etc.)
There can be one or more variables in each set.
The goal can be focused on explanation, prediction or both.

5.4.8 Classification of regression analysis: The regression analysis can be classified on the following bases: -
(i) Change in Proportion; and
(ii) Number of variables
Basis of Change in Proportions: On the basis of proportions the regression can be classified into the
following categories: -
1. Linear regression and
2. Non-linear regression

Linear regression quantifies the relationship between one or more *predictor variable(s)* and one *outcome variable*. Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). Linear regression is also known as *multiple regression*, *multivariate regression*, *ordinary least squares (OLS)*, and *regression*. This post will show you examples of linear regression, including an example of *simple linear regression* and an example of *multiple linear regression*.
Nonlinear Regression: Nonlinear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function. Simple linear regression relates two variables (X and Y) with a straight line (y = mx + b), while nonlinear regression must generate a line (typically a curve) as if every value of Y was a random variable.
The goal of the model is to make the sum of the squares as small as possible. The sum of squares is a measure that tracks how much observations vary from the mean of the data set. It

is computed by first finding the difference between the mean and every point of data in the set. Then, each of those differences is squared. Lastly, all of the squared figures are added together. The smaller the sum of these squared figures, the better the function fits the data points in the set. Nonlinear regression uses logarithmic functions, trigonometric functions, exponential functions, power functions, Lorenz curves, Gaussian functions, and other fitting methods.

Nonlinear regression modeling is similar to linear regression modeling in that both seek to track a particular response from a set of variables graphically. Nonlinear models are more complicated than linear models to develop because the function is created through a series of approximations (iterations) that may stem from trial-and-error. Mathematicians use several established methods, such as the Gauss-Newton method and the Levenberg-Marquardt method.

On the basis of Number of Variables: On the basis of number of variables regression analysis can be classified as under:

1. Simple regression: When only two variables are studied to find the regression relationships, it
is known as simple regression analysis. Of these variables, one is treated as an independent variable while the other as dependent one.

2. Partial Regression: When more than two variables are studied in a functional relationship but
the relationship of only two variables is analyzed at a time, keeping other variables as constant, such a regression analysis is called partial regression.

3. Multiple Regression: When more than two variables are studied and their relationships are simultaneously worked out, it is a case of multiple regression.

5.4.9 Regression Lines: A regression line is a line that best describes the behavior of a set of data. In other words, it's a line that best fits the trend of a given data.

Regression lines are very useful for forecasting procedures. The purpose of the line is to describe the interrelation of a dependent variable (Y variable) with one or many independent variables (X variable). By using the equation obtained from the regression line an analyst can forecast future behaviors of the dependent variable by inputting different values for the independent ones. Regression lines are widely used in the financial sector and in business in general.

5.4.10 Methods of Drawing Regression Lines: The regression lines can be drawn by two methods as given below: -

1. Free Hand Curve Method: This is a familiar concept, and is briefly described for drawing frequency curves. In case of a time series a scatter diagram of the given observations is plotted against time on the horizontal axis and a freehand smooth curve is drawn through the plotted points. The curve is so drawn that most of the points concentrate around the curve, however, smoothness should not be scarified in trying to let the points fall exactly on the curve. It is

better to draw a straight line through the plotted points instead of a curve, if possible. The curve fitted by this method eliminates the short term and long term oscillations and the irregular movements from the time series, and elevates the general trend. After having drawn such a curve or line, the trend values or the estimated YY values, which may be denoted by YY, can be read from the graph corresponding to each time period.

One of the major disadvantages of this method is that different individuals draw curves or lines that differ in slope and intercept, and hence no two conclusions are identical. However, it is the most simple and quickest method of isolating the trend. This method is generally employed in situations where the scatter diagram of the original data conforms to some well define trends.

2. The method of Least Squares: The Method of Least Squares is a procedure to determine the best fit line to data; the proof uses simple calculus and linear algebra. The basic problem is to find the best fit straight line y = ax + b given that, for n ∈ {1, . . . , N}, the pairs (xn, yn) are observed. The method easily generalizes to finding the best fit of the form.

5.4.11 Methods of Calculating Regression Equations or Derivation of Regression Lines:

Following are the two methods to form the two regression equations, that equation for Y on X and, for X on Y.

1. Regression equations through normal equations
2. Regression equations through regression co-efficient.

Regression Equations through Normal Equations: The two main equations generally used in

regression analysis is:

(i) Y on X,

(ii) X on Y

for Y on X, the equation is Yc = a + bX

for X on Y, the equation is Xc = a + bY

Where:'a' and 'b' are constant values and 'a' is called the intercept.

Regression Equation of Y on X:

The regression equation of Y on X can be written as Yc = a + bX

Regression Equation of X on Y

The regression of X on Y is expressed as

Xc = a + b Y

5.4.12 Regression Equations through Regression Coefficients:

Regression coefficient refers to the constant value multiplied to the independent variable in a given
relation. In a relation, Y = a + bx, here b (the slope of the regression line) is the regression coefficient,
since it is a multiple of independent variable x, Regression equations or lines can easily be arrived at
by the use of regression coefficients.

5.4.13 Properties of Regression Coefficients

1. Correlation coefficient is the geometric mean between the regression coefficients.

$$r_{XY} = \sqrt{b_{XY} \times b_{YX}}$$

2. It is clear from the property 1, both regression coefficients must have the same sign. *i.e.,* either they will positive or negative.
3. If one of the regression coefficients is greater than unity, the other must be less than unity.
4. The correlation coefficient will have the same sign as that of the regression coefficients.
5. Arithmetic mean of the regression coefficients is greater than the correlation coefficient.

$$\frac{b_{XY} + b_{YX}}{2} \geq r_{XY}$$

6. Regression coefficients are independent of the change of origin but not of scale.

5.4.14 Properties of regression equation

1. If $r = 0$, the variables are uncorrelated, the lines of regression become perpendicular to each other.
2. If $r = 1$, the two lines of regression either coincide or parallel to each other.
3. Angle between the two regression lines is $\theta = \tan^{-1} (m_1 - m_2 \ / \ 1+m_1m_2)$ where m1 and m2 are the slopes of regression lines X on Y and Y on X respectively.
4. The angle between the regression lines indicates the degree of dependence between the variable.

5.4.15 Difference between Correlation and Regression: Correlation and Regression are the two-analysis based on multivariate distribution. A multivariate distribution is described as a distribution of multiple variables. **Correlation** is described as the analysis which lets us know the association or the absence of the relationship between two variables ‘x’ and ‘y’. On the other end, **Regression** analysis, predicts the value of the dependent variable based on the known value of the independent variable, assuming that average mathematical relationship between two or more variables.

BASIS FOR COMPARISON	CORRELATION	REGRESSION
Meaning	Correlation is a statistical measure which determines co-relationship or of two association variables.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

Sub Unit - 5 PROBABILITY

5.5.1: Approaches to Probability: The three approaches to probability are:

- Classical approach
- Frequency-based (or empirical) approach
- Subjective approach
  - **Classical Approach:** This approach traces back to the field where probability was first systematically employed, which is gambling (flipping coins, tossing dice and so forth). Gambling problems are characterized by random experiments which have *n* possible outcomes, equally likely to occur. It means that none of them is more or less likely to occur than other ones, hence they are said to be in a symmetrical position.
    The classical approach is pretty intuitive, nevertheless it suffers from some pitfalls:
    The assumption of symmetry is far too strong and unrealistic. Namely, imagine you want to know the probability of the event "tomorrow I will have a car accident". The possible outcomes of this scenario are two: having a car accident or not having a car accident. Given that *k*=having a car accident, the probability of that event is 1/2, which, besides being a bit worrying, is not representative of the real likelihood of the event.
    In this approach, there is no space for the concept of information, which is strictly related to probability. Let's think about the previous example of the dice. Imagine you are told this dice is loaded and, instead of having the number "one", it has two "six" (so the faces will be 2,3,4,5,6,6). Provided with this information, which probability would you attribute to the event "one"? Since it is impossible, the probability is equal to zero and not 1/6. Hence, probability does depend on the available information (the intuition will be clearer in the subjective approach)
  - **Frequency-based (or empirical) approach:** This approach was formally introduced in the field of natural science, where the assumption of symmetric position poorly fails. Instead, the idea on which this approach is based is that several experiments can be run under certain conditions considered as equivalent. Each experiment might lead either to success or to an in success.
    This approach is not lacking of criticisms. Again, there is one big assumption which is the convergence property of the frequency, whose limit might not exist. Repeating experiments under equivalent conditions might not be possible. There are events extremely rare, for which is impossible to run many simulations (think about extreme natural events like *tsunami*).
  - **Subjective approach:** Developed by probabilist B. de Finetti, this is the most intuitive definition of probability. Indeed, according to that approach, the probability of an event is the degree of belief a person attaches to that event, based on his/her available information. This last approach does not count serious criticisms, since it resolves some pitfalls of the previous approaches (like the impossibility of repeating experiments under equivalent conditions, because of the uniqueness of many events) and, at the same time, does not contrast with other theories. Indeed, the evaluator who has to decide the price

of the lottery is not prevented from running experiments, compute the frequency of successes and use this information to propose a price. Basically, what in other approaches was a rule, in the subjective approach is an option.

5.5.2: Bayes' Theorem: In statistics and probability theory, the Bayes' theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes' theorem describes the probability.

Total Probability Rule: The Total Probability Rule (also known as the law of total probability) is a fundamental rule in statistics relating to conditional and marginal of an event based on prior knowledge of the conditions that might be relevant to the event.

The theorem is named after English statistician Thomas Bayes, who discovered the formula in 1763. It is considered the foundation of the special statistical inference approach called the Bayes' inference.

Besides statistics, basic Statistics concepts for finance, a solid understanding of statistics is crucially important in helping us better understand finance. Moreover, statistics concepts can help investors monitor, the Bayes' theorem is also used in various disciplines, with medicine and pharmacology as the most notable examples. In addition, the theorem is commonly employed in different fields of finance. Some of the applications include, but are not limited to, modeling the risk of lending money to borrowers or forecasting the probability of the success of an investment.

Formula for Bayes' Theorem: The Bayes' theorem is expressed in the following formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

- P(A|B) – the probability of event A occurring, given event B has occurred
- P(B|A) – the probability of event B occurring, given event A has occurred
- P(A) – the probability of event A
- P(B) – the probability of event B

Note that events A and B are independent eventsIndependent EventsIn statistics and probability theory, independent events are two events wherein the occurrence of one event does not affect the occurrence of another event (i.e., the probability of the outcome of event A does not depend on the probability of the outcome of event B).

A special case of the Bayes' theorem is when event A is a binary variable. In such a case, the theorem is expressed in the following way:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A^-)P(A^-) + P(B|A^+)P(A^+)}$$

Where:

- $P(B|A^-)$ – the probability of event B occurring given that event A^- has occurred
- $P(B|A^+)$ – the probability of event B occurring given that event A^+ has occurred

In the special case above, events A^- and A^+ are mutually exclusive outcomes of event A.

Sub Unit - 5.6 PROBABILITY DISTRIBUTION

5.6.1 Binomial Distribution:

What is Binomial Distribution? Binomial distribution is a common probability distribution that models the probability of obtaining one of two outcomes under a given number of parameters. It summarizes the number of trials when each trial has the same chance of attaining one specific outcome. The value of binomial is obtained by multiplying the number of independent trials by the successes.

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one. However, for N much larger than n, the binomial distribution remains a good approximation, and is widely used.

Bernoulli distribution: The Bernoulli distribution is a special case of the binomial distribution, where $n = 1$. Symbolically, $X \sim B(1, p)$ has the same meaning as $X \sim \text{Bernoulli}(p)$. Conversely, any binomial distribution, $B(n, p)$, is the distribution of the sum of n Bernoulli trials, $\text{Bernoulli}(p)$, each with the same probability p.

Poisson approximation: The binomial distribution converges towards the Poisson distribution as the number of trials goes to infinity while the product np remains fixed or at least p tends to zero. Therefore, the Poisson distribution with parameter $\lambda = np$ can be used as an approximation to $B(n, p)$ of the binomial distribution if n is sufficiently large and p is sufficiently small. According to two rules of thumb, this approximation is good if $n \geq 20$ and $p \leq 0.05$, or if $n \geq 100$ and $np \leq 10$.

Normal approximation: If n is large enough, then the skew of the distribution is not too great. In this case a reasonable approximation to $B(n, p)$ is given by the normal distribution and this basic approximation can be improved in a simple way by using a suitable continuity correction. The basic approximation generally improves as n increases (at least 20) and is better when p is not near to 0 or 1. Various rules of thumb may be used to decide whether n is large enough, and p is far enough from the extremes of zero or one:

Criteria of Binomial Distribution: Binomial distribution models the probability of occurrence of an event when the specific criteria are met. Binomial distribution involves the following rules that must be present in the process in order to use the binomial probability formula:

1. Fixed trials: The process under investigation must have a fixed number of trials that cannot be altered in the course of the analysis. During the analysis, each trial must be performed in a uniform manner, although each trial may yield a different outcome.

In the binomial probability formula, the number of trials is represented by the letter “n.” An example of a fixed trial may be coin flips, free throws, wheel spins, etc. The number of times that each trial is conducted is known from the start. If a coin is flipped 10 times, each flip of the coin is a trial.

2. Independent trials: The other condition of a binomial probability is that the trials are independent of each other. In simple terms, the outcome of one trial should not affect the outcome of the subsequent trials.

When using certain sampling methods, there is a possibility of having trials that are not completely independent of each other, and binomial distribution may only be used when the size of the population is large vis-a-vis the sample size.

An example of independent trials may be tossing a coin or rolling a dice. When tossing a coin, the first event is independent of the subsequent events.

3. Fixed probability of success: In a binomial distribution, the probability of getting a success must remain the same for the trials we are investigating. For example, when tossing a coin, the probability of flipping a coin is ½ or 0.5 for every trial we conduct, since there are only two possible outcomes.

In some sampling techniques like sampling without replacement, the probability of success from each trial may vary from one trial to the other. For example, assume that there are 50 boys in a population of 1,000 students. The probability of picking a boy from that population is 0.05. In the next trial, there will be 49 boys out of 999 students. The probability of picking a boy in the next trial is 0.049. It shows that in subsequent trials, the probability from one trial to the next will vary slightly from the prior trial.

4. Two mutually exclusive outcomes: In binomial probability, there are only two mutually exclusive outcomes, i.e., success or failure. While success is generally a positive term, it can be used to mean that the outcome of the trial agrees with what you have defined as a success, whether it is a positive or negative outcome.

For example, when a business receives a consignment of lamps with a lot of breakages, the business can define success for the trial to be every lamp that has broken glass. A failure can be defined as when the lamps have zero broken glasses.

In our example, the instances of broken lamps may be used to denote success as a way of showing that a high proposition of the lamps in the consignment is broken. and that there is a low probability of getting a lamp with zero breakages.

Properties of binomial distribution:

1. Binomial distribution is applicable when the trials are independent and each trial has just two outcomes success and failure.

It is applied in coin tossing experiments, sampling inspection plan, genetic experiments and so on.

2. Binomial distribution is known as bi-parametric distribution as it is characterized by two parameters n and p.

This means that if the values of n and p are known, then the distribution is known completely.

3. The mean of the binomial distribution is given by $\mu = np$

4. Depending on the values of the two parameters, binomial distribution may be uni-modal or bi-modal. To know the mode of binomial distribution, first we have to find the value of (n+1)p. **(n+1) p is a non integer --------> Uni-modal**

Here, the mode = the largest integer contained in (n+1)p. **(n+1)p is a integer --------> Bi-modal.** Here, the mode = (n+1|)p, (n+1)p - 1

5. The variance of the binomial distribution is given by $\sigma^2 = npq$

6. Since p and q are numerically less than or equal to 1, **npq < np**

That is, variance of a binomial variable is always less than its mean.

7. Variance of binomial variable X attains its maximum value at p = q = 0.5 and this maximum value is n/4.

8. Additive property of binomial distribution. Let X and Y be the two independent binomial variables. X is having the parameters n_1 and p and Y is having the parameters n_2 and p. Then (X+Y) will also be a binomial variable with the parameters ($n_1 + n_2$) and p

5.6.2 Poisson Distribution: In probability theory, the Poisson distribution is a very common discrete probability distribution. A Poisson distribution helps in describing the chances of occurrence of a number of events in some given time interval or given space conditionally that the value of average number of occurrence of the event is known. This is a major and only condition of Poisson distribution.

The distribution that arises from the Poisson experiment is termed as Poisson distribution. The number of successes that are resulting from a Poisson experiment is termed as a Poisson random variable.

Cumulative Poisson Probability: A **cumulative Poisson probability** refers to the probability that the Poisson random variable is greater than some specified lower limit and less than some specified upper limit.

Condition for Poisson distribution: Poisson distribution is the limiting case of binomial distribution under the following assumptions.

1. The number of trials n should be indefinitely large ie., $n \to \infty$
2. The probability of success p for each trial is indefinitely small.
3. $np = \lambda$, should be finite where λ is constant.

Properties

1. Poisson distribution is defined by single parameter λ.
2. Mean = λ
3. Variance = λ. Mean and Variance are equal.

Application of Poisson distribution

1. It is used in quality control statistics to count the number of defects of an item.
2. In biology, to count the number of bacteria.
3. In determining the number of deaths in a district in a given period, by rare disease.
4. The number of error per page in typed material.
5. The number of plants infected with a particular disease in a plot of field.
6. Number of weeds in particular species in different plots of a field.

5.6.3 Normal Distribution: Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

- A normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- In reality, most pricing distributions are not perfectly normal.

The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

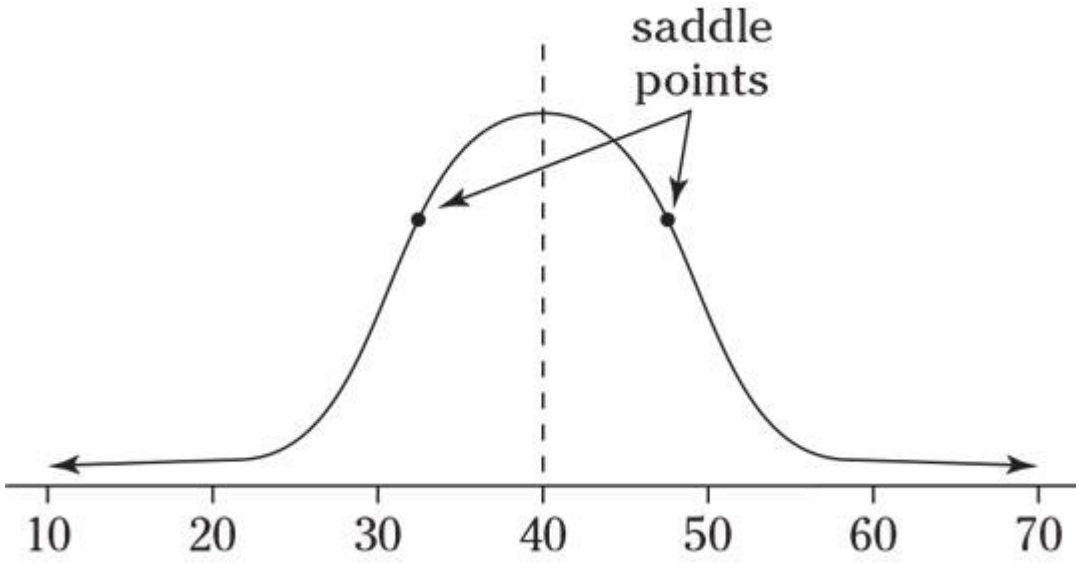
Condition of Normal Distribution

- Normal distribution is a limiting form of the binomial distribution under the following conditions.
  a) n, the number of trials is indefinitely large ie., nà¥ and
  b) Neither p nor q is very small.
- Normal distribution can also be obtained as a limiting form of Poisson distribution with parameter mà¥
- Constants of normal distribution are mean = m, variation =s2 , Standard deviation = s

Normal probability curve: The curve representing the normal distribution is called the normal probability curve. The curve is symmetrical about the mean (m), bell-shaped and the two tails on the right and left sides of the mean extends to the infinity.

Properties of normal distribution

- The normal curve is bell shaped in appearance.
- There is one maximum point of normal curve which occur at mean.
- As it has only one maximum curve so it is unimodal.
- In binomial and possion distribution the variable is discrete while in this it is continuous.
- Here mean= median =mode.
- The total area of normal curve is 1. The area to the left and the area to the right of the curve is 0.5.
- No portion of curve lies below x-axis so probability can never be negative.
- The curve becomes parallel to x-axis which is supposed to meet it at infinity.
- Here mean deviation = 4/5 standard deviation.
- Quartile deviation = 5/6 mean deviation
- Quartile deviation : mean deviation : standard deviation
  10 : 12 : 15
- 4 standard deviation = 5 mean deviation = 6 quartile deviation
- The standard deviation is the distance from the center to the *saddle point* (the place where the curve changes from an “upside-down-bowl” shape to a “right-side-up-bowl” shape. (The standard deviation of the population is designated by the Greek letter σ.)



- About 68 percent of the values lie within one standard deviation of the mean, about 95 percent lie within two standard deviations, and most of the values (99.7 percent or more) lie within three standard deviations by the empirical rule.
- Each normal distribution has a different mean and standard deviation that make it look a little different from the rest, yet they all have the same bell shape.

Shape of Normal Distribution: A normal distribution is symmetric from the peak of the curve, where the mean is. It means that most of the observed data is clustered near the mean, while the data becomes less frequent when it is far away from the mean. When plotted on a graph, the resultant graph appears as bell-shaped where the mean, median and mode are of the same values and appear at the peak of the curve.

The graph is a perfect symmetry such that, if you fold it at the middle, you will get two equal halves since one-half of the observable data points fall on each side of the graph.

Parameters of Normal Distribution: The two main parameters of a (normal) distribution are the mean and standard deviation. The parameters determine the shape and probabilities of the distribution. The shape of the distribution changes as the parameter values.

1. Mean: The mean is used by researchers as a measure of central tendency. It can be used to describe the distribution of variables measured as ratios or intervals. In a normal distribution graph, the mean defines the location of the peak, and most of the data points are clustered around the mean. Any changes made to the value of the mean move the curve either to the left or right along the X-axis.

2. Standard Deviation: The standard deviation measures the dispersion of the data points relative to the mean. It determines how far away from the mean the data points are positioned and represents the distance between the mean and the observations.

On the graph, the standard deviation determines the width of the curve, and it tightens or expands the width of the distribution along the x-axis. Typically, a small standard deviation relative to the mean produces a steep curve, while a large standard deviation relative to the mean produces a flat curve.

Importance of Normal Distribution

1) It has one of the important properties called central theorem. Central theorem means relationship between shape of population distribution and shape of sampling distribution of mean. This means that sampling distribution of mean approaches normal as sample size increase.

2) In case the sample size is large the normal distribution serves as good approximation.

3) Due to its mathematical properties it is more popular and easier to calculate.

4) It is used in statistical quality control in setting up of control limits.

5) The whole theory of sample tests t, f and chi-square test is based on the normal distribution.

Sub Unit - 5.7: RESEARCH

5.7.1 Concept and Types

Concept of Research: According to the American sociologist Earl Robert Babbie, "Research is a systematic inquiry to describe, explain, predict, and control the observed phenomenon. Research involves inductive and deductive methods."

Inductive research methods are used to analyze an observed event. Deductive methods are used to verify the observed event. Inductive approaches are associated with qualitative research and deductive methods are more commonly associated with quantitative research.

Research is conducted with a purpose to understand:

- What do organizations or businesses really want to find out?
- What are the processes that need to be followed to chase the idea?
- What are the arguments that need to be built around a concept?
- What is the evidence that will be required for people to believe in the idea or concept?

Characteristics of Research

- A systematic approach must be followed for accurate data. Rules and procedures are an integral part of the process that set the objective. Researchers need to practice ethics and a code of conduct while making observations or drawing conclusions.
- Research is based on logical reasoning and involves both inductive and deductive methods.
- The data or knowledge that is derived is in real time from actual observations in natural settings.
- There is an in-depth analysis of all data collected so that there are no anomalies associated with it.
- Research creates a path for generating new questions. Existing data helps create more opportunities for research.
- Research is analytical in nature. It makes use of all the available data so that there is no ambiguity in inference.
- Accuracy is one of the most important aspects of research. The information that is obtained should be accurate and true to its nature. For example, laboratories provide a controlled environment to collect data. Accuracy is measured in the instruments used, the calibrations of instruments or tools, and the final result of the experiment.

Types of Research

- **Basic Research:** A basic research definition is data collected to enhance knowledge. The main motivation is knowledge expansion. It is a non-commercial research that doesn't facilitate in creating or inventing anything. For example: an experiment to determine a simple fact.
- **Applied Research:** Applied research focuses on analyzing and solving real-life problems. This type refers to the study that helps solve practical problems using scientific methods. Studies play an important role in solving issues that impact the overall well-being of humans. For example: finding a specific cure for a disease.

- **Problem Oriented Research:** As the name suggests, problem-oriented research is conducted to understand the exact nature of a problem to find out relevant solutions. The term "problem" refers to multiple choices or issues when analyzing a situation.
- **Problem Solving Research**: This type of research is conducted by companies to understand and resolve their own problems. The problem-solving method uses applied research to find solutions to the existing problems.
- **Qualitative Research:** Qualitative research is a process that is about inquiry. It helps create in-depth understanding of problems or issues in their natural settings. This is a non-statistical method.

  Qualitative research is heavily dependent on the experience of the researchers and the questions used to probe the sample. The sample size is usually restricted to 6-10 people. Open-ended questions are asked in a manner that encourages answers that lead to another question or group of questions. The purpose of asking open-ended questions is to gather as much information as possible from the sample.

  The following are the methods used for qualitative research:

  - One-to-one interview
  - Focus groups
  - Ethnographic research
  - Content/Text Analysis
  - Case study research

- **Quantitative Research:** Qualitative research is a structured way of collecting data and analyzing it to draw conclusions. Unlike qualitative methods, this method uses a computational and statistical process to collect and analyze data. Quantitative data is all about numbers.

Quantitative research involves a larger population — more people means more data. With more data to analyze, you can obtain more accurate results. This method uses close-ended questions because the researchers are typically looking to gather statistical data.

Online surveys, questionnaires, and polls are preferable data collection tools used in quantitative research. There are various methods of deploying surveys or questionnaires.

Online surveys allow survey creators to reach large amounts of people or smaller focus groups for different types of research that meet different goals. Survey respondents can receive surveys on mobile phones, in emails, or can simply use the internet to access surveys.

Types of research methods and research example

Qualitative Methods: Qualitative research is a method that collects data using conversational methods. Participants are asked open-ended questions. The responses collected are essentially non-numerical. This method not only helps a researcher understand what participants think but also why they think in a particular way.

Types of qualitative methods include:

- **One-to-one Interview:** This interview is conducted with one participant at a given point in time. One-to-one interviews need a researcher to prepare questions in advance. The researcher asks only the most important questions to the participant. This type of interview lasts anywhere between 20 minutes to half an hour. During this time the researcher collects as many meaningful answers as possible from the participants to draw inferences.
- **Focus Groups:** Focus groups are small groups comprising of around 6-10 participants who are usually experts in the subject matter. A moderator is assigned to a focus group who facilitates the discussion amongst the group members. A moderator's experience in conducting the focus group plays an important role. An experienced moderator can probe the participants by asking the correct questions that will help them collect a sizable amount of information related to the research.
- **Ethnographic Research:** Ethnographic research is an in-depth form of research where people are observed in their natural environment without This method is demanding due to the necessity of a researcher entering a natural environment of other people. Geographic locations can be a constraint as well. Instead of conducting interviews, a researcher experiences the normal setting and daily life of a group of people.

- **Text Analysis:** Text analysis is a little different from other qualitative methods as it is used to analyze social constructs by decoding words through any available form of documentation. The researcher studies and understands the context in which the documents are written and then tries to draw meaningful inferences from it. Researchers today follow activities on a social media platform to try and understand patterns of thoughts.
- **Case Study:** Case study research is used to study an organization or an entity. This method is one of the most valuable options for modern This type of research is used in fields like the education sector, philosophical studies, and psychological studies. This method involves a deep dive into ongoing research and collecting data.

Quantitative Research Methods: Quantitative methods deal with numbers and measurable forms. It uses a systematic way of investigating events or data. It is used to answer questions in terms of justifying relationships with measurable variables to explain, predict, or control a phenomenon.

There are three methods that are often used by researchers:

- **Survey Research** — The ultimate goal of survey research is to learn about a large population by deploying a survey. Today, online surveys are popular as they are convenient and can be sent in an email or made available on the internet. In this method, a researcher designs a survey with the most relevant survey questions and distributes the survey. Once the researcher receives responses, they summarize them to tabulate meaningful findings and data.
- **Descriptive Research** — Descriptive research is a method which identifies the characteristics of an observed phenomenon and collects more information. This method is designed to depict the participants in a very systematic and accurate manner. In simple words, descriptive research is all about describing the phenomenon, observing it, and drawing conclusions from it.
- **Correlational Research**— Correlational research examines the relationship between two or more variables. Consider a researcher is studying a correlation between cancer and married Married women have a negative correlation with cancer. In this example, there are two variables: cancer and married women. When we say negative correlation, it means women who are married are less likely to develop cancer. However, it doesn’t mean that marriage directly avoids cancer.

5.7.2 Research Designs

Definition: Research design is the framework of research methods and techniques chosen by a researcher. The design allows researchers to hone in on research methods that are suitable for the subject matter and set up their studies up for success.

The design of a research topic explains the type of research (experimental, survey, correlational, semi-experimental, review) and also its sub-type (experimental design, research problem, descriptive case-study).

There are three main types of research design: Data collection, measurement, and analysis.

The type of research problem an organization is facing will determine the research design and not vice-versa. The design phase of a study determines which tools to use and how they are used.

An impactful research design usually creates a minimum bias in data and increases trust in the accuracy of collected data. A design that produces the least margin of error in experimental research is generally considered the desired outcome. The essential elements of the research design are:

- Accurate purpose statement
- Techniques to be implemented for collecting and analyzing research
- The method applied for analyzing collected details
- Type of research methodology
- Probable objections for research
- Settings for the research study
- Timeline
- Measurement of analysis

Proper research design sets your study up for success. Successful research studies provide insights that are accurate and unbiased. You'll need to create a survey that meets all of the main characteristics of a design.

Characteristics of Research Design: There are four key characteristics of research design:

- **Neutrality:** When you set up your study, you may have to make assumptions about the data you expect to collect. The results projected in the research design should be free from bias and neutral. Understand opinions about the final evaluated scores and conclusion from multiple individuals and consider those who agree with the derived results.
- **Reliability:** With regularly conducted research, the researcher involved expects similar results every time. Your design should indicate how to form research questions to ensure the standard of results. You'll only be able to reach the expected results if your design is reliable.
- **Validity:** There are multiple measuring tools available. However, the only correct measuring tools are those which help a researcher in gauging results according to the objective of the research. The questionnaire developed from this design will then be valid.

- **Generalization:** The outcome of your design should apply to a population and not just a restricted sample. A generalized design implies that your survey can be conducted on any part of a population with similar accuracy.

The above factors affect the way respondents answer the research questions and so all the above characteristics should be balanced in a good design.

A researcher must have a clear understanding of the various types of research design to select which model to implement for a study. Like research itself, the design of your study can be broadly classified into quantitative and qualitative.

Qualitative research design: Qualitative research determines relationships between collected data and observations based on mathematical calculations. Theories related to a naturally existing phenomenon can be proved or disproved using statistical methods. Researchers rely on qualitative research design methods that conclude "why" a particular theory exists along with "what" respondents have to say about it.

Quantitative research design: Quantitative research is for cases where statistical conclusions to collect actionable insights are essential. Numbers provide a better perspective to make critical business decisions. Quantitative research design methods are necessary for the growth of any organization. Insights drawn from hard numerical data and analysis prove to be highly effective when making decisions related to the future of the business.

You can further break down the types of research design into five categories:

1. Descriptive research design: In a descriptive design, a researcher is solely interested in describing the situation or case under their research study. It is a theory-based design method which is created by gathering, analyzing, and presenting collected data. This allows a researcher to provide insights into the why and how of research. Descriptive design helps others better understand the need for the research. If the problem statement is not clear, you can conduct exploratory research.

2. Experimental research design: Experimental research design establishes a relationship between the cause and effect of a situation. It is a causal design where one observes the impact caused by the independent variable on the dependent variable. For example, one monitors the influence of an independent variable such as a price on a dependent variable such as customer satisfaction or brand loyalty. It is a highly practical research design method as it contributes to solving a problem at hand. The independent variables are manipulated to monitor the change it has on the dependent variable. It is often used in social sciences to observe human behavior by analyzing two groups. Researchers can have participants change their actions and study how the people around them react to gain a better understanding of social psychology.

3. Correlational research design: Correlational research is a non-experimental research design technique that helps researchers establish a relationship between two closely connected variables. This type of research requires two different groups. There is no assumption while evaluating a relationship between two different variables, and statistical analysis techniques calculate the relationship between them.

A correlation coefficient determines the correlation between two variables, whose value ranges between -1 and +1. If the correlation coefficient is towards +1, it indicates a positive relationship between the variables and -1 means a negative relationship between the two variables.

4. Diagnostic research design: In diagnostic design, the researcher is looking to evaluate the underlying cause of a specific topic or phenomenon. This method helps one learn more about the factors that create troublesome situations.

This design has three parts of the research:

· Inception of the issue

· Diagnosis of the issue

· Solution for the issue

5. Explanatory research design: Explanatory design uses a researcher's ideas and thoughts on a subject to further explore their theories. The research explains unexplored aspects of a subject and details about what, how, and why of research questions.

Sub Unit - 5.8: DATA

5.8.1 Collection of Data

Definition of Data Collection: Data collection is defined as the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information.

The most critical objective of data collection is ensuring that information-rich and reliable data is collected for statistical analysis so that data-driven decisions can be made for research.

Data Collection Methods:

1. Closed-ended Surveys and Online Quizzes: Closed-ended surveys and online quizzes are based on questions that give respondents predefined answer options to opt for. There are two main types of closed-ended surveys – those based on categorical and those based on interval/ratio questions.

Categorical survey questions can be further classified into dichotomous ('yes/no'), multiple-choice questions, or checkbox questions and can be answered with a simple "yes" or "no" or a specific piece of predefined information.

Interval/ratio questions, on the other hand, can consist of rating-scale, Likert-scale, or matrix questions and involve a set of predefined values to choose from on a fixed scale. To learn more, we have prepared a guide on different types of closed-ended survey questions.

Qualitative data collection methods

2. Open-Ended Surveys and Questionnaires: Opposite to closed-ended are open-ended surveys and questionnaires. The main difference between the two is the fact that closed-ended surveys offer predefined answer options the respondent must choose from, whereas open-ended surveys allow the respondents much more freedom and flexibility when providing their answers.

Compared to closed-ended surveys, one of the quantitative data collection methods, the findings of open-ended surveys are more difficult to compile and analyze due to the fact that there are no uniform answer options to choose from.

3. 1-on-1 Interviews: One-on-one (or face-to-face) interviews are one of the most common types of data collection methods in qualitative research. Here, the interviewer collects data directly from the interviewee. Due to it being a very personal approach, this data collection technique is perfect when you need to gather highly-personalized data.

Depending on your specific needs, the interview can be informal, unstructured, conversational, and even spontaneous (as if you were talking to your friend) – in which case it's more difficult and time-consuming to process the obtained data – or it can be semi-structured and standardized to a certain extent (if you, for example, ask the same series of open-ended questions).

4. Focus groups: The focus groups data collection method is essentially an interview method, but instead of being done 1-on-1, here we have a group discussion.

Whenever the resources for 1-on-1 interviews are limited (whether in terms of people, money, or time) or you need to recreate a particular social situation in order to gather data on people's attitudes and behaviors, focus groups can come in very handy.

Ideally, a focus group should have 3-10 people, plus a moderator. Of course, depending on the research goal and what the data obtained is to be used for, there should be some common denominators for all the members of the focus group.

5. Direct observation: Direct observation is one of the most passive qualitative data collection methods. Here, the data collector takes a participatory stance, observing the setting in which the subjects of their observation are while taking down notes, video/audio recordings, photos, and so on.

Due to its participatory nature, direct observation can lead to bias in research, as the participation may influence the attitudes and opinions of the researcher, making it challenging for them to remain objective. Plus, the fact that the researcher is a participant too can affect the naturalness of the actions and behaviors of subjects who know they're being observed.

5.8.2 Classification of Data: Data classification is the process of organizing data into categories that make it is easy to retrieve, sort and store for future use.

A well-planned data classification system makes essential data easy to find and retrieve. This can be of particular importance for risk management, legal discovery and compliance. Written procedures and guidelines for data classification policies should define what categories and criteria the organization will use to classify data and specify the roles and responsibilities of employees within the organization regarding data stewardship. Once a data-classification scheme has been created, security standards that specify appropriate handling practices for each category and storage standards that define the data's lifecycle requirements need to be addressed.

Purpose of Data Classification: On top of making data easier to locate and retrieve, a carefully planned data classification system also makes essential data easy to manipulate and track. While some combination of all of the following attributes may be achieved, most businesses and data professionals focus on a particular goal when they approach a data classification project. The most common goals include but are not limited to the following:

- **Confidentiality.** A classification system that values confidentiality above other attributes will mostly focus on security measures, including user permissions and encryption.
- **Integrity of data.** A system that focuses on data integrity will require more storage, user permissions and proper channels of access.
- **Availability of data.** When security and integrity do not need to be perfected, it is easiest to make data more easily accessible to users.

Common steps of Data Classification: Most commonly, not all data needs to be classified, and some is even better destroyed. It is important to begin by prioritizing which types of data need to go through the classification and reclassification processes.

Next, data scientists and other professionals create a framework within which to organize the data. They assign metadata or other tags to the information, which allow machines and software to instantly sort it in different groups and categories. It is important to maintain at every step that all data classification schemes adhere to company policies as well as local and federal regulations around the handling of the data.

In addition, companies need to always consider the ethical and privacy practices that best reflect their standards and the expectations of clients and customers:

- **Scan.** This step involves taking stock of an entire database and making a digital game plan to tackle the organization process.
- **Identify.** Anything from file type to character units to size of packets of data may be used to sort the information into searchable, sortable categories.
- **Separate.** Once the data is categorized with a system the data science professional implements, it can be separated by those categories whenever the system is called to bring them up.

Unauthorized disclosure of information that falls within one of the protected categories of a company's data classification systems is likely a breach of protocol and, in some countries, may even be considered a serious crime. In order to enforce proper protocols, the protected data needs to first be sorted into its category of sensitivity.

Data classification can be used to further categorize structured data, but it is an especially important process for getting the most out of unstructured data by maximizing its usefulness for an organization.

Types of Data Classification: In computer programming, file parsing is a method of splitting packets of information into smaller sub-packets, making them easier to move, manipulate and categorize or sort. Different parsing styles help a system to determine what kind of information is input. For instance, dates are split up by day, month or year, and words may be separated by spaces.

Within data classification, there are many kinds of intervals that can be applied, including but not limited to the following:

- **Manual intervals.** Using manual intervals involves a human going through the entire data set and entering class breaks by observing where they make the most sense. This is a perfectly fine system for smaller data sets, but may prove problematic for larger collections of information.
- **Defined intervals.** Defined intervals specify a number of characters to include in a packet. For example, information might be broken into smaller packets every three units.
- **Equal intervals.** Equal intervals divide an entire data set into a specified number of groups, distributing the amount of information over those groups evenly.
- **Quantiles.** Using quantiles involves setting a number of data values allowed per class type.

- **Natural breaks.** Programs are able to determine wherever large changes in the data occur on their own and use those indicators as a way of determining where to break up the data.
- **Geometric intervals.** For geometric intervals, the same number of units is allowed per class category.
- **Standard deviation intervals.** These are determined by how much the attributes of an entry differ from the norm. There are set number values to show each entry's deviations.
- **Custom ranges.** Custom ranges can be created and set by a user and changed at any point.

Classification is an important part of data management that varies slightly from data characterization. Classification is all about sorting information and data, while categorization involves the actual systems that hold that information and data.

There are certain data classification standard categories. Each one of these standards may have federal and local laws about how they need to be handled. They inlcude the following:

- **Public information.** This standard is maintained by state institutions and subject to disclosure as part of certain laws.
- **Confidential information.** This may have legal restrictions about the way it is handled, or there may be other consequences around the way it is handled.
- **Sensitive information.** This is any information stored or handled by state institutions that include authorization requirements and other rigid rules around its use.
- **Personal information.** Generally, peoples' personal information is considered protected by law, and it needs to be handled following certain protocols and rules for proper use. Sometimes there are gaps between the moral requirements and contemporary legislative protections for their use.

Steps for Effective Data Classification

- **Understand the Current Setup:** Taking a detailed look at the location of current data and all regulations that pertain to your organization is perhaps the best starting point for effectively classifying data. You must know what data you have before you can classify it.
- **Creating a Data Classification Policy:** Staying compliant with data protection principles in an organization is nearly impossible without proper policy. Creating a policy should be your top priority.
- **Prioritize and Organize Data:** Now that you have a policy and a picture of your current data, it’s time to properly classify the data. Decide on the best way to tag your data based on its sensitivity and privacy.

There are **more benefits** to data classification than simply making data easier to find. Data classification is necessary to enable modern enterprises to make sense of the vast amounts of data available at any given moment.

Data classification provides a clear picture of all data within an organization’s control and an understanding of where data is stored, how to easily access it, and the best way to protect it from potential security risks. Once implemented, data classification provides an organized framework that facilitates more adequate data protection measures and promotes employee compliance with security policies.

Sub Unit - 5.9: SAMPLING AND ESTIMATION

5.9.1 Concept of Sampling and Estimation

Concept of Sampling: Sampling is the process of converting continuous signal to discrete form. Ex- Conversion of a sound wave into sequence of samples. A sample denotes a set of values at a point in time and/or space. A sampler is a subsystem that extracts samples from a continuous signal. Its application includes audio sampling, sampling rate, quantization, speech sampling, video sampling. Also there are other types of sampling such as Oversampling, Undersampling, Complex sampling.

Concept of Estimation: The procedure of making judgment or decision about a population parameter is referred to as statistical estimation or simply estimation. Statistical estimation procedures provide estimates of population parameter with a desired degree of confidence. The degree of confidence can be controlled in part,

- by the size the sample (larger sample greater accuracy of the estimate) and
- by the type of the estimate made. Population parameters are estimated from sample data because it is not possible (it is impracticable) to examine the entire population in order to make such an exact determination.

The statistical estimation of the population parameter is further divided into two types,

(i) Point Estimation, and

(ii) Interval Estimation

Point Estimation: The objective of point estimation is to obtain a single number from the sample which will represent the unknown value of the population parameter. Population parameters (population mean, variance, etc) are estimated from the corresponding sample statistics (sample mean, variance, etc).

A statistic used to estimate a parameter is called a point estimator or simply an estimator, the actual numerical value obtained by estimator is called an estimate.

Interval Estimation: A point estimator (such as sample mean) calculated from the sample data provides a single number as an estimate of the population parameter, which can not be expected to be exactly equal to the population parameter because the mean of a sample taken from a population may assume different values for different samples. Therefore we estimate an interval/ range of values (set of values) within which the population parameter is expected to lie with a certain degree of confidence. This range of values used to estimate a population parameter is known as interval estimate or estimate by a confidence interval, and is defined by two numbers, between which a population parameter is expected to lie.

5.9.2 Methods of Sampling – Probability and Non-probability Methods

Probability Sampling: A **probability sampling** scheme is one in which each unit in the population has a chance (greater than zero) of being selected in the sample, and this possibility can be accurately determined.

The combination of these behaviors makes it possible to produce unbiased estimations of population totals, by weighting sampled units rendering to their probability of selection.

Probability sampling may be of the following types:

- **Simple Random Sampling:** In a simple random sample ('SRS') of a given size, all such subsets of the frame are given an equal probability. Each component of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned. This minimizes bias and simplifies analysis of results.
- **Systematic sampling:** Systematic sampling depend on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards.
  In this case, k=(population size/sample size).
  It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list.
  A simple example would be to select every 10^{th} name from the telephone directory (an 'every 10^{th}' sample, also referred to as 'sampling with a skip of 10').
- **Stratified Sampling:** The sampling where the population embraces several distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected. Dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.
- **Cluster Sampling:** It is an example of 'two-stage sampling' or 'multistage sampling': in the first stage a sample of areas is chosen; in the second stage a sample of respondents within those areas is selected.
- **Multistage Sampling:** Multistage sampling is a complex form of cluster sampling in which two or more levels of units are embedded one in the other. The first stage consists of constructing the clusters that will be used to sample from. In the second stage, a sample of primary units is randomly selected from each cluster (rather than using all units contained in all selected clusters). In following stages, in each of those selected clusters, additional samples of units are selected, and so on. All ultimate units (individuals, for instance) selected at the last step of this procedure are then surveyed. This technique, thus, is essentially the process of taking random samples of preceding random samples. It is not as effective as true random sampling, but it probably solves more of the problems inherent to random sampling. Moreover, It is an effective strategy because it banks on multiple randomizations. As such, it is extremely useful. Multistage sampling is used frequently when a complete list of all members of the population does not exist and is inappropriate. Moreover, by avoiding the use of all sample units in all selected clusters, multistage sampling avoids the large, and perhaps unnecessary, costs associated traditional cluster sampling.

Non-Probability Sampling: Non-probability sampling is any sampling technique where some elements of the population have no definite chance of selection, or where the probability of selection can't be correctly determined.

Non-probability sampling may be of the following types:

- Quota Sampling: In quota sampling the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example, interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for many years.

- Convenience Sampling (grab or opportunity sampling): Convenience sampling is a type of non-probability sampling which involves the sample being drawn from that part of the population which is close to hand? That is, a sample population selected because it is readily available and convenient. It may be through meeting the person or including a person in the sample when one meets them or chosen by finding them through technological means such as the internet or through phone. The researcher using such a sample cannot scientifically generalize about the total population from this sample because it would not be representative enough. For example, if the interviewer was to conduct such a survey at a shopping centre early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey was to be conducted at different times of day and several times per week. This type of sampling is most useful for pilot testing.

5.9.3 Sampling Distribution

What is a Sampling Distribution? A sampling distribution is a graph of a statistic for your sample data. While, technically, you could choose any statistic to paint a picture, some common ones you'll come across are:

- Mean
- Mean absolute value of the deviation from the mean
- Range
- Standard deviation of the sample
- Unbiased estimate of variance
- Variance of the sample

Sampling Distribution of the Mean: Suppose we draw all possible samples of size *n* from a population of size *N*. Suppose further that we compute a mean score for each sample. In this way, we create a sampling distribution of the mean.

We know the following about the sampling distribution of the mean. The mean of the sampling distribution (μ_x) is equal to the mean of the population (μ). And the standard error of the sampling distribution (σ_x) is determined by the standard deviation of the population (σ), the population size (N), and the sample size (n).

Sampling Distribution of the Proportion: In a population of size *N*, suppose that the probability of the occurrence of an event (dubbed a "success") is P; and the probability of the event's non-occurrence (dubbed a "failure") is Q. From this population, suppose that we draw all possible samples of size *n*. And finally, within each sample, suppose that we determine the proportion of successes *p* and failures *q*. In this way, we create a sampling distribution of the proportion.

5.9.4 Central Limit Theorem: The **central limit theorem** states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough.

How large is "large enough"? The answer depends on two factors.

- Requirements for accuracy. The more closely the sampling distribution needs to resemble a normal distribution, the more sample points will be required.
- The shape of the underlying population. The more closely the original population resembles a normal distribution, the fewer sample points will be required.

In practice, some statisticians say that a sample size of 30 is large enough when the population distribution is roughly bell-shaped. Others recommend a sample size of at least 40. But if the original population is distinctly not normal (e.g., is badly skewed, has multiple peaks, and/or has outliers), researchers like the sample size to be even larger.

5.9.5 Standard Error: The Standard Error (SE) is very similar to standard deviation. Both are measures of spread. The higher the number, the more spread out your data is. To put it simply, the two terms are essentially equal — but there is one important difference. While the standard error uses **statistics** (sample data) standard deviations use **parameters** (population data). (What is the difference between a statistic and a parameter?).

In statistics, you'll come across terms like "the standard error of the mean" or "the standard error of the median." The SE tells you how far your sample statistic (like the sample mean) deviates from the actual population mean. The larger your sample size, the smaller the SE. In other words, the larger your sample size, the closer your sample mean is to the actual population mean.

5.9.6 Statistical Estimation: Statistical inference is the process of making judgment about a population based on sampling properties. An important aspect of statistical inference is using **estimates** to approximate the value of an unknown population parameter. Another type of inference involves choosing between two opposing views or statements about the population; this process is called **hypothesis testing**.

Point Estimate vs. Interval Estimate: Statisticians use sample statistics to estimate population parameters. For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

An estimate of a population parameter may be expressed in two ways:

Point Estimate: A point estimate of a population parameter is a single value of a statistic.

Interval Estimate: An interval estimate is defined by two numbers, between which a population parameter is said to lie.

Confidence Intervals: Statisticians use a confidence interval to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

- A confidence levels.
- A statistic.
- A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic + margin of error*. Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Confidence Level: The probability part of a confidence interval is called a **confidence level**. The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

Here is how to interpret a confidence level. Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

Margin of Error: In a confidence interval, the range of values above and below the sample statistic is called the **margin of error**.

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Sub Unit - 5.10: HYPOTHESIS TESTING

A statistical hypothesis is an assumption about a population which may or may not be true. Hypothesis testing is a set of formal procedures used by statisticians to either accept or reject statistical hypotheses. Statistical hypotheses are of two types:

- **Null hypothesis, *H*0** - represents a hypothesis of chance basis.
- **Alternative hypothesis, *Ha*** - represents a hypothesis of observations which are influenced by some non-random cause.

Hypothesis Tests: Following formal process is used by statistician to determine whether to reject a null hypothesis, based on sample data. This process is called hypothesis testing and is consists of following **four steps:**

5. **State the hypotheses** - This step involves stating both null and alternative hypotheses. The hypotheses should be stated in such a way that they are mutually exclusive. If one is true then other must be false.
6. **Formulate an analysis plan** - The analysis plan is to describe how to use the sample data to evaluate the null hypothesis. The evaluation process focuses around a single test statistic.
7. **Analyze sample data** - Find the value of the test statistic (using properties like mean score, proportion, t statistic, z-score, etc.) stated in the analysis plan.
8. **Interpret results** - Apply the decisions stated in the analysis plan. If the value of the test statistic is very unlikely based on the null hypothesis, then reject the null hypothesis.

5.10.1 z – test: A ***Z*-test** is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Z-test tests the mean of a distribution in which we already know the population variance σ^2 . Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. For each significance level in the confidence interval, the *Z*-test has a single critical value (for example, 1.96 for 5% two tailed) which makes it more convenient than the Student's *t*-test which has separate and different critical values for each sample size (for different sample size, it would have different degree of freedom, which may determine the value of the critical values). Therefore, many statistical tests can be conveniently performed as approximate *Z*-tests if the sample size is large or the population variance is known. If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large ($n < 30$), the Student's *t*-test may be more appropriate.

How to perform a Z test when *T* is a statistic that is approximately normally distributed under the null hypothesis is as follows:

First, estimate the expected value μ of *T* under the null hypothesis, and obtain an estimate *s* of the standard deviation of *T*.

Second, determine the properties of *T* : one tailed or two tailed.

For Null hypothesis H_0: $\mu \geq \mu_0$ vs alternative hypothesis H_1: $\mu < \mu_0$, it is upper/left-tailed (one tailed).
For Null hypothesis H_0: $\mu \leq \mu_0$ vs alternative hypothesis H_1: $\mu > \mu_0$, it is lower/right-tailed (one tailed).
For Null hypothesis H_0: $\mu = \mu_0$ vs alternative hypothesis H_1: $\mu \neq \mu_0$, it is two-tailed.

5.10.2 t – test: The ***t*-test** is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the null hypothesis.
A *t*-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's *t* distribution. The *t*-test can be used, for example, to determine if the means of two sets of data are significantly different from each other.
A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the statistical significance. To conduct a test with three or more means, one must use an analysis of variance.
T-Test Assumptions:

1. The first assumption made regarding t-tests concerns the scale of measurement. The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale, such as the scores for an IQ test.
2. The second assumption made is that of a simple random sample, that the data is collected from a representative, randomly selected portion of the total population.
3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.
4. The final assumption is the homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

Correlated (or Paired) T-Test: The correlated t-test is performed when the samples typically consist of matched pairs of similar units, or when there are cases of repeated measures. For example, there may be instances of the same patients being tested repeatedly—before and after receiving a particular treatment. In such cases, each patient is being used as a control sample against themselves.
This method also applies to cases where the samples are related in some manner or have matching characteristics, like a comparative analysis involving children, parents or siblings. Correlated or paired t-tests are of a dependent type, as these involve cases where the two sets of samples are related.

5.10.3 ANOVA: Analysis of variance (**ANOVA**) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the *t*-test beyond two means.

Assumptions: There are four basic assumptions used in ANOVA.

- the expected values of the errors are zero
- the variances of all errors are equal to each other
- the errors are independent
- they are normally distributed

One-way ANOVA: The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. This guide will provide a brief introduction to the one-way ANOVA, including the assumptions of the test and when you should use this test. If you are familiar with the one-way ANOVA, but would like to carry out a one-way ANOVA analysis, go to our guide: One-way ANOVA in SPSS Statistics.

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

where μ = group mean and k = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (H_A), which is that there are at least two group means that are statistically significantly different from each other.

At this point, it is important to realize that the one-way ANOVA is an **omnibus** test statistic and cannot tell you which specific groups were statistically significantly different from each other, only that at least two groups were. To determine which specific groups differed from each other, you need to use a **post hoc test**. Post hoc tests are described later in this guide.

Two-way Analysis of Variance: In statistics, the **two-way analysis of variance** (**ANOVA**) is an extension of the one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable. The two-way ANOVA not only aims at assessing the main effect of each independent variable but also if there is any interaction between them.

5.10.4 Chi-square test: The term "chi-squared test," also written as χ^2 **test**, refers to certain types of statistical hypothesis tests that are valid to perform when the test statistic is chi-squared distributed under the null hypothesis. Often, however, the term is used to refer to *Pearson's* chi-squared test and variants thereof. Pearson's chi-squared test is used to determine whether there is a statistically significant difference (i.e., a magnitude of difference that is unlikely to be due to chance alone) between the expected frequencies and the observed frequencies in one or more categories of a so-called contingency table.

In the standard applications of this test, the observations are classified into mutually exclusive classes. If the so-called null hypothesis is true, the test statistic computed from the observations follows a χ^2 distribution. The purpose of the test is to evaluate how likely the observed frequencies would be assuming the null hypothesis is true.

Test statistics that follow a χ^2 distribution occur when the observations are independent and normally distributed, which assumptions are often justified under the central limit theorem. There are also χ^2 tests for testing the null hypothesis of independence of a pair of random variables based on observations of the pairs.

The term "chi-squared test" is often used to refer to tests for which the distribution of the test statistic approaches the χ^2 distribution *asymptotically*, i.e., the sampling distribution (if the null hypothesis is true) of the test statistic approximates a chi-squared distribution more and more closely as sample sizes increase.

Tests for Different Purposes

1. ***Chi square test for testing goodness of fit*** is used to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value.
   For example, given a sample, we may like to test if it has been drawn from a normal population. This can be tested using chi square goodness of fit procedure.
2. ***Chi square test for independence of two attributes:*** Suppose N observations are considered and classified according two characteristics say A and B. We may be interested to test whether the two characteristics are independent. In such a case, we can use Chi square test for independence of two attributes.
   The example considered above testing for independence of success in the English test vis a vis immigrant status is a case fit for analysis using this test.
3. ***Chi square test for single variance*** is used to test a hypothesis on a specific value of the population variance. Statistically speaking, we test the null hypothesis H0: $\sigma = \sigma 0$ against the research hypothesis H1: $\sigma \# \sigma 0$ where σ is the population mean and $\sigma 0$ is a specific value of the population variance that we would like to test for acceptance.
   In other words, this test enables us to test if the given sample has been drawn from a population with specific variance $\sigma 0$. This is a small sample test to be used only if sample size is less than 30 in general.

Assumptions: The Chi square test for single variance has an assumption that the population from which the sample has been is normal. This normality assumption need not hold for chi square goodness of fit test and test for independence of attributes.

However, while implementing these two tests, one has to ensure that expected frequency in any cell is not less than 5. If it is so, then it has to be pooled with the preceding or succeeding cell so that expected frequency of the pooled cell is at least 5.

Non-Parametric and Distribution Free: It has to be noted that the Chi square goodness of fit test and test for independence of attributes depend only on the set of observed and expected frequencies and degrees of freedom. These two tests do not need any assumption regarding distribution of the parent population from which the samples are taken.

Since these tests do not involve any population parameters or characteristics, they are also termed as non parametric or distribution free tests. An additional important fact on these two tests is they are sample size independent and can be used for any sample size as long as the assumption on minimum expected cell frequency is met.

Uses of Chi Square: The chi-squared distribution has many uses in statistics, including:

- Confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.
- Independence of two criteria of classification of qualitative variables.
- Relationships between categorical variables (contingency tables).
- Sample variance study when the underlying distribution is normal.
- Tests of deviations of differences between expected and observed frequencies (one-way tables).
- The chi-square test (a goodness of fit test).

5.10.5 Mann-Whitney test (U- test): In statistics, the **Mann–Whitney *U* test** (also called the **Mann–Whitney–Wilcoxon** (**MWW**), **Wilcoxon rank-sum test**, or **Wilcoxon–Mann–Whitney test**) is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one population will be less than or greater than a randomly selected value from a second population.

This test can be used to investigate whether two *independent* samples were selected from populations having the same distribution. A similar nonparametric test used on *dependent* samples is the Wilcoxon signed-rank test.

Assumptions of Mann–Whitney–Wilcoxon (MWW), Wilcoxon Rank-Sum Test

Although Mann and Whitney developed the Mann–Whitney *U* test under the assumption of continuous responses with the alternative hypothesis being that one distribution is stochastically greater than the other, there are many other ways to formulate the null and alternative hypotheses such that the Mann–Whitney *U* test will give a valid test.

A very general formulation is to assume that:

1. All the observations from both groups are independent of each other,
2. The responses are ordinal (i.e., one can at least say, of any two observations, which is the greater),
3. Under the null hypothesis H_0, the distributions of both populations are equal.
4. The alternative hypothesis H_1 is that the distributions are not equal.

5.10.6 Kruskal-Wallis test (H-test): The **Kruskal–Wallis test** by ranks, **Kruskal–Wallis *H* test** (named after William Kruskal and W. Allen Wallis), or **one-way ANOVA on ranks** is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes. It extends the Mann–Whitney *U* test, which is used for comparing only two groups. The parametric equivalent of the Kruskal–Wallis test is the one-way analysis of variance (ANOVA).

A significant Kruskal–Wallis test indicates that at least one sample stochastically dominates one other sample. The test does not identify where this stochastic dominance occurs or for how many pairs of groups stochastic dominance obtains. For analyzing the specific sample pairs for stochastic dominance, Dunn's test, pairwise Mann-Whitney tests without Bonferroni correction, or the more powerful but less well known Conover–Iman test are sometimes used.

Since it is a non-parametric method, the Kruskal–Wallis test does not assume a normal distribution of the residuals, unlike the analogous one-way analysis of variance. If the researcher can make the assumptions of an identically shaped and scaled distribution for all groups, except for any difference in medians, then the null hypothesis is that the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group.

Assumptions for the Kruskal Wallis Test

- One independent variable with two or more levels (independent groups). The test is more commonly used when you have three or more levels. For two levels, consider using the Mann Whitney U Test instead.
- Ordinal scale, Ratio Scale or Interval scale dependent variables.
- Your observations should be independent. In other words, there should be no relationship between the members in each group or between groups. For more information on this point, see: Assumption of Independence.
- All groups should have the same shape distributions. Most software (i.e. SPSS, Minitab) will test for this condition as part of the test.

5.10.7 Rank correlation test: A Spearman's Rank correlation test is a non-parametric measure of rank correlation. It is a statistical test used to determine the strength and direction of the association between two ranked variables.

Sub Unit - 5.11: REPORT WRITING

What is Report Writing: It is a formal style of writing elaborately on a topic. The tone of a report is always formal. The audience it is meant for is always thought out section. For example – report writing about a school event, report writing about a business case, etc.

A report can be defined as a testimonial or account of some happening. It is purely based on observation and analysis. A report gives an explanation of any circumstance. In today's corporate world, reports play a crucial role. They are a strong base for planning and control in an organization, i.e., reports give information which can be utilized by the management team in an organization for making plans and for solving complex issues in the organization.

A report discusses a particular problem in detail. It brings significant and reliable information to the limelight of top management in an organization. Hence, on the basis of such information, the management can make strong decisions. Reports are required for judging the performances of various departments in an organization.

The essentials of good/effective report writing are as follows-

1. Know your objective, i.e., be focused.
2. Analyze the niche audience, i.e., make an analysis of the target audience, the purpose for which audience requires the report, kind of data audience is looking for in the report, the implications of report reading, etc.
3. Decide the length of report.
4. Disclose correct and true information in a report.
5. Discuss all sides of the problem reasonably and impartially. Include all relevant facts in a report.
6. Concentrate on the report structure and matter. Pre-decide the report writing style. Use vivid structure of sentences.
7. The report should be neatly presented and should be carefully documented.
8. Highlight and recap the main message in a report.
9. Encourage feedback on the report from the critics. The feedback, if negative, might be useful if properly supported with reasons by the critics. The report can be modified based on such feedback.
10. Use graphs, pie-charts, etc to show the numerical data records over years.
11. Decide on the margins on a report. Ideally, the top and the side margins should be the same (minimum 1 inch broad), but the lower/bottom margins can be one and a half times as broad as others.
12. Attempt to generate reader's interest by making appropriate paragraphs, giving bold headings for each paragraph, using bullets wherever required, etc.

Step in Report Writing: An effective report can be written going through the following steps-

8. Determine the objective of the report, i.e., identify the problem.
9. Collect the required material (facts) for the report.
10. Study and examine the facts gathered.
11. Plan the facts for the report.
12. Prepare an outline for the report, i.e., draft the report.
13. Edit the drafted report.
14. Distribute the draft report to the advisory team and ask for feedback and recommendations.