MRCET CAMPUS

ICET CODE MLRD

RESEARCH METHODOLOGY & STATISTICAL ANALYSIS

Digital Notes

Compiled by

DR. G. NAVEEN KUMAR

DR. A. KAVYA

DR. P. NAGA JYOTHI

MBA

2023-24

Hold on to the ideal. March on! do not look back upon little mistakes and things. In this battle field of ours the dust of mistakes must be raised. Those who are so thin-skinned that they cannot bear the dust, let them get out of the ranks.

- Swami Vivekananda

Life is not just a series of calculations and a sum total of statistics, it's about experience, it's about participation, it is something more complex and more interesting than what is obvious.

- Daniel Libeskind

SUBJECT EXPERTS

Dr. G. NAVEEN KUMAR
B.Sc. (C.S.Engg.), MBA, SLET, PhD
Head of Department,
Department of Business Management,
Malla Reddy College of Engineering & Technology.

Dr. A. KAVYA
B.Tech (CSE) MBA, PhD
Assistant Professor,
Department of Business Management,
Malla Reddy College of Engineering & Technology.

Dr. P.NAGAJYOTHI
MBA, PhD
Assistant Professor,
Department of Business Management,
Malla Reddy College of Engineering & Technology.

Advice to the Students:

1. Use distributive practice rather than massed practice. That is, set aside one to two hours at the same time each day for six days out of the week (Take the seventh day off) for studying statistics. Do not cram your study for four or five hours into one or two sittings each week. This is a cardinal principle.
2. Study in triads or quads of students at least once every week. Verbal interchange and interpretation of concepts and skills with other students really cements a greater depth of understanding.
3. Don't try to memorize formulas (A good instructor will never ask you to do this). Study CONCEPTS CONCEPTS CONCEPTS. Remember, later in life when you need to use a statistical technique you can always look the formula up in a textbook.
4. Work as many and varied problems and exercises as you possibly can. Hopefully your textbook is accompanied by a workbook. You can not learn statistics by just reading about it. You must push the pencil and practice your skills repeatedly.
5. Look for reoccurring themes in statistics. There are probably only a handful of important skills that keep popping up over and over again. Ask your instructor to emphasize these if need be.
6. Must Carry Calculators and Statistical Tables.

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous Institution-UGC, Govt. of India)

Course Title	**: RESEARCH METHODOLOGY & STATISTICAL ANALYSIS**
Course Code	**: R22MBA04 Course**
Year/Semester	**: MBA I Year I Semester**
Course Type	**: Core**
Course Credits	**4**
Course Aim/s:	

- To encourage thinking statistically.
- To develop the abilities to understand and use data.
- To develop expertise in a standard set of statistical and graphical techniques that will be useful in analyzing data.
- To learn to apply these techniques in a number of areas of management.

Learning Outcome/s:

- Appreciate that the collection and statistical analysis of data improves business decisions and reduces the risk of implementing solutions that waste resources and effort.
- Select and deploy the correct statistical method for a given data analysis requirement.
- Achieve a practical level of competence in building statistical models that suit business applications.
- Recognize, develop and distinguish between models for cross-sectional analysis at a single point in time and models for time series analysis at multiple points in time.

Unit-I: Introduction to Research

Introduction to Research: Meaning, Scope, Types of Research, Research Process.
Data collection techniques - Questionnaire Design.

Research Design: Research Problem, Purpose of Research Design, Characteristics of Good Research Design, Sampling and its Applications.

Unit-II: Measures of Central Tendency, Dispersion &

Introduction to Statistics - Measurement of Central Tendency- Mean- Median – Mode;

Measures of Dispersion - Range - Quartile Deviation - Mean Deviation - Standard Deviation and Co-efficient of Variation.
Measures of Skewness.

Unit-III: Tabulation and Graphical Presentation of Data

Classification and Tabulation: Univariate - Bivariate - Multivariate Data - Data Classification and Tabulation.

Graphical Presentation of Data: Diagrammatic and Graphical Representation of Data - One Dimensional - Two Dimensional - Three Dimensional Diagrams and Graphs.

Unit-IV: Correlation and Regression

Correlation Analysis: Introduction: Karl Pearson's Coefficient of Correlation - Spearman's Rank Correlation**.** Scatter Diagram - Positive and Negative Correlation - Limits for Coefficient of Correlation - - Concept of Multiple and Partial Correlation.

Regression Analysis: Concept - Least Square Method - Two Lines of Regression.

Trend analysis - Free Hand Curve - Moving Averages. **Time Series Analysis and Report writing**

Unit-V: Small Sample Tests

Sample Test: t-Distribution - Properties and Applications - Testing for One and Two Means - Paired t-test.

Analysis of Variance: One Way and Two Way ANOVA.

Chi-Square distribution: Test for a specified Population variance - Test for Independence of Attributes.

REFERENCES:

- Levin R.I., Rubin S. David, "Statistics for Management", Pearson.
- Beri, "Business Statistics", TMH.
- Gupta S.C, "Fundamentals of Statistics", HPH.
- Amir D. Aczel and Jayavel Sounder pandian, "Complete Business Statistics", TMH,
- Levine, Stephan , Krehbiel , Berenson - Statistics for Managers using Microsoft Excel, PHI.
- J. K Sharma, "Business Statistics", Pearson.

Note: Refer Class Notes also

UNIT-1

INTRODUCTION TO RESEARCH

OBJECTIVE

To know essence of research in Management

INTRODUCTION TO RESEARCH

- Meaning and Scope
- Types of Research
- Research Process
- Data collection techniques
- Questionnaire Design

RESEARCH DESIGN

- Research Problem
- Purpose of Research Design
- Characteristics of Good Research Design
- Sampling and its Applications

MEANING OF RESEARCH

Research in simple terms refers to **search for knowledge.**
It is a **scientific and systematic search** for information on a **particular topic or issue.**
It is also known as the **art of scientific investigation.**
Several social scientists have defined research in different ways.

- According to Redman and Mory (1923), research is a ―systematized effort to **gain new knowledge‖**. It is an **academic activity (Mini projects, Major projects and report writing)** and therefore the term should be used in a **technical sense (t test, z test, ANOVA, CHI Square test, correlation and regression analysis)**.

- According to Clifford Woody (Kothari, 1988), research comprises **―defining and redefining problems, formulating hypotheses** or **suggested solutions**; **collecting(primary and secondary data),** organizing and evaluating data; making deductions and reaching conclusions; and finally, carefully testing the conclusions to determine whether they fit the formulated hypotheses.

OBJECTIVES OF RESEARCH

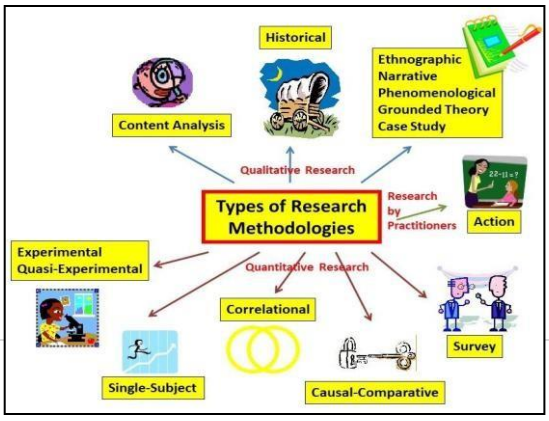
- The objective of research is to **find answers to the questions by applying scientific procedures.**

  In other words, although every research study has its own specific objectives, the research objectives may be broadly grouped as follows:
- To gain familiarity with new insights into a phenomenon (i.e. formulative Research studies);
- To accurately portray the characteristics of a particular individual, group, or a situation(i.e., descriptive research studies);
- To analyze the frequency with which something occurs (i.e., diagnostic research studies)
- To examine the hypothesis of a causal relationship between two variables (i.e., hypothesis- testing research studies).

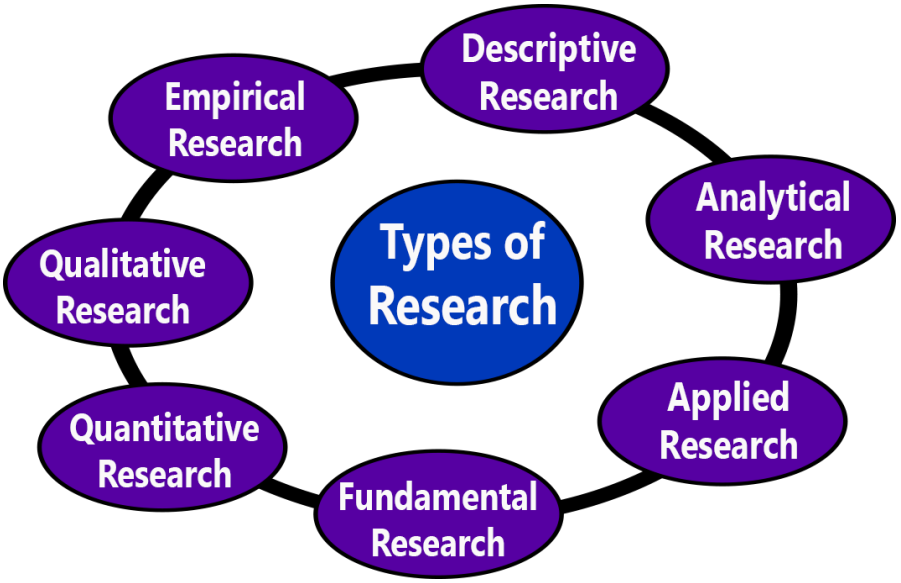
SCOPE OF RESEARCH

Scope of the study refers to the **elements that will be covered in a research project**. It defines the boundaries of the research. The main purpose of the scope of the study is that it explains the extent to which the research area will be explored and thus specifies the parameters that will be observed within the study.

TYPES OF RESEARCH



- Fundamental research.
- Applied research.
- Qualitative research.
- Quantitative research
- Mixed research.
- Exploratory research.
- Field research.
- Cross- sectional Research



- Laboratory research.
- Fixed research

Fundamental Research:

Fundamental, or basic, research is designed to help researchers better understand certain phenomena in the world; **This research attempts to broaden your understanding and expand scientific theories and explanations.** For example, fundamental research could include a company's study of how different product sales. This study provides information and is knowledge-based. Tata motors sales 2020-21 FA APRIL 1ST TO MARCH 31STRELIANCE ALSO

Applied Research:

Applied research is designed to identify solutions to specific problems or find answers to specific questions. For instance, applied research may include a study on ways to increase student involvement in the classroom. This research focuses on a defined problem and is solution-based.

Qualitative Research:

Qualitative research involves **non numerical data,** such as opinions and literature.

Examples of qualitative data may include:

- Focus groups(Team)
- Surveys (consumers, customers, employees and employers)
- Participant comments.
- Observations
- Interviews

Quantitative Research:

Quantitative research **depends on numerical data,** such as statistics and measurements. For example, a car manufacturer may compare the number of sales ofred CARS compared to white CARS. The research uses objective data—the sales figures for red and white CARS—to draw conclusions.

Mixed Research:

Mixed research includes both qualitative and quantitative data. Consider the car manufacturer

comparing AUDI sales. The company could also ask car buyers to complete a survey after buying a red or white sedan that asks how much the color impacted their decision and other opinion-based questions.

Exploratory Research:

Exploratory research is designed to examine what is already known about a topic and what additional information may be relevant. It rarely answers a specific question.

Longitudinal Research:

Longitudinal research focuses on how certain measurements change over time without manipulating any variables. For instance, a researcher may examine if and how employee satisfaction changes in the same employees after one year, three years and five years with the same company.

Cross-sectional Research:

Cross-sectional research studies a group or subgroup at one point in time. Participants are generally chosen based on certain shared characteristics, such as **age, gender or income,** and researchers examine the **similarities and differences** within **groups and between groups**.

Field Research:

Field research takes place wherever the participants or **subjects are, or "on location."** This type of research requires onsite observation and data collection.

RESEARCH PROCESS

Research process consists of a series of steps or actions required for effectively conducting research. The following are the steps that provide useful procedural guidelines regarding the conduct of research:

- Formulating the research problem;
- Extensive literature survey;
- Developing hypothesis;
- Preparing the research design;
- Determining sample design;
- Collecting data;
- Execution of the project;
- Analysis of data;
- Hypothesis testing;
- Generalization and interpretation, and Preparation of the report.
- In other words, it involves the formal write-up of conclusions.

DATA COLLECTION TECHNIQUES - QUESTIONNAIRE DESIGN

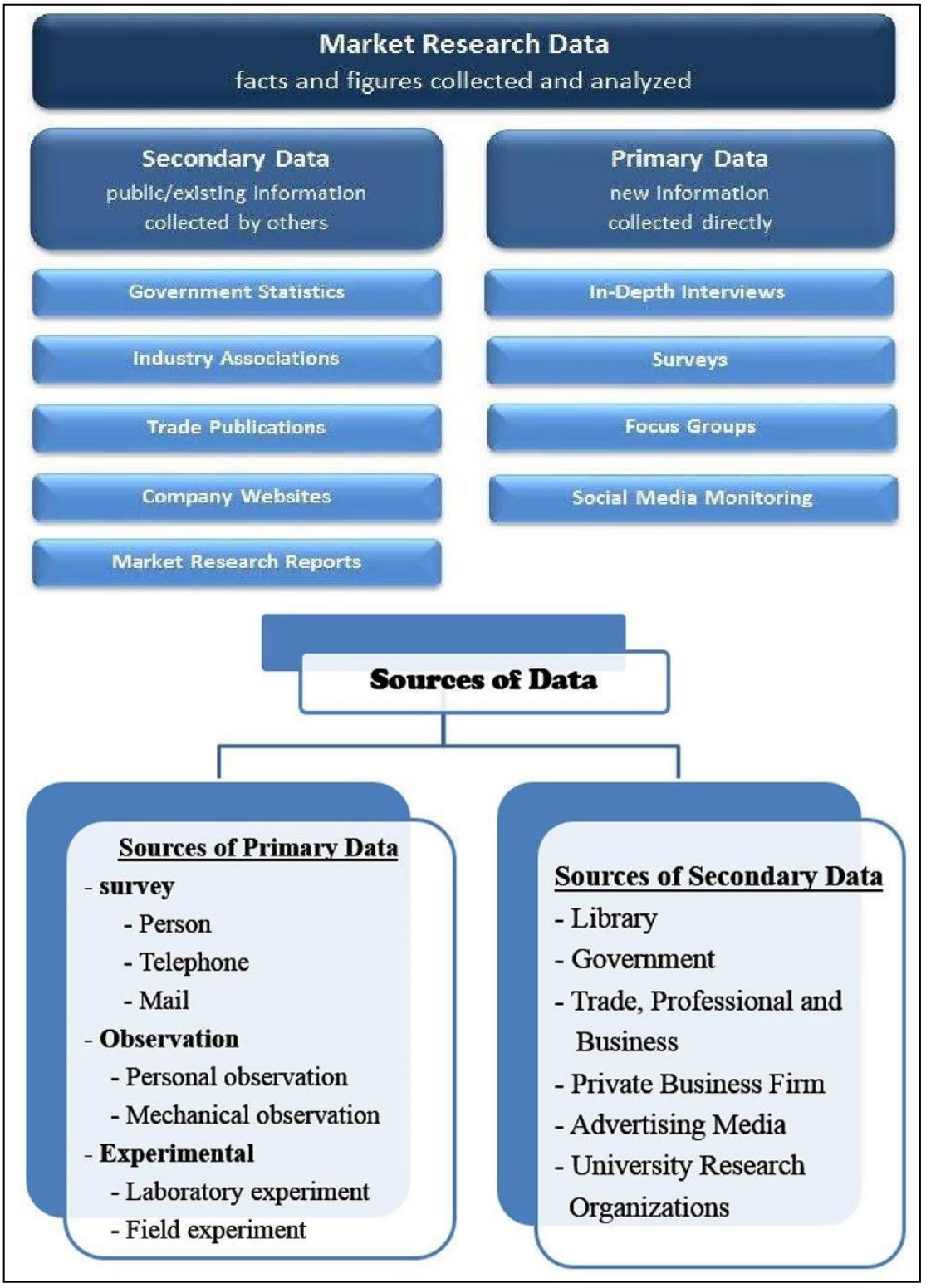
DATA:

Research data is any information that has been collected, observed, generated or created to validate original research findings.

Meaning:

Primary data refers to the first hand data gathered by the researcher himself.

- Source Surveys, observations, experiments, questionnaire, personal interview, etc.
- Secondary data means data collected by someone else earlier.
- Government publications, websites, books, journal articles, internal records etc.



Uses of DATA:

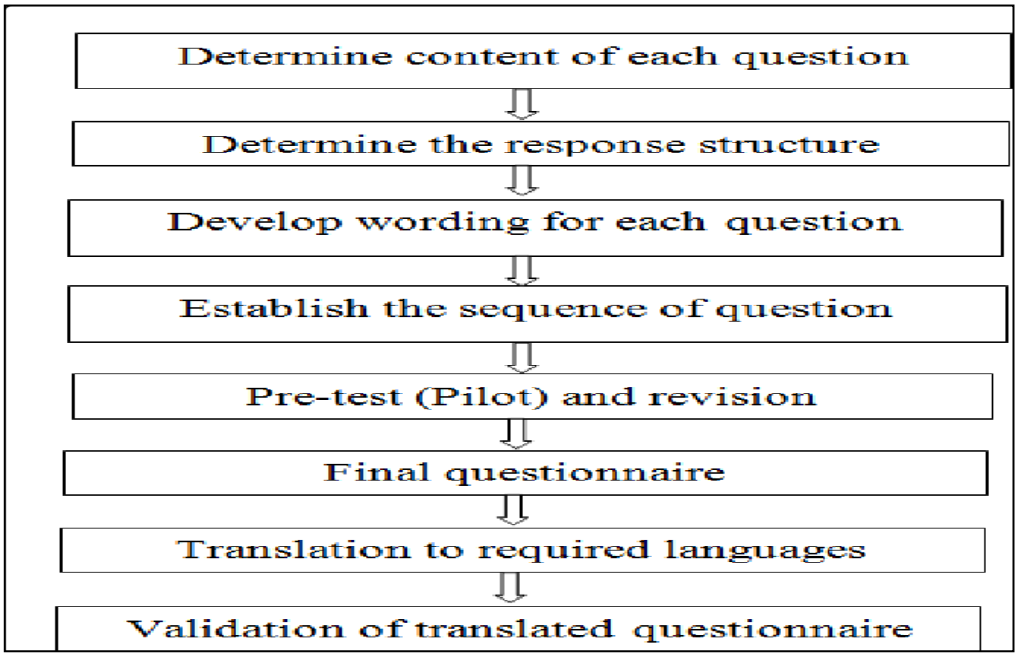
- Data allows organizations to more effectively determine the cause of problems.
- Data allows organizations to visualize relationships between what is happening in different locations, departments, and systems.
- Any information collected, stored, and processed to produce and validate original research results.
- Data might be used to prove or disprove a theory, or to further the knowledge around a specific topic or problem.
- Observe: Take a snapshot of the world.
- Reason: Draw conclusions about how the world works.
- Act: Physically change the world.

Advantages of Primary Data

- Answers a specific research question
- Data are current
- Source of Data is known
- Secrecy can be maintained

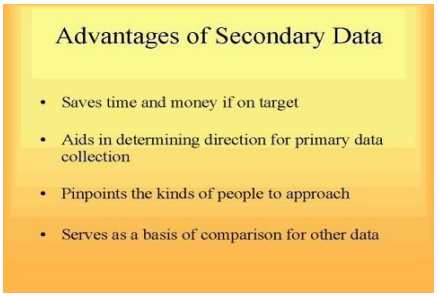
Questionnaire:

A questionnaire is a research tool featuring a series of questions used to collect useful information from respondents. These instruments include either written or oral questions and comprise an interview-style format. ... Questionnaires feature either open or closed questions and sometimes employ a mixture of both.



Schedule of Questions:

A schedule is **a structure of a set of questions on a given topic which are asked by the interviewer or investigator** personally. The order of questions, the language of the questions and the arrangement of parts of the schedule are not changed.



Here are following types of questionnaires:

Computer questionnaire

Respondents are asked to answer the questionnaire which is sentby mail.

Telephone questionnaire

Surveying is a **way to collect information directly from** project stakeholders, participants or beneficiaries in a systematic, standardised way, and rely on the use of questionnaires distributed to respondents.

In-house survey.

The survey will also include a **written description of the property, the street address, the location of buildings and adjacent properties, and any improvements a homeowner can make to the land**. A property survey also includes things like right-of-ways and easements.

Mail Questionnaire

Mail questionnaire is a form of questionnaire which **is mailed to targeted individuals**, which has a collection of questions on a particular topic asked tothem as a part of interview or survey which is used for conducting research on that topic.

Open question questionnaires

Open-ended questions are **questions that allow someone to give a free-form answer**. Closed-ended questions can be answered with "Yes" or "No," or they have a limited set of possible answers (such as: A, B, C, or All of the Above).

Dichotomous Questions.

The dichotomous question is a **question that can have two possible answers**. Dichotomous questions are usually used in a survey that asks for a Yes/No, True/False, Fair/Unfair or Agree/Disagree answers. They are used for a clear distinction of qualities, experiences, or respondent's opinions.

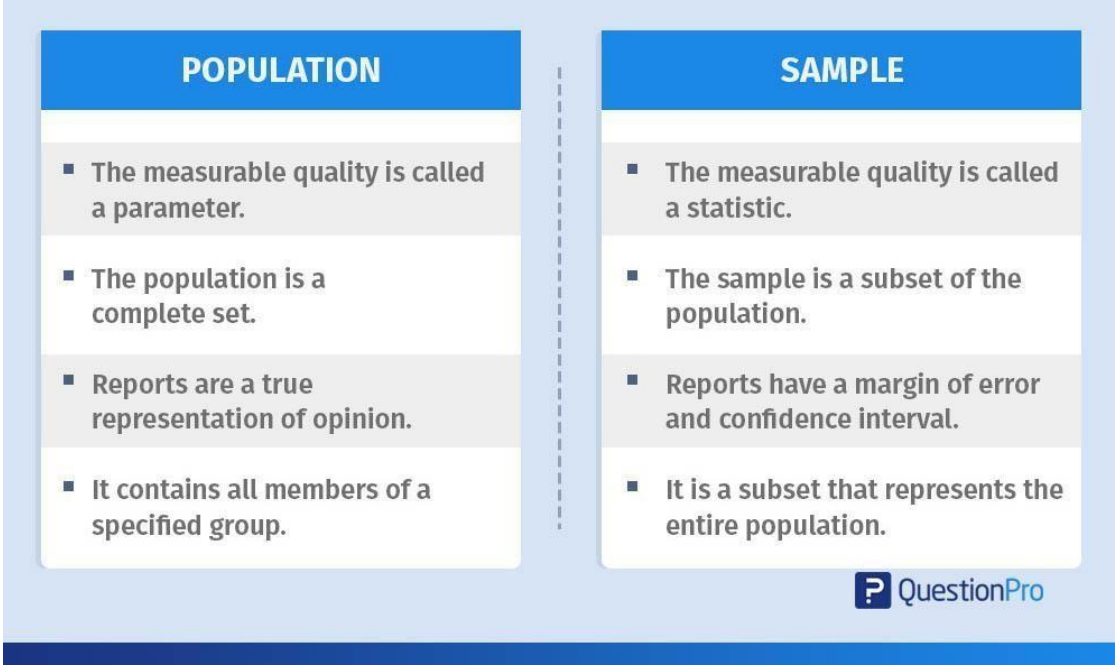
Scaling Questions

Scaling questions **ask clients to consider their position on a scale** (usually from 1 to 10, with one being the least desirable situation and 10 being the most desirable). Scaling questions can be a helpful way to track coaches' progress toward goals and monitor incremental change.

SAMPLING METHODS

Random - Non-Random Techniques - Attitude.

A research population is generally **a large collection of individuals or objects** that is the main focus of a scientific query ..... A research population is also known as a well-defined collection of individuals or objects known to have similar characteristics.



METHODS OF SAMPLING:

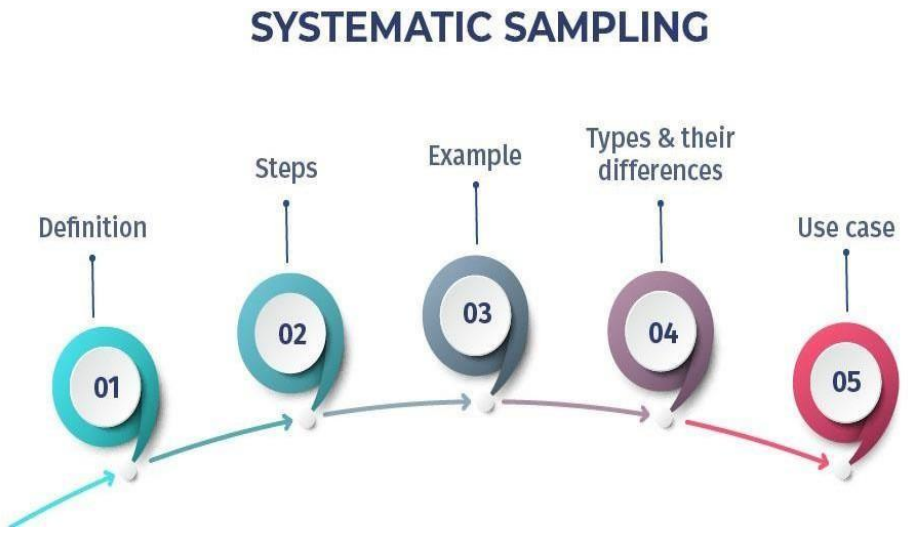
Simple random sampling. Simple random sampling is **a type of probability sampling in which the researcher randomly selects a subset of participants from a population**.

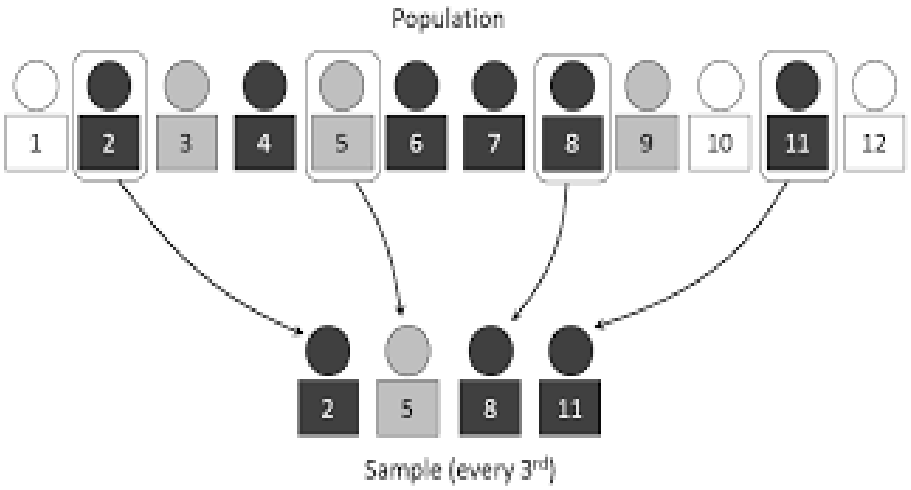
Random Number Table

1	69	24	40	68	29	39	95	60	30
97	23	70	59	79	4	47	19	38	20
13	44	5	71	12	99	78	34	9	96
34	55	83	21	72	3	37	85	61	2
22	80	18	82	54	32	84	16	46	88
7	43	6	48	11	92	63	53	86	28
56	90	36	91	64	45	15	73	10	87
49	65	50	14	51	33	89	52	74	57
98	17	100	58	5	8	77	25	62	31
27	76	66	81	26	93	41	94	67	42

Systematic Sampling

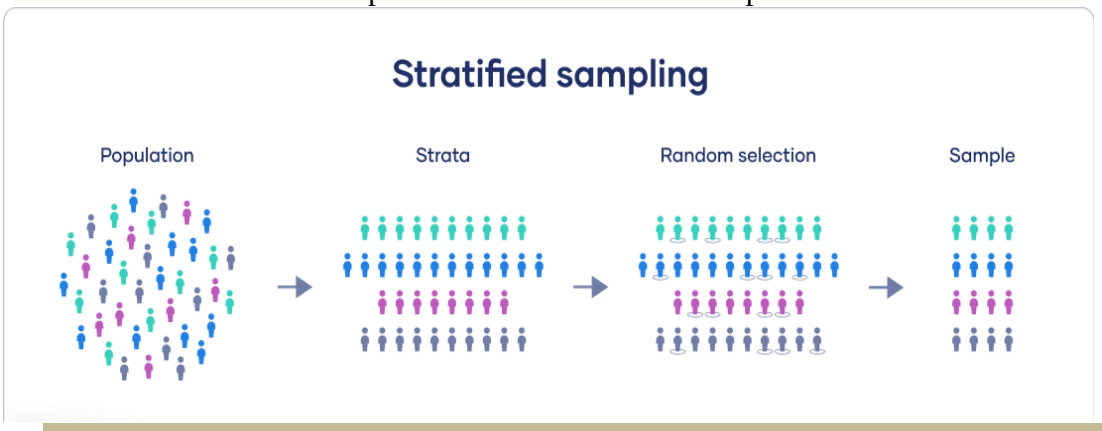
Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size.





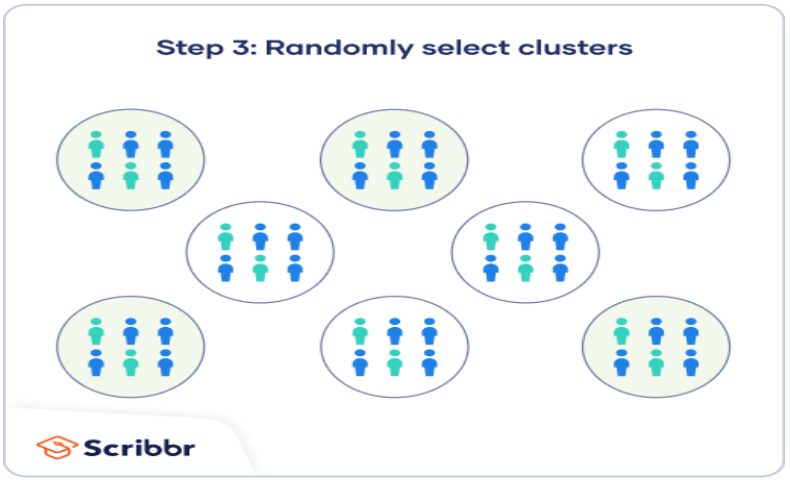
Stratified Sampling

A stratified random sampling **involves dividing the entire population into homogeneous groups called strata** (plural for stratum).....A random sample from each stratum is taken in a number proportional to the stratum's size when compared to the population. These subsets of the strata are then pooled to form a random sample.



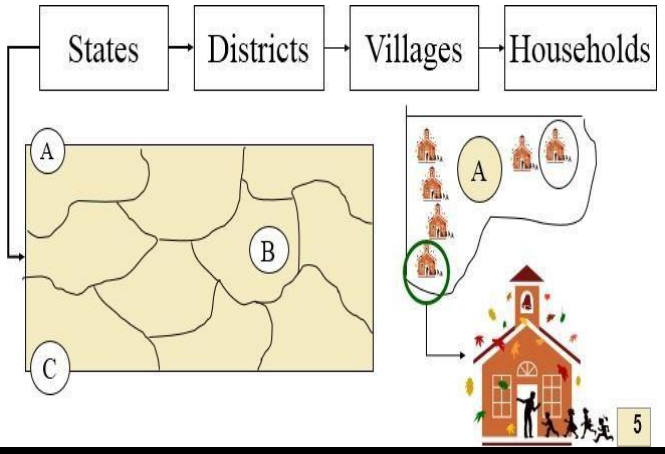
Clustered Sampling

Cluster sampling is a probability sampling technique where researchers divide the population into multiple groups (clusters) for research. Researchers then select random groups with a simple random or systematic random sampling technique for data collection and data analysis.



Multi Stage Sampling

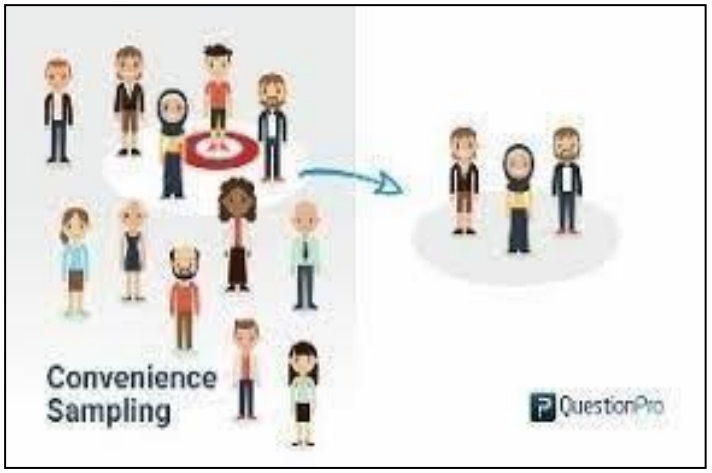
Multistage sampling is defined as **a sampling method that divides the populationinto groups (or clusters) for conducting research**. ... During this sampling method, significant clusters of the selected people are split into sub-groups at various stages to make it simpler for primary data collection.

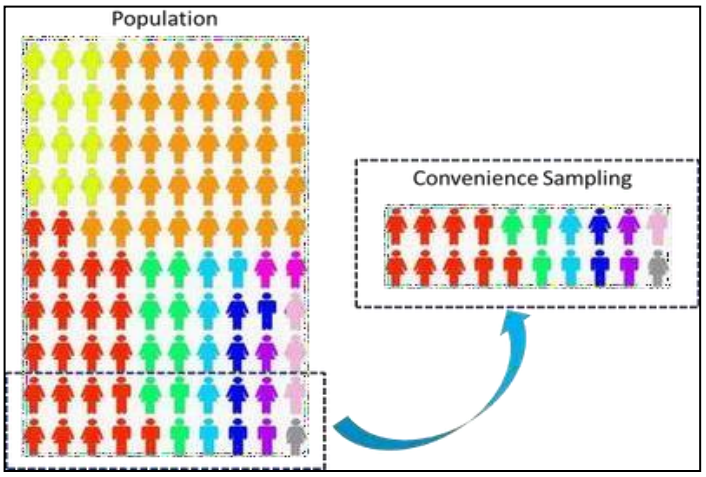


NON PROBABILITY SAMPLING

Convenience sampling:

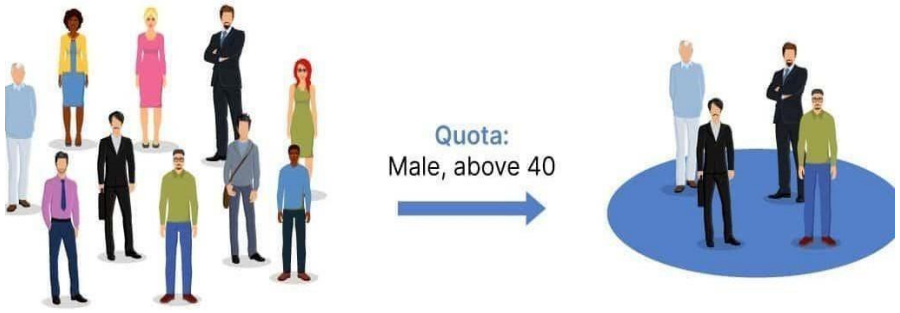
Convenience sampling (also known as grab sampling, accidental sampling, or opportunity sampling) is a type of non-probability sampling that involves the sample being drawn from that part of the population that is close to hand. This type of sampling is most useful for pilot testing.





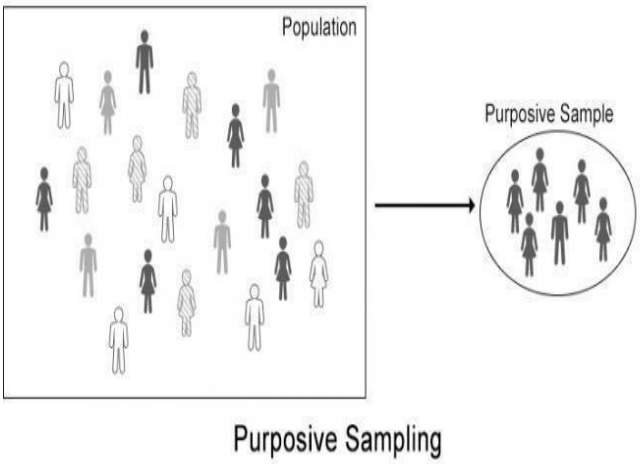
Quota sampling:

Quota sampling is defined as a **non-probability sampling method in which researchers create a sample involving individuals that represent a population**. ... They decide and create quotas so that the market research samples can be useful in collecting data. These samples can be generalized to the entire population.



Judgement (or Purposive) Sampling:

Purposive sampling, also known as judgmental, selective, or subjective sampling, is a **form of non-probability sampling** in which researchers rely on their own judgment when choosing members of the population to participate in their surveys.



UNIT-2

MEASURES OF CENTRAL TENDENCY

OBJECTIVE

To Know the Concepts and their Calculation Methods

CENTRAL TENDENCY MEASURES

- Mean
- Median
- Mode

MEASURES OF DISPERSION

- Range
- Quartile Deviation
- Mean deviation
- Standard Deviation
- Coefficient of Variation

SKEWNESS

- Karl Pearson
- Bowley's
- Kelly's Coefficients

MEASURES OF CENTRAL TENDENCY

One of the important objectives of statistics is to find out various numerical values which explain the inherent characteristics of a frequency distribution. The first of such measures is averages. The averages are the measures which condense a huge unwieldy set of numerical data into single numerical values which represent the entire distribution. The inherent inability of the human mind to remember a large body of numerical data compels us to few constants that will describe the data. Averages provide us the gist and give a bird's eye view of the huge mass of unwieldy numerical data. Averages are the typical values around which other items of the distribution congregate. This value lies between the two extreme observations of the distribution and give us an idea about the concentration of the values in the central part of the distribution. They are called the measures of central tendency.

Averages are also called measures of location since they enable us to locate the position or place of the distribution in question. Averages are statistical constants which enables us to comprehend in a single value the significance of the whole group. According to Croxlon and Cowden, an average value is a single value within the range of the data that is used to represent all the values in that series. Since an average is somewhere within the range of data, it is sometimes called a measure of central value. An average is the most typical representative item of the group to which it belongs and which is capable of revealing all important characteristics of that group or distribution.

Measures of central tendency, Mean, Median, Mode, etc., indicate the central position of a series. They indicate the general magnitude of the data but fail to reveal all the peculiarities and characteristics of the series. In other words, they fail to reveal the degree of the spread out or the extent of the variability in individual items of the distribution. This can be explained by certain other measures, known as „Measures of Dispersion" or Variation.

The study of statistics does not show much interest in things which are constant. The total area of the Earth may not be very important to a research-minded, person but the area covered by different crops, forests, residential and commercial buildings are figures of great importance, because these figures keep on changing from time to time and from place to place. Many experts are engaged in the study of changing phenomena.

Experts working in different countries keep a watch on forces which are responsible for bringing changes in the fields of human interest. Agricultural, industrial and mineral production and their transportation from one area to other parts of the world are of great interest to economists, statisticians, and other experts. Changes in human populations, changes in standards of living, changes in literacy rates and changes in prices attract experts to perform detailed studies and then correlate these changes to human life. Thus, variability or variation is connected with human life and its study is very important for mankind.

Objects of Central Tendency:

The most important object of calculating an average or measuring central tendency is to determine a single figure which may be used to represent a whole series involving magnitudes of the same variable. Second object is that an average represents the empire data; it facilitates comparison within one group or between groups of data. Thus, the performance of the members of a group can be compared with the average performance of different groups. Third object is that an average helps in computing various other statistical measures such as dispersion, skewness, kurtosis etc.

Essential of a Good Average:

An average represents the statistical data and it is used for purposes of comparison, it must possess the following properties.

1. It must be rigidly defined and not left to the mere estimation of the observer. If the definition is rigid, the computed value of the average obtained by different persons shall be similar.
2. The average must be based upon all values given in the distribution. If the item is not based on all value it might not be representative of the entire group of data.
3. It should be easily understood. The average should possess simple and obvious properties. It should be too abstract for the common people.
4. It should be capable of being calculated with reasonable care and rapidity.
5. It should be stable and unaffected by sampling fluctuations.
6. It should be capable of further algebraic manipulation.

Different methods of measuring "Central Tendency" provide us with different kinds of averages. The following are the main types of averages that are commonly used:

1. Mean
2. Median
3. Mode

Arithmetic Mean

Arithmetic mean is the most commonly used average or measure of the central tendency applicable only in case of quantitative data; it is also simply called the "mean". Arithmetic mean is defined as: **"***Arithmetic mean is a quotient of sum of the given values and number of the given values***".**

Arithmetic mean can be computed for both ungrouped data (raw data: data without any statistical treatment) and grouped data (data arranged in tabular form containing different groups).

Pros and Cons of Arithmetic Mean:

Pros:

- It is rigidly defined
- It is easy to calculate and simple to follow
- It is based on all the observations
- It is determined for almost every kind of data
- It is finite and not indefinite
- It is readily used in algebraic treatment
- It is least affected by fluctuations of sampling

Cons:

- The arithmetic mean is highly affected by extreme values
- It cannot average the ratios and percentages properly
- It is not an appropriate average for highly skewed distributions
- It cannot be computed accurately if any item is missing
- The mean sometimes does not coincide with any of the observed values

Median

The median is that value of the variable which divides the group in two equal parts. One part comprising the values greater than and the other all values less than median. Median of a distribution may be defined as that value of the variable which exceeds and is exceeded by the same number of observations. It is the value such that the number of observations above it is equal to the number of observations below it. Thus' we know that the arithmetic mean is based on all items of the

distribution, the median is positional average, i.e., it depends upon the position occupied by a value in the frequency distribution. When the items of a series are arranged in ascending or descending order of magnitude the value of the middle item in the series is known as median in the case of individual observation.

Symbolically, Median = size of n^{th} item

It the number of items is even, and then there is no value exactly in the middle of the series. In such a situation the median is arbitrarily taken to be halfway between the two middle items.

Advantages of Median:

(1) It is very simple to understand and easy to calculate. In some cases it is obtained simply by inspection.
(2) Median lies at the middle part of the series and hence it is not affected by the extreme values.
(3) It is a special average used in qualitative phenomena like intelligence or beauty which are not quantified but ranks are given. Thus, we can locate the person whose intelligence or beauty is the average.
(4) In grouped frequency distribution it can be graphically located by drawing gives.
(5) It is especially useful in open-ended distributions since the position rather than the value of item that matters in median.

Disadvantages of Median:

(1) In simple series, the item values have to be arranged. If the series contains large number of items, then the process becomes tedious.
(2) It is a less representative average because it does not depend on all the items in the series.
(3) It is not capable of further algebraic treatment. For example, we cannot find a combined median of two or more groups if the median of different groups are given.
(4) It is affected more by sampling fluctuations than the mean as it is concerned with on1y one item i.e. the middle item.
(5) It is not rigidly defined. In simple series having even number of items, median cannot be exactly found. Moreover, the interpolation formula applied in the continuous series is based on the unrealistic assumption that the frequency of the median class is evenly spread over the magnitude of the class interval of the median group.

Mode

Mode is that value of the variable which occurs or repeats itself maximum number of item. The mode is most "fashionable" size in the sense that it is the most common and typical and is defined by Zizek as "the value occurring most frequently in series of items and around which the other items are distributed most densely." In the words of Croxton and Cowden, the mode of a distribution is the value at the point where the items tend to be most heavily concentrated. According to A.M. Tuttle, Mode is the value which has the greater frequency density in its immediate neighborhood. In the case of individual observations, the mode is that value which is repeated the maximum number of times in the series. The value of mode can be denoted by the alphabet z also.

Graphic Location of Mode:

Since mode is a positional average, it can be located graphically by the following process:

- A histogram of the frequency distribution is drawn.
- In the histogram, the highest rectangle represents the model class.

- The top left corner of the highest rectangle is joined with the top left corner of the following rectangle and the top right corner of the highest rectangle is joined with the top right corner of the preceding rectangle respectively.
- From the point of intersection of both the lines a perpendicular is drawn on the X-axis, and check that point on the X-axis. This will be the required value of mode.

Advantages and Disadvantages of Mode:

Advantages:

- It is easy to understand and simple to calculate.
- It is not affected by extremely large or small values.
- It can be located just by inspection in ungrouped data and discrete frequency distribution.
- It can be useful for qualitative data.
- It can be computed in an open-end frequency table.
- It can be located graphically.

Disadvantages:

- It is not well defined.
- It is not based on all the values.
- It is stable for large values so it will not be well defined if the data consists of a small number of values.
- It is not capable of further mathematical treatment.
- Sometimes the data has one or more than one mode and sometimes the data has no mode at all.

MEASURES OF DISPERSION

Dispersion:

The word dispersion has a technical meaning in statistics. The average measures the center of the data, and it is one aspect of observation. Another feature of the observation is how the observations are spread about the center. The observations may be close to the center or they may be spread away from the center. If the observations are close to the center (usually the arithmetic mean or median), we say that dispersion, scatter or variation is small. If the observations are spread away from the center, we say dispersion is large.

The study of dispersion is very important in statistical data. If in a certain factory there is consistency in the wages of workers, the workers will be satisfied. But if some workers have high wages and some have low wages, there will be unrest among the low paid workers and they might go on strike and arrange demonstrations. If in a certain country some people are very poor and some are very rich, we say there is economic disparity. This means that dispersion is large.

The idea of dispersion is important in the study of workers' wages, price of commodities, standards of living of different people, distribution of wealth, distribution of land among framers, and many other fields of life. Some brief definitions of dispersion are:

- The degree to which numerical data tend to spread about an average value is called the dispersion or variation of the data.
- Dispersion or variation may be defined as a statistic signifying the extent of the scatteredness of items around a measure of central tendency.
- Dispersion or variation is the measurement of the size of the scatter of items in a series about the average.

Properties of a good measure of Dispersion:

There are certain pre-requisites for a good measure of dispersion:

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be rigidly defined.
4. It should be based on each individual item of the distribution.
5. It should be capable of further algebraic treatment.
6. It should have sampling stability.
7. It should not be unduly affected by the extreme items.

For the study of dispersion, we need some measures which show whether the dispersion is small or large. There are two types of measure of dispersion, which are:

(a) Absolute Measures of Dispersion
(b) Relative Measures of Dispersion

Absolute Measures of Dispersion:

These measures give us an idea about the amount of dispersion in a set of observations. They give the answers in the same units as the units of the original observations. When the observations are in kilograms, the absolute measure is also in kilograms. If we have two sets of observations, we cannot always use the absolute measures to compare their dispersions. We shall explain later as to when the absolute measures can be used for comparison of dispersion in two or more sets of data. The absolute measures which are commonly used are:

1. The Range
2. The Quartile Deviation

3. The Mean Deviation
4. The Standard Deviation and Variance

Relative Measures of Dispersion:

These measures are calculated for the comparison of dispersion in two or more sets of observations. These measures are free of the units in which the original data is measured. If the original data is in dollars or kilometers, we do not use these units with relative measures of dispersion. These measures are a sort of ratio and are called coefficients. Each absolute measure of dispersion can be converted into its relative measure. Thus the relative measures of dispersion are:

1. Coefficient of Range or Coefficient of Dispersion
2. Coefficient of Quartile Deviation or Quartile Coefficient of Dispersion
3. Coefficient of Mean Deviation or Mean Deviation of Dispersion
4. Coefficient of Standard Deviation or Standard Coefficient of Dispersion
5. Coefficient of Variation (a special case of Standard Coefficient of Dispersion)

Range:

Range is a simplest method of studying dispersion. It takes lesser time to compute the 'absolute' and 'relative' range. Range does not take into account all the values of a series, i.e. it considers only the extreme items and middle items are not given any importance. Therefore, Range cannot tell us anything about the character of the distribution. Range cannot be computed in the case of "open ends" distribution i.e., a distribution where the lower limit of the first group and upper limit of the higher group is not given. The concept of range is useful in the field of quality control and to study the variations in the prices of the shares etc.

Quartile Deviation:

The quartile deviation is a slightly better measure of absolute dispersion than the range, but it ignores the observations on the tails. If we take difference samples from a population and calculate their quartile deviations, their values are quite likely to be sufficiently different. This is called sampling fluctuation, and it is not a popular measure of dispersion. The quartile deviation calculated from the sample data does not help us to draw any conclusion (inference) about the quartile deviation in the population.

Advantages of Quartile Deviation:

- It is easy to calculate. We are required simply to find the values of Q1 and Q3 and then apply the formula of absolute and coefficient of quartile deviation.
- It has better results than range method. While calculating range, we consider only the extreme values that make dispersion erratic, in the case of quartile deviation; we take into account middle 50% items.
- The quartile deviation is not affected by the extreme items.

Disadvantages:

- It is completely dependent on the central items. If these values are irregular and abnormal the result is bound to be affected.
- All the items of the frequency distribution are not given equal importance in finding the values of Q1 and Q3.
- Because it does not take into account all the items of the series, considered to be inaccurate.

Average Deviation:

Average deviation is defined as a value which is obtained by taking the average of the deviations of various items from a measure of central tendency Mean or Median or Mode, ignoring negative signs. Generally, the measure of central tendency from which the deviations arc taken, is specified in the problem. If nothing is mentioned regarding the measure of central tendency specified than deviations are taken from median because the sum of the deviations (after ignoring negative signs) is minimum. This method is more effective during the reports presented to the general public or to groups who are not familiar with statistical methods.

Steps to Compute Average Deviation:

1. Calculate the value of Mean or Median or Mode
2. Take deviations from the given measure of central-tendency and they are shown as d.
3. Ignore the negative signs of the deviation that can be shown as \d\ and add them to find S|d|.
4. Apply the formula to get Average Deviation about Mean or Median or Mode.

Advantages of Average Deviations

- Average deviation takes into account all the items of a series and hence, it provides sufficiently representative results.
- It simplifies calculations since all signs of the deviations are taken as positive.
- Average Deviation may be calculated either by taking deviations from Mean or Median or Mode.
- Average Deviation is not affected by extreme items.
- It is easy to calculate and understand.
- Average deviation is used to make healthy comparisons.

Disadvantages of Average Deviations

- It is illogical and mathematically unsound to assume all negative signs as positive signs.
- Because the method is not mathematically sound, the results obtained by this method are not reliable.
- This method is unsuitable for making comparisons either of the series or structure of the series.

Standard Deviation:

The standard deviation, which is shown by greek letters (read as sigma) is extremely useful in judging the representativeness of the mean. The concept of standard deviation, which was introduced by Karl Pearson has a practical significance because it is free from all defects, which exists in a range, quartile deviation or average deviation.

Standard deviation is calculated as the square root of average of squared deviations taken from actual mean. It is also called root mean square deviation. The square of standard deviation i.e., S^2 is called „variance".

Calculation of standard deviation in case of raw data

There are four ways of calculating standard deviation for raw data:

1. When actual values are considered;
2. When deviations are taken from actual mean;
3. When deviations are taken from assumed mean; and
4. When „step deviations" are taken from assumed mean.

Advantages of Standard Deviation:

- Standard deviation is the best measure of dispersion because it takes into account all the items and is capable of future algebraic treatment and statistical analysis.

- It is possible to calculate standard deviation for two or more series.
- This measure is most suitable for making comparisons among two or more series about variability.

Disadvantages:

- It is difficult to compute.
- It assigns more weights to extreme items and less weight to items that are nearer to mean. It is because of this fact that the squares of the deviations which are large in size would be proportionately greater than the squares of those deviations which are comparatively small.

Coefficient of Standard Deviation

The standard deviation is the absolute measure of dispersion. Its relative measure is called the standard coefficient of dispersion or coefficient of standard deviation.

Coefficient of Variation

The most important of all the relative measures of dispersion is the coefficient of variation. This word is variation not variance. There is no such thing as coefficient of variance.

Thus CV is the value of SD when mean is assumed equal to 100. It is a pure number and the unit of observation is not mentioned with its value. It is written in percentage form like 20% or 25%. When its value is 20%, it means that when the mean of the observations is assumed equal to 100, their standard deviation will be 20. The C.V is used to compare the dispersion in different sets of data particularly the data which differ in their means or differ in their units of measurement. The wages of workers may be in dollars and the consumption of meat in families may be in kilograms. The standard deviation of wages in dollars cannot be compared with the standard deviation of amount of meat in kilograms. Both the standard deviations need to be converted into a coefficient of variation for comparison. Suppose the value of C.V for wages is 10% and the values of C.V for kilograms of meat are 25%. This means that the wages of workers are consistent. Their wages are close to the overall average of their wages. But the families consume meat in quite different quantities. Some families consume very small quantities of meat and some others consume large quantities of meat. We say that there is greater variation in their consumption of meat. The observations about the quantity of meat are more dispersed or more variant.

Uses of Coefficient of Variation

- Coefficient of variation is used to know the consistency of the data. By consistency we mean the uniformity in the values of the data/distribution from the arithmetic mean of the data/distribution. A distribution with a smaller C.V than the other is taken as more consistent than the other.
- C.V is also very useful when comparing two or more sets of data that are measured in different units of measurement.

SKEWNESS

Measures of Skewness and Kurtosis, like measures of central tendency and dispersion, study the characteristics of a frequency distribution. Averages tell us about the central value of the distribution and measures of dispersion tell us about the concentration of the items around a central value. These measures do not reveal whether the dispersal of value on either side of an average is symmetrical or not. If observations are arranged in a symmetrical manner around a measure of central tendency, we get a symmetrical distribution; otherwise, it may be arranged in an asymmetrical order which gives asymmetrical distribution. Thus, skewness is a measure that studies the degree and direction of departure from symmetry.

A symmetrical distribution, when presented on the graph paper, gives a 'symmetrical curve', where the value of mean, median and mode are exactly equal. On the other hand, in an asymmetrical distribution, the values of mean, median and mode are not equal. When two or more symmetrical distributions are compared, the difference in them is studied with 'Kurtosis'. On the other hand, when two or more symmetrical distributions are compared, they will give different degrees of Skewness. These measures are mutually exclusive i.e. the presence of skewness implies absence of kurtosis and vice-versa.

Tests of Skewness:

There are certain tests to know whether skewness does or does not exist in a frequency distribution. They are:

1. In a skewed distribution, values of mean, median and mode would not coincide. The values of mean and mode are pulled away and the value of median will be at the Centre. In this distribution, mean-Mode = 2/3 (Median - Mode).
2. Quartiles will not be equidistant from median.
3. When the asymmetrical distribution is drawn on the graph paper, it will not give a bell-shaped curve.
4. Sum of the positive deviations from the median is not equal to sum of negative deviations.
5. Frequencies are not equal at points of equal deviations from the mode.

Nature of Skewness:

Skewness can be positive or negative or zero.

1. When the values of mean, median and mode are equal, there is no skewness.
2. When mean > median > mode, skewness will be positive.
3. When mean < median < mode, skewness will be negative.

Characteristic of a good measure of skewness:

1. It should be a pure number in the sense that its value should be independent of the unit of the series and also degree of variation in the series.
2. It should have zero-value, when the distribution is symmetrical.
3. It should have a meaningful scale of measurement so that we could easily interpret the measured value.

Methods of ascertaining Skewness:

Skewness can be studied graphically and mathematically. When we study skewness graphically, we can find out whether skewness is positive or negative or zero. This can be shown with the help of a diagram:

Mathematically skewness can be studied as:

(a) Absolute Skewness

(b) Relative or coefficient of skewness

When the skewness is presented in absolute term i.e, in units, it is absolute skewness. If the value of skewness is obtained in ratios or percentages, it is called relative or coefficient of skewness. When skewness is measured in absolute terms, we can compare one distribution with the other if the units of measurement are same. When it is presented in ratios or percentages, comparison become easy. Relative measures of skewness is also called coefficient of skewness.

Mathematical measures of skewness can be calculated by:

(a) Bowley's Method
(b) Karl-Pearson's Method
(c) Kelly's method

(a) Bowley's Method:

Bowley's method of skewness is based on the values of median, lower and upper quartiles. This method suffers from the same limitations which are in the case of median and quartiles. Wherever positional measures are given, skewness should be measured by Bowley,s method. This method is also used in case of „open-end series", where the importance of extreme ‘

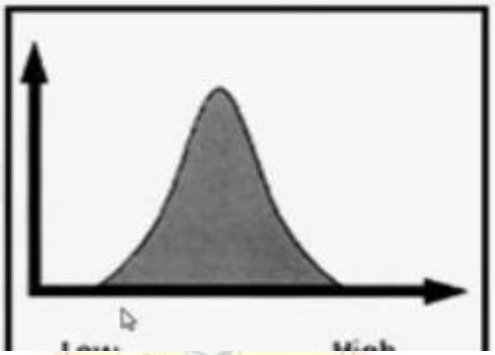
It would be better for a measure to cover the entire data especially because in measuring scenes we are often interested in the more extreme items bowley's measure can be extended by taking any two deciles equidistant from the median or any percentiles equidistant from the median. Kelly has suggested the following formula for measuring skewers upon the 10th and the 90th percentiles (or the first and ninth deciles): This measure of skewers ha one theoretical attraction if skewers is to be based on percentiles however this method is not popular in practice and generally Karl Pearson method is used.

Shape of a Data Distribution

- **Skewness and Kurtosis**

Skewness

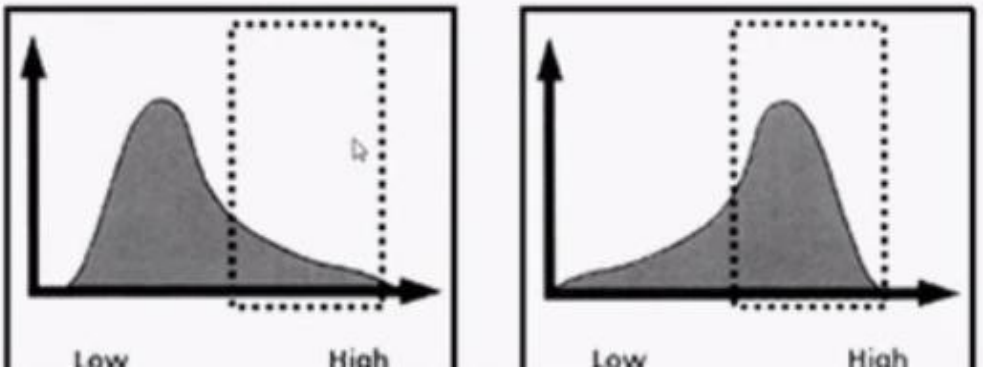
- Skewness is the measure of the symmetry of distribution
- The normal distribution is symmetric and has value of zero for skewness



Shape of the Distribution

Skewness

- A distribution with a long right tail has positive skewness
- A distribution with a long left tail has negative skewness



Shape of the Distribution

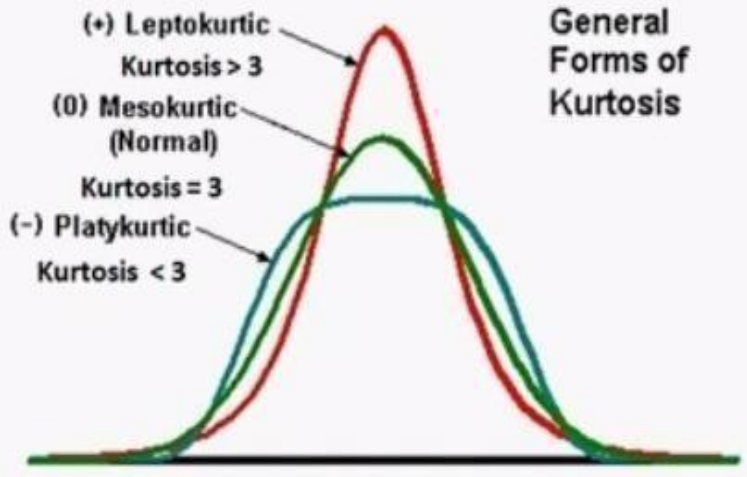
How to Interpret Values of Skewness

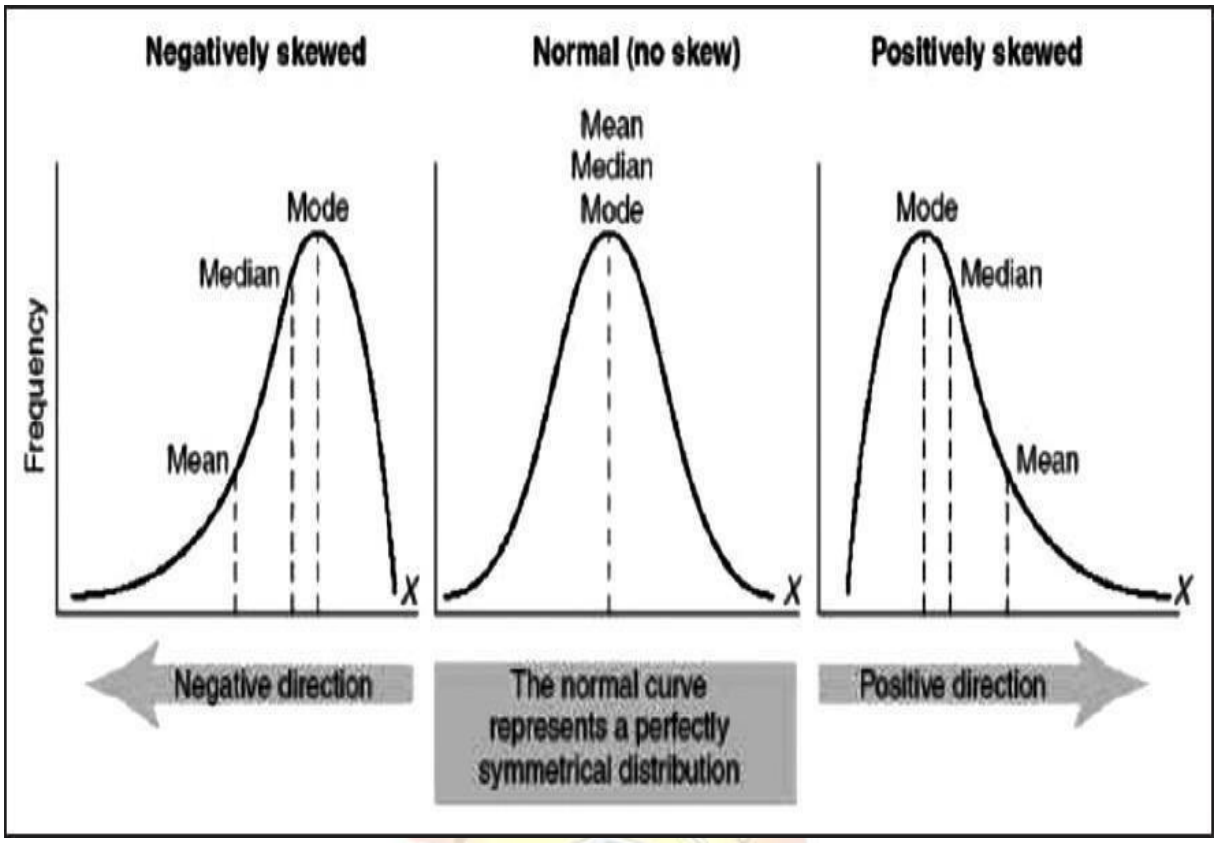
- According to Balmer, M.G. "Principles of Statistics",
  - •If skewness is less than −1 or greater than +1, the distribution is **highly skewed**
  - • If skewness is between −1 and −½ or between +½ and +1, the distribution is **moderately skewed**
  - • If skewness is between −½ and +½, the distribution is **approximately symmetric**

Shape of the Distribution

Kurtosis

- Kurtosis – Refers to peakedness or flatness of the distribution





BASIS FOR COMPARISON	SKEWNESS	KURTOSIS
Meaning	Skewness alludes the tendency of a distribution that determines its symmetry about the mean.	Kurtosis means the measure of the respective sharpness of the curve, in the frequency distribution.
Measure for	Degree of lopsidedness in the distribution.	Degree of tailedness in the distribution.
What is it?	It is an indicator of lack of equivalence in the frequency distribution.	It is the measure of data, which is either peaked or flat in relation to the normal distribution.
Represents	Amount and direction of the skew.	How tall and sharp the central peak is?

LEVELS (or) SCALES OF MEASUREMENT

The level of measurement refers to the relationship among the values that are assigned to the attributes for a variable. Each scale of measurement has certain properties which in turn determine the appropriateness for use of certain statistical analyses. It is important for the researcher to understand the different levels of measurement, as these levels of measurement, together with how the research question is phrased, dictate what statistical analysis is appropriate.

There are typically four levels of measurement that are defined:
1. Nominal
2. Ordinal
3. Interval
4. Ratio

The first level of measurement is **NOMINAL Level of Measurement**. In this level of measurement, the numbers in the variable are used only to classify the data. In this level of measurement, words, letters, and alpha-numeric symbols can be used. Suppose there are data about people belonging to three different gender categories. In this case, the person belonging to the female gender could be classified as F, the person belonging to the male gender could be classified as M, and transgendered classified as T. This type of assigning classification is nominal level of measurement.

The second level of measurement is the **ORDINAL Level of Measurement**. This level of measurement depicts some ordered relationship among the variable's observations. Suppose a student scores the highest grade of 100 in the class. In this case, he would be assigned the first rank. Then, another classmate scores the second highest grade of an 92; she would be assigned the second rank. A third student scores a 81 and he would be assigned the third rank, and so on. The ordinal level of measurement indicates an ordering of the measurements.

The third level of measurement is the **INTERVAL Level of Measurement**. The interval level of measurement not only classifies and orders the measurements, but it also specifies that the distances between each interval on the scale are equivalent along the scale from low interval to high interval. For example, an interval level of measurement could be the measurement of anxiety in a student between the score of 10 and 11; this interval is the same as that of a student who scores between 40 and 41. A popular example of this level of measurement is temperature in centigrade, where, for example, the distance between 94^0C and 96^0C is the same as the distance between 100^0C and 102^0C.

The fourth level of measurement is the **RATIO Level of Measurement**. In this level of measurement, the observations, in addition to having equal intervals, can have a value of zero as well. The zero in the scale makes this type of measurement unlike the other types of measurement, although the properties are similar to that of the interval level of measurement. In the ratio level of measurement, the divisions between the points on the scale have an equivalent distance between them.

Stevens (1946, 1951) proposed that measurements can be classified into four different types of scales

Scale Type	Permissible Statistics	Admissible Scale Transformation	Mathematical Structure
Nominal	Mode, Chi-Square	One to One (Equality(=))	Standard Set Structure (Unordered)
Ordinal	Median, Percentile	Monotonic Increasing (Order(<))	Totally Ordered Set
Interval	Mean, SD, Correlation, Regression, ANOVA	Positive Linear (Affine)	Affine Line
Ratio	All Statistics permitted for Interval Scales plus the following: GM, HM, Coefficient of Variation, Logarithms	Positive Similarities (Multiplication)	Field

UNIT-3

CLASSIFICATION AND TABULATION

OBJECTIVE

To Know the Tabulation Methods for representing Data in Précised manner

CLASSIFICATION AND TABULATION

- Univariate
- Bivariate
- Multivariate
- Classification and Tabulation

GRAPHICAL PRESENTATION OF DATA

- One Dimensional
- Two Dimensional
- Three Dimensional Diagrams and Graphs

CLASSIFICATION OF DATA

Classification:

The collected data, also known as raw data or ungrouped data are always in an un organised form and need to be organised and presented in meaningful and readily comprehensible form in order to facilitate further statistical analysis. It is, therefore, essential for an investigator to condense a mass of data into more and more comprehensible and assimilable form. The process of grouping into different classes or sub classes according to some characteristics is known as classification, tabulation is concerned with the systematic arrangement and presentation of classified data. Thus classification is the first step in tabulation. For Example, letters in the post office are classified according to their destinations viz., Delhi, Madurai, Bangalore, Mumbai etc.,

Objects of Classification:

The following are main objectives of classifying the data.

1. It condenses the mass of data in an easily assimilable form.
2. It eliminates unnecessary details.
3. It facilitates comparison and highlights the significant aspect of data.
4. It enables one to get a mental picture of the information and helps in drawing inferences.
5. It helps in the statistical treatment of the information collected.

Types of classification:

Statistical data are classified in respect of their characteristics. Broadly there are four basic types of classification namely

a) Chronological Classification;
b) Geographical Classification;
c) Qualitative Classification;
d) Quantitative Classification.

a) Chronological Classification: In chronological classification the collected data are arranged according to the order of time expressed in years, months, weeks, etc., The data is generally classified in ascending order of time. For example, the data related with population, sales of a firm, imports and exports of a country are always subjected to chronological classification.

Year	2005	2006	2007	2008	2009	2010	2011
Birth rate	36.8	36.9	36.6	34.6	34.5	35.2	34.2

b) Geographical Classification: In this type of classification the data are classified according to geographical region or place. For instance, the production of paddy in different states in India, production of wheat in different countries etc.,

Country	America	China	Denmark	France	India
Yield of Wheat in (kg/acre)	1925	893	225	439	862

c) Qualitative Classification: In this type of classification data are classified on the basis of same attributes or quality like sex, literacy, religion, employment etc., such attributes cannot be measured along with a scale. For example, if the population to be classified in respect to one attribute, say sex, then we can classify them into two namely that of males and females. Similarly, they can also be classified into ‘employed’ or ‘unemployed’ on the basis of another attribute ‘employment’. Thus

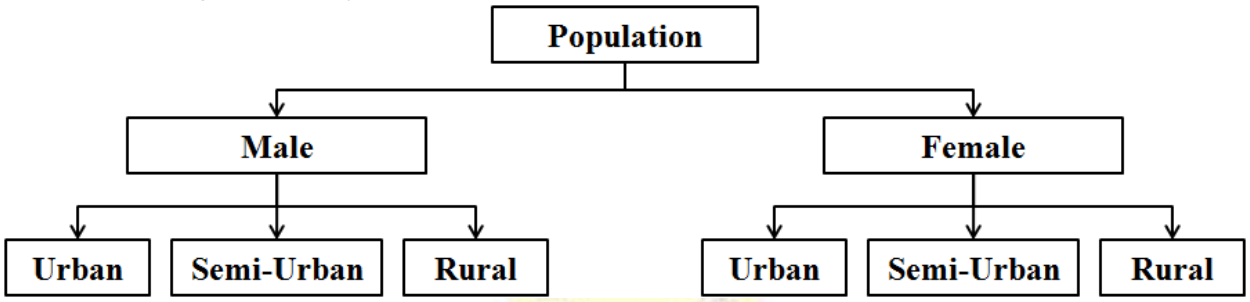
when the classification is done with respect to one attribute, which is dichotomous in nature, two classes are formed, one possessing the attribute and the other not possessing the attribute. This type of classification is called simple or dichotomous classification.

The classification, where two or more attributes are considered and several classes are formed, is called a manifold classification. For example, if we classify population simultaneously with respect to two attributes, e.g. sex and employment, then population are first classified with respect to 'sex' into 'males' and 'females'. Each of these classes may then be further classified into 'Urban', 'Semi-Urban' and 'Rural' on the basis of attribute 'employment' and as such Population are classified into four classes namely.

(i) Male in Urban Area
(ii) Male Semi-Urban Area
(iii) Male in Rural Area
(iv) Female in Urban Area
(v) Female Semi-Urban Area
(vi) Female in Rural Area

Still the classification may be further extended by considering other attributes like marital status etc.

This can be explained by the following chart



S. No.	Management	Number of Schools
1	Government	4
2	Local Body	8
3	Private Aided	10
4	Private Unaided	2
Total		24

d) Quantitative Classification: Quantitative classification refers to the classification of data according to some characteristics that can be measured such as height, weight, etc., For example the students of a college may be classified according to weight. In this type of classification there are two elements, namely

(i) The variable (i.e.) the weight in the above example, and
(ii) The frequency in the number of students in each class.

There are 50 students having weights ranging from 90 to 100 lb, 200 students having weight ranging between 100 to 110 lb and so on.

Weight (in lbs)	90-100	100-110	110-120	120-130	130-140	140-150	**Total**
No. of Students	50	200	260	360	90	40	**1000**

TABULATION OF DATA

Tabulation is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information. A table is a systematic arrangement of classified data in columns and rows. Thus, a statistical table makes it possible for the investigator to present a huge mass of data in a detailed and orderly form. It facilitates comparison and often reveals certain patterns in data which are otherwise not obvious. Classification and „Tabulation", as a matter of fact, are not two distinct processes. Actually, they go together. Before tabulation data are classified and then displayed under different columns and rows of a table.

Advantages of Tabulation:

Statistical data arranged in a tabular form serve following objectives:

1. It simplifies complex data and the data presented are easily understood.
2. It facilitates comparison of related facts.
3. It facilitates computation of various statistical measures like averages, dispersion, correlation etc.
4. It presents facts in minimum possible space and unnecessary repetitions and explanations are avoided. Moreover, the needed information can be easily located.
5. Tabulated data are good for references and they make it easier to present the information in the form of graphs and diagrams.

Preparing a Table:

The making of a compact table itself an art. This should contain all the information needed within the smallest possible space. What the purpose of tabulation is and how the tabulated information is to be used are the main points to be kept in mind while preparing for a statistical table. An ideal table should consist of the following main parts:

1. Table number
2. Title of the table
3. Captions or column headings
4. Stubs or row designation
5. Body of the table
6. Footnotes
7. Sources of data

Table Number: A table should be numbered for easy reference and identification. This number, if possible, should be written in the center at the top of the table. Sometimes it is also written just before the title of the table.

Title: A good table should have a clearly worded, brief but unambiguous title explaining the nature of data contained in the table. It should also state arrangement of data and the period covered. The title should be placed centrally on the top of a table just below the table number (or just after table number in the same line).

Captions or column Headings: It stands for brief and self-explanatory headings of vertical columns. Captions may involve headings and sub-headings as well. The unit of data contained should also be given for each column. Usually, a relatively less important and shorter classification should be tabulated in the columns.

Stubs or Row Designations: Stubs stands for brief and self-explanatory headings of horizontal rows. Normally, a relatively more important classification is given in rows. Also, a variable with a large number of classes is usually represented in rows. For example, rows may stand for score of classes

and columns for data related to sex of students. In the process, there will be many rows for scores classes but only two columns for male and female students.

Body: It contains the numerical information of frequency of observations in the different cells. This arrangement of data is according to the description of captions and stubs.

Footnotes: They are given at the foot of the table for explanation of any fact or information included in the table which needs some explanation. Thus, they are meant for explaining or providing further details about the data, which have not been covered in title, captions and stubs.

Sources of data: Lastly one should also mention the source of information from which data are taken. This may preferably include the name of the author, volume, page and the year of publication. This should also state whether the data contained in the table is of „primary or secondary" nature.

Requirements of a Good Table:

A good statistical table is not merely a careless grouping of columns and rows but should be such that it summarizes the total information in an easily accessible form in minimum possible space. Thus while preparing a table, one must have a clear idea of the information to be presented, the facts to be compared and he points to be stressed.

Though, there is no hard and fast rule for forming a table yet a few general points should be kept in mind:

1. A table should be formed in keeping with the objects of statistical enquiry.
2. A table should be carefully prepared so that it is easily understandable.
3. A table should be formed so as to suit the size of the paper. But such an adjustment should not be at the cost of legibility.
4. If the figures in the table are large, they should be suitably rounded or approximated. The method of approximation and units of measurements too should be specified.
5. Rows and columns in a table should be numbered and certain figures to be stressed may be put in 'box' or 'circle' or in bold letters.
6. The arrangements of rows and columns should be in a logical and systematic order. This arrangement may be alphabetical, chronological or according to size.
7. The rows and columns are separated by single, double or thick lines to represent various classes and sub-classes used. The corresponding proportions or percentages should be given in adjoining rows and columns to enable comparison. A vertical expansion of the table is generally more convenient than the horizontal one.
8. The averages or totals of different rows should be given at the right of the table and that of columns at the bottom of the table. Totals for every sub-class too should be mentioned.
9. In case it is not possible to accommodate all the information in a single table, it is better to have two or more related tables.

Type of Tables:

Tables can be classified according to their purpose, stage of enquiry, nature of data or number of characteristics used. On the basis of the number of characteristics, tables may be classified as follows:

1. Simple or one-way table
2. Two way table
3. Manifold table

Simple or one-way Table:

A simple or one-way table is the simplest table which contains data of one characteristic only. A simple table is easy to construct and simple to follow. For example, the blank table given below may be used to show the number of adults in different occupations in a locality.

Occupations	No. of Adults
Total	

Two-way Table:

A table, which contains data on two characteristics, is called a two-way table. In such case, therefore, either stub or caption is divided into two co-ordinate parts. In the given table, as an example the caption may be further divided in respect of „sex". This subdivision is shown in two-way table, which now contains two characteristics namely, occupation and sex.

Occupations	No. of Adults		Total
	Male	Female	
Total			

Manifold Table:

Thus, more and more complex tables can be formed by including other characteristics. For example, we may further classify the caption sub-headings in the above table in respect of "marital status", "religion" and "socio-economic status" etc. A table, which has more than two characteristics of data, is considered as a manifold table. For instance, table shown below shows three characteristics namely, occupation, sex and marital status. Manifold tables, though complex are good in practice as these enable full information to be incorporated and facilitate analysis of all related facts. Still, as a normal practice, not more than four characteristics should be represented in one table to avoid confusion. Other related tables may be formed to show the remaining characteristics

Occupations	No. of Adults						Total
	Male			Female			
	Married	Unmarried	Total	Married	Unmarried	Total	
Total							

UNIT-4

CORRELATION AND REGRESSION

OBJECTIVE
To Know the Correlation and Regression Analysis of Data

CORRELATION ANALYSIS

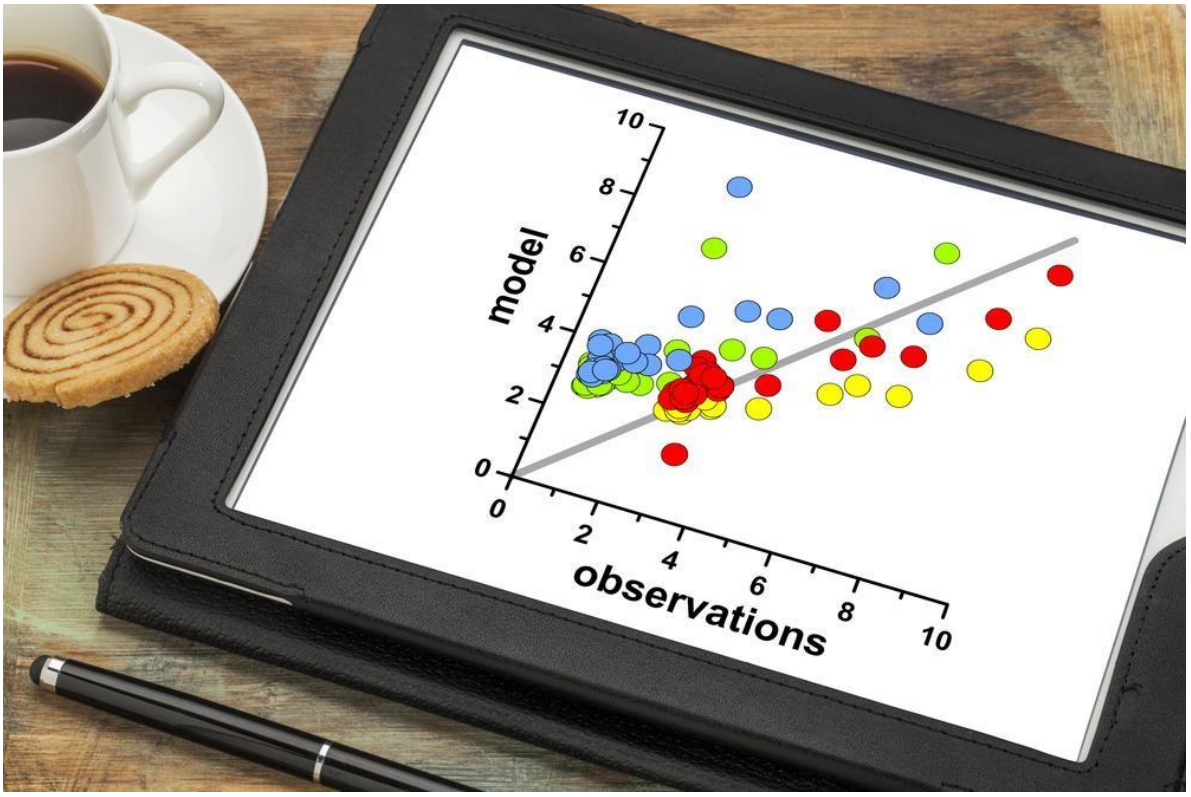
- Types of Correlation
- Limits of Correlation
- Karl Pearson's Coefficient of Correlation
- Spearman's Coefficient of Correlation

REGRESSION ANALYSIS

- Least Square Method
- Two lines of Regression
- Properties of regression Coefficients

TIME SERIES ANALYSIS

- Trend Analysis
- Free Hand curve
- Moving Averages



CORRELATION

Introduction:

The term correlation is used by a common man without knowing that he is making use of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

The study related to the characteristics of only variable such as height, weight, ages, marks, wages, etc., is known as Univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bivariate Analysis. Sometimes the variables may be inter-related. In health sciences we study the relationship between blood pressure and age, consumption level of some nutrient and weight gain, total income and medical expenditure, etc., the nature and strength of relationship may be examined by correlation and Regression analysis. Thus, Correlation refers to the relationship of two variables or more. (eg.) relation between height of father and son, yield and rainfall, wage and price index, share and debentures etc.

Correlation is statistical Analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. Price and supply, income and expenditure are correlated.

Definitions:

1. Croxton and Cowden, "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation".
2. A.M. Tuttle, "Correlation is an analysis of the co-variation between two or more variables."
3. W.A. Neiswanger, "Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective."
4. L. R. Conner, "If two or more quantities vary in sympathy so that the movement in one tends to be accompanied by corresponding movements in others than they are said to be correlated."

Uses of correlation:

1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity etc. Businessmen estimates costs, sales, price etc. using correlation.
3. It is helpful in measuring the degree of relationship between the variables like income and expenditure, price and supply, supply and demand etc.
4. Sampling error can be calculated.
5. It is the basis for the concept of regression.

Types of Correlation

Correlation can be categorised as one of the following:

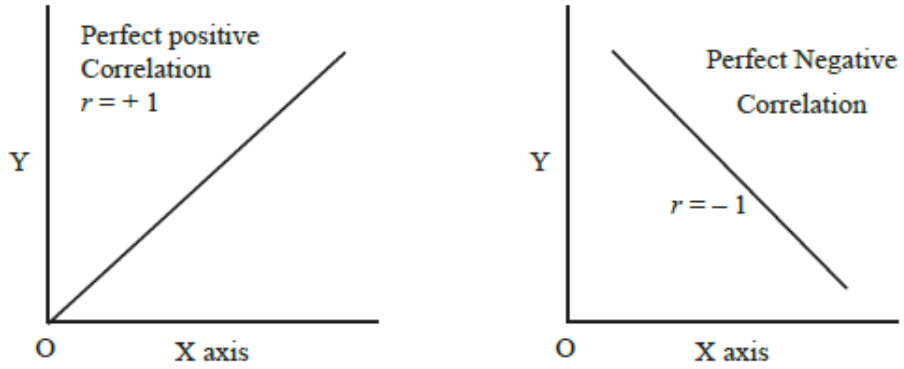
1. Positive and Negative.
2. Simple and Multiple.
3. Partial and Total.
4. Linear and Non-Linear (Curvilinear).

1. **Positive and Negative Correlation:** Positive or direct Correlation refers to the movement of variables in the same direction. The correlation is said to be positive when the increase (decrease) in the value of one variable is accompanied by an increase (decrease) in the value of other variable also. Negative or inverse correlation refers to the movement of the variables in opposite direction. Correlation is said to be negative, if an increase (decrease) in the value of one variable is accompanied by a decrease (increase) in the value of other.
2. **Simple and Multiple Correlation:** Under simple correlation, we study the relationship between two variables only i.e., between the yield of wheat and the amount of rainfall or between demand and supply of a commodity. In case of multiple correlations, the relationship is studied among three or more variables. For example, the relationship of yield of wheat may be studied with both chemical fertilizers and the pesticides.
3. **Partial and Total Correlation:** There are two categories of multiple correlation analysis. Under partial correlation, the relationship of two or more variables is studied in such a way that only one dependent variable and one independent variable is considered and all others are kept constant. For example, coefficient of correlation between yield of wheat and chemical fertilizers excluding the effects of pesticides and manures is called partial correlation. Total correlation is based upon all the variables.
4. **Linear and Non-Linear Correlation:** When the amount of change in one variable tends to keep a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. But if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable then the correlation is said to be non-linear. The distinction between linear and non-linear is based upon the consistency of the ratio of change between the variables.

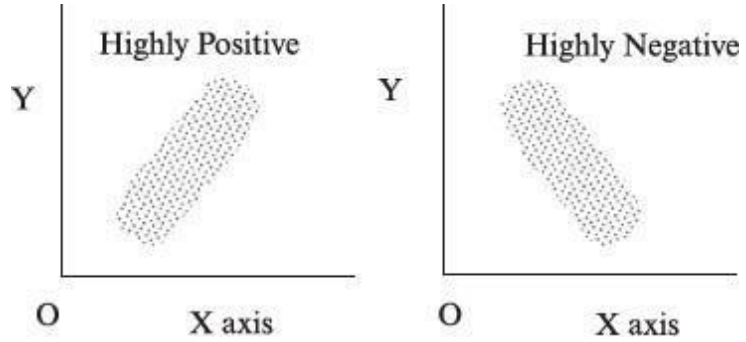
Scatter Diagram:

It is the simplest method of studying the relationship between two variables diagrammatically. One variable is represented along the horizontal axis and the second variable along the vertical axis. For each pair of observations of two variables, we put a dot in the plane. There are as many dots in the plane as the number of paired observations of two variables. The direction of dots shows the scatter or concentration of various points. This will show the type of correlation.

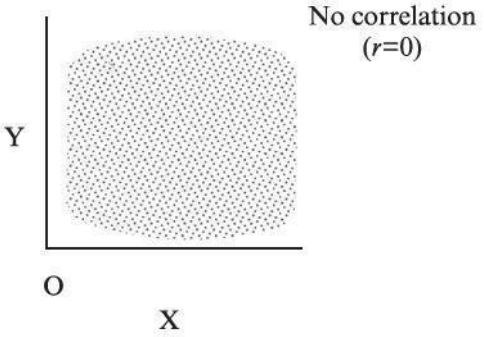
1. If all the plotted points form a straight line from lower left hand corner to the upper right hand corner then there is Perfect positive correlation. We denote this as $r = +$.



2. If all the plotted dots lie on a straight line falling from upper left hand corner to lower right hand corner, there is a perfect negative correlation between the two variables. In this case the coefficient of correlation takes the value $r = -1$.
3. If the plotted points in the plane form a band and they show a rising trend from the lower left hand corner to the upper right hand corner the two variables are highly positively correlated.



4. If the points fall in a narrow band from the upper left hand corner to the lower right hand corner, there will be a high degree of negative correlation.
5. If the plotted points in the plane are spread all over the diagram there is no correlation between the two variables.



Merits:

1. It is a simplest and attractive method of finding the nature of correlation between the two variables.
2. It is a non-mathematical method of studying correlation. It is easy to understand.
3. It is not affected by extreme items.
4. It is the first step in finding out the relation between the two variables.
5. We can have a rough idea at a glance whether it is a positive correlation or negative correlation.

Demerits:

1. By this method we cannot get the exact degree or correlation between the two variables.

Karl Pearson‟s Co-efficient of Correlation

Popularly known as Pearsonian co-efficient of correlation, is most widely applied in practice to measure correlation. The Pearsonian co-efficient of correlation is represented by the symbol r. According to Karl Pearson‟s method, co-efficient of correlation between the variables is obtained by dividing the sum of the products of the corresponding deviations of the various items of two series from their respective means by the product of their standard deviations and the number of pairs of observations. Symbolically, r = where r stands for coefficient of correlation ...(i) where x1, x2, x3, x4..... xn are the deviations of various items of the first variable from the mean, y1, y2, y3,. .....yn are the deviations of all items of the second variable from mean, Sxy is the sum of products of these corresponding deviations. N stands for the number of pairs, sx stands for the standard deviation of X variable and sy stands for the standard deviation of Y variable. sx = and sy = If we substitute the value of sx and sy in the above written formula of computing r, we get r = or r = Degree of correlation varies between + 1 and –1; the result will be + 1 in case of perfect positive correlation and – 1 in case of perfect negative correlation. Computation of correlation

coefficient can be simplified by dividing the given data by a common factor. In such a case, the final result is not multiplied by the common factor because coefficient of correlation is independent of change of scale and origin.

Assumptions:

Karl Pearson based his formula on following basic assumptions:

1. Two variables are affected by many independent causes and form a normal distribution.
2. The cause and effect relationship exists between two variables.
3. The relationship between two variables is linear. It is often denoted by r.

Merits and Demerits of Pearson's method of studying correlation:

Merits

1. This method indicates the presence or absence of correlation between two variables and gives the exact degree of their correlation.
2. In this method, we can also ascertain the direction of the correlation; positive, or negative.
3. This method has many algebraic properties for which the calculation of co-efficient of correlation, and other related factors, are made easy.

Demerits:

1. It is more difficult to calculate than other methods of calculations.
2. It is much affected by the values of the extreme items.
3. It is based on a many assumptions, such as: linear relationship, cause and effect relationship etc. which may not always hold good.
4. It is very much likely to be misinterpreted in case of homogeneous data.

Spearman‟s Rank Correlation:

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc. The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by Edward Spearman in 1904. The value of r lies between –1 and +1. If r = +1, there is complete agreement in order of ranks and the direction of ranks is also same. If r = -1, then there is complete disagreement in order of ranks and they are in opposite directions. Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an average rank is to be given to each individual item. The sign of the Spearman correlation indicates the direction of association between X (the independent variable) and Y (the dependent variable). If Y tends to increase when X increases, the Spearman correlation coefficient is positive. If Y tends to decrease when X increases, the Spearman correlation coefficient is negative. A Spearman correlation of zero indicates that there is no tendency for Y to either increase or decrease when X increases. The Spearman correlation increases in magnitude as X and Y become closer to being perfect monotone functions of each other. When X and Y are perfectly monotonically related, the Spearman correlation coefficient becomes 1. A perfect monotone increasing relationship implies that for any two pairs of data values X_i, Y_i and X_j, Y_j, that $X_i - X_j$ and $Y_i - Y_j$ always have the same sign. A perfect monotone decreasing relationship implies that these differences always have opposite signs. The Spearman correlation coefficient is often described as being "nonparametric". This can have two meanings: First, a perfect Spearman correlation results when X and Y are related by any monotonic function. Contrast this with the Pearson correlation, which only gives a perfect value when X and Y are related by a *linear* function. The other sense in which the Spearman correlation is nonparametric in that its exact sampling distribution can be obtained without requiring knowledge (*i.e.*, knowing the parameters) of the joint probability distribution of X and Y.

TIME SERIES ANALYSIS

Time Series is a sequence of well-defined data points measured at consistent time intervals over a period of time. Data collected on an ad-hoc basis or irregularly does not form a time series. Time series analysis is the use of statistical methods to analyze time series data and extract meaningful statistics and characteristics about the data. Time series Analysis helps us understand what are the underlying forces leading to a particular trend in the time series data points and helps us in forecasting and monitoring the data points by fitting appropriate models to it.

Statistical data which is recorded with its time of occurrence is called a time series. The yearly output of wheat recorded for the last twenty five years, the weekly average price of eggs recorded for the last 52 weeks, the monthly average sales of a firm recorded for the last 48 months or the quarterly average profits recorded for the last 40 quarters etc., are examples of time series data. It may be observed that this data undergoes changes with the passage of time. A number of factors can be isolated which contribute to changes occurring over time in such a series. In the fields of economics and business, data such as income, imports, exports, production, consumption, and prices depend on time. All of these data are dependent on seasonal changes as well as regular cyclical changes over the time period. To evaluate the changes in business and economics, the analysis of time series plays an important role in this regard. It is necessary to associate time with time series, because time is one basic variable in time series analysis.

Benefits and Applications of Time Series Analysis:

Time series analysis aims to achieve various objectives and the tools and models used vary accordingly. The various types of time series analysis include –

- **Descriptive analysis** – to determine the trend or pattern in a time series using graphs or other tools. This helps us identify cyclic patterns, overall trends, turning points and outliers.
- **Spectral analysis** – is also referred to as frequency domain and aims to separate periodic or cyclical components in a time series. For example, identifying cyclical changes in sales of a product.
- **Forecasting** – used extensively in business forecasting, budgeting, etc. based on historical trends.
- **Intervention Analysis** – is used to determine if an event can lead to a change in the time series, for example, an employee's level of performance has improved or not after an intervention in the form of training – to determine the effectiveness of the training program.
- **Explanative Analysis** – studies the cross correlation or relationship between two time series and the dependence of one on another. For example the study of employee turnover data and employee training data to determine if there is any dependence of employee training programs on employee turnover rates over time.

The Components of Time Series

The factors that are responsible for bringing about changes in a time series, also called the components of time series, are as follows:

1. Secular Trends (or General Trends)
2. Seasonal Movements
3. Cyclical Movements
4. Irregular Fluctuations or Erratic Trends

1. **Secular Trend or Long Term Variation:**
   It is a longer term change. Here we take into account the number of observations available and make a subjective assessment of what is ***long term***. To understand the meaning of long term, let

for example, climate variables sometimes exhibit *cyclic variation* over a very long time period such as 50 years. If one just had 20 years data, this long term oscillation would appear to be a trend, but if several hundred years of data is available, then long term oscillations would be visible. These movements are systematic in nature where the movements are broad, steady, showing slow rise or fall in the same direction. The ***trend*** may be linear or non-linear (curvilinear). Some examples of secular trend are: Increase in prices, increase in pollution, increase in the need of wheat, increase in literacy rate, and decrease in deaths due to advances in science. Taking averages over a certain period is a simple way of detecting trend in seasonal data. Change in averages with time is evidence of a trend in the given series, though there are more formal tests for detecting ***trend in time series***.

2. **Seasonal Variation or Seasonal Fluctuations:**
   Many of the *time series data* exhibits a ***seasonal variation*** which is annual period, such as sales and temperature readings. This type of variation is easy to understand and can be easily measured or removed from the data to give *de-seasonalized data*. ***Seasonal Fluctuations*** describes any *regular variation* (*fluctuation*) with a period of less than one year for example cost of variation types of fruits and vegetables, cloths, unemployment figures, average daily rainfall, increase in sale of tea in winter, increase in sale of ice cream in summer etc., all show ***seasonal variations***. The changes which repeat themselves within a fixed period, are also called ***seasonal variations***, for example, traffic on roads in morning and evening hours, Sales at festivals like EID etc., increase in the number of passengers at weekend etc. ***Seasonal variations*** are caused by climate, social customs, religious activities etc.
3. **Cyclical Variation or Cyclic Fluctuations:**
   Time series exhibits ***Cyclical Variations*** at a fixed period due to some other physical cause, such as daily variation in temperature. ***Cyclical variation*** is a non-seasonal component which varies in recognizable cycle. Sometime series exhibits oscillations which do not have a fixed period but are predictable to some extent. For example, economic data affected by business cycles with a period varying between about 5 and 7 years. In weekly or monthly data, the ***cyclical component*** may describe any *regular variation (fluctuations)* in *time series data*. The ***cyclical variation***s are periodic in nature and repeat themselves like business cycle, which has four phases
   (i) *Peak* (ii) *Recession* (iii) *Trough/Depression* (iv) *Expansion*.
4. **Irregular Fluctuations:**
   When *trend* and *cyclical variations* are removed from a set of time series data, the *residual* left, which may or may not be random. Various techniques for analyzing series of this type examine to see "if irregular variation may be explained in terms of probability models such as *moving average* or *autoregressive models*, i.e. we can see if any *cyclical variation* is still left in the ***residuals.*** These variation occur due to sudden causes are called ***residual variation*** (*irregular variation* or *accidental* or *erratic fluctuations*) and are unpredictable, for example rise in prices of steel due to strike in the factory, accident due to failure of break, flood, earth quick, war etc.

Mathematical Statement of the Composition of Time Series:

A time series may not be affected by all type of variations. Some of these types of variations may affect a few time series, while the other series may be affected by all of them. Hence, in analysing time series, these effects are isolated. In classical time series analysis it is assumed that any given observation is made up of trend, seasonal, cyclical and irregular movements and these four components have multiplicative relationship.

Symbolically:

$$O = T \times S \times C \times I$$

Where O refers to original data,

T refers to trend.

S refers to seasonal variations,
C refers to cyclical variations and
I refers lo irregular variations.
This is the most commonly used model in the decomposition of time series.

There is another model called Additive model in which a particular observation in a time series is the sum of these four components.

$$O = T + S + C + I$$

To prevent confusion between the two models, it should be made clear that in Multiplicative model S, C, and I are indices expressed as decimal percents whereas in Additive model S, C and I are quantitative deviations about trend that can be expressed as seasonal, cyclical and irregular in nature. If in a multiplicative model. T = 500, S = 1.4, C = 1.20 and I = 0.7 then

$$O = T \times S \times C \times I$$

By substituting the values we get

$$O = 500 \times 1.4 \times 1.20 \times 0.7 = 608$$

In additive model, T = 500, S = 100, C = 25, I = –50

$$O = 500 + 100 + 25 – 50 = 575$$

The assumption underlying the two schemes of analysis is that whereas there is no interaction among the different constituents or components under the additive scheme, such interaction is very much present in the multiplicative scheme. Time series analysis, generally, proceed on the assumption of multiplicative formulation.

Methods of Measuring Trend

Trend can be determined:

1. Free hand curve method ;
2. Moving averages method ;
3. Semi-averages method; and
4. Least-squares Method.

Each of these methods is described below:

1. Freehand Curve Method:

The term freehand is used to any non-mathematical curve in statistical analysis even if it is drawn with the aid of drafting instruments. This is the simplest method of studying trend of a time series. The procedure for drawing free hand curve is as follows:

(i) The original data are first plotted on a graph paper.
(ii) The direction of the plotted data is carefully observed.
(iii)A smooth line is drawn through the plotted points.

While fitting a trend line by the freehand method, an attempt should be made that the fitted curve conforms to these conditions.

- The curve should be smooth either a straight line or a combination of long gradual curves.
- The trend line or curve should be drawn through the graph of the data in such a way that the areas below and above the trend line are equal to each other.
- The vertical deviations of the data above the trend line must equal to the deviations below the line.
- Sum of the squares of the vertical deviations of the observations from the trend should be minimum.

Merits:

- It is very simplest method for study trend values and easy to draw trend.
- Sometimes the trend line drawn by the statistician experienced in computing trend may be considered better than a trend line fitted by the use of a mathematical formula.
- Although the free hand curves method is not recommended for beginners, it has considerable merits in the hands of experienced statisticians and widely used in applied situations.

Demerits:

- This method is highly subjective and curve varies from person to person who draws it.
- The work must be handled by skilled and experienced people.
- Since the method is subjective, the prediction may not be reliable.
- While drawing a trend line through this method a careful job has to be done.

2. Method of Moving Averages:

The moving average is a simple and flexible process of trend measurement which is quite accurate under certain conditions. This method establishes a trend by means of a series of averages covering overlapping periods of the data.

The process of successively averaging, say, three years data, and establishing each average as the moving-average value of the central year in the group, should be carried throughout the entire series. For a five-item, seven-item or other moving averages, the same procedure is followed: the average obtained each time being considered as representative of the middle period of the group. The choice of a 5-year, 7-year, 9-year, or other moving average is determined by the length of period necessary to eliminate the effects of the business cycle and erratic fluctuations. A good trend must be free from such movements, and if there is any definite periodicity to the cycle, it is well to have the moving average to cover one cycle period. Ordinarily, the necessary periods will range between three and ten years for general business series but even longer periods are required for certain industries.

In the preceding discussion, the moving averages of odd number of years were representatives of the middle years. If the moving average covers an even number of years, each average will still be representative of the midpoint of the period covered, but this mid-point will fall halfway between the two middle years. In the case of a four-year moving average, for instance each average represents a point halfway between the second and third years. In such a case, a second moving average may be used to 're-centre' the averages.

That is, if the first moving average gives averages centering half-way between the years, a further two-point moving average will re-centre the data exactly on the years. This method, however, is valuable in approximating trends in a period of transition when the mathematical lines or curves may be inadequate. This method provides a basis for testing other types of trends, even though the data are not such as to justify its use otherwise.

Merits

1. This is a very simple method.
2. The element of flexibility is always present in this method as all the calculations have not to be altered if same data is added. It only provides additional trend values.
3. If there is a coincidence of the period of moving averages and the period of cyclical fluctuations, the fluctuations automatically disappear.
4. The pattern of moving average is determined in the trend of data and remains unaffected by the choice of method to be employed.
5. It can be put to utmost use in case of series having strikingly irregular trend.

Limitations:

1. It is not possible to have a trend value for each and every year. As the period of moving average increases, there is always an increase in the number of years for which trend values cannot be calculated and known. For example, in a five yearly moving average, trend value cannot be obtained for the first two years and last two years, in a seven yearly moving average for the first three years and last three years and so on. But usually values of the extreme years are of great interest.
2. There is no hard and fast rule for the selection of a period of moving average.
3. Forecasting is one of the leading objectives of trend analysis. But this objective remains unfulfilled because moving average is not represented by a mathematical function.
4. Theoretically it is claimed that cyclical fluctuations are ironed out if period of moving average coincide with period of cycle, but in practice cycles are not perfectly periodic.

3. Method of Semi Averages:

In this method the whole data is divided in two equal parts with respect to time. For example if we are given data from 1979 to 1996 i.e. over a period of 18 years the two equal parts will be first nine years i.e. from 1979 to 1987 and 1988 to 1996. In case of odd number of years like 9, 13, 17 etc. two equal parts can be made simply by omitting the middle year. For example if the data are given for 19 years from 1978 to 1996 the two equal parts would be from 1978 to 1986 and from 1988 to 1996, the middle year 1987 will be omitted. After the data have been divided into two parts, an average (arithmetic mean) of each part is obtained. We thus get two points. Each point is plotted against the mid year of the each part. Then these two points are joined by a straight line which gives us the trend line. The line can be extended downwards or upwards to get intermediate values or to predict future values.

Merits:

- This method is simple to understand as compare to moving average method and method of least squares.
- This is an objective method of measuring trend as everyone who applies this method is bound to get the same result.

Demerits:

- The method assumes straight line relationship between the plotted points regardless of the fact whether that relationship exists or not.
- The main drawback of this method is if we add some more data to the original data then whole calculation is to be done again for the new data to get the trend values and the trend line also changes.
- As the A.M of each half is calculated, an extreme value in any half will greatly affect the points and hence trend calculated through these points may not be precise enough for forecasting the future.

4. Method of Least Squares:

If a straight line is fitted to the data it will serve as a satisfactory trend, perhaps the most accurate method of fitting is that of least squares. This method is designed to accomplish two results.

(i) The sum of the vertical deviations from the straight line must equal zero.

(ii) The sum of the squares of all deviations must be less than the sum of the squares for any other conceivable straight line.

There will be many straight lines which can meet the first condition. Among all different lines, only one line will satisfy the second condition. It is because of this second condition that this method

is known as the method of least squares. It may be mentioned that a line fitted to satisfy the second condition, will automatically satisfy the first condition.

The formula for a straight-line trend can most simply be expressed as

$$\mathbf{Y_c = a + bX}$$

Where X represents time variable, Yc is the dependent variable for which trend values are to be calculated and a and b are the constants of the straight tine to be found by the method of least squares.

Constant is the Y-intercept. This is the difference between the point of the origin (O) and the point of the trend line and Y-axis intersect. It shows the value of Y when X = 0, constant b indicates the slope which is the change in Y for each unit change in X. Let us assume that we are given observations of Y for n number of years. If we wish to find the values of constants a and b in such a manner that the two conditions laid down above are satisfied by the fitted equation.

Merits:

- This is a mathematical method of measuring trend and as such there is no possibility of subjectiveness i.e. everyone who uses this method will get same trend line.
- The line obtained by this method is called *the line of best fit.*
- Trend values can be obtained for all the given time periods in the series.

Demerits:

- Great care should be exercised in selecting the type of trend curve to be fitted i.e. linear, parabolic or some other type. Carelessness in this respect may lead to wrong results.
- The method is more tedious and time consuming.
- Predictions are based on only long term variations i.e. trend and the impact of cyclical, seasonal and irregular variations is ignored.

REGRESSION ANALYSIS

The Regression Analysis is a statistical tool used to determine the probable change in one variable for the given amount of change in another. This means, the value of the unknown variable can be estimated from the known value of another variable. The degree to which the variables are correlated to each other depends on the Regression Line. The regression line is a single line that best fits the data, i.e. all the points plotted are connected via a line in the manner that the distance from the line to the points is the smallest.

The regression also tells about the relationship between the two or more variables, then what is the difference between regression and correlation? Well, there are two important points of differences between Correlation and Regression. These are:

- The Correlation Coefficient measures the "degree of relationship" between variables, say X and Y whereas the Regression Analysis studies the "nature of relationship" between the variables.
- Correlation coefficient does not clearly indicate the cause-and-effect relationship between the variables, i.e. it cannot be said with certainty that one variable is the cause, and the other is the effect. Whereas, the Regression Analysis clearly indicates the cause-and-effect relationship between the variables.

The regression analysis is widely used in all the scientific disciplines. In economics, it plays a significant role in measuring or estimating the relationship among the economic variables. For example, the two variables – price (X) and demand (Y) are closely related to each other, so we can find out the probable value of X from the given value of Y and similarly the probable value of Y can be found out from the given value of X.

The relationship between two variables may be interested in estimating (predicting) the value of one variable given the value of another. The variable predicted on the basis of other variables is called the "dependent" or the „explained‟ variable and the other the „independent‟ or the „predicting‟ variable. The prediction is based on average relationship derived statistically by regression analysis. The equation, linear or otherwise, is called the regression equation or the explaining equation.

For example, if we know that advertising and sales are correlated we may find out expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales.

The relationship between two variables can be considered between, say, rainfall and agricultural production, price of an input and the overall cost of product consumer expenditure and disposable income. Thus, regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

Types of Regression:

The regression analysis can be classified in to:

a) Simple and Multiple
b) Linear and Non –Linear
c) Total and Partial

a) Simple and Multiple:

In case of simple relationship only two variables are considered, for example, the influence of advertising expenditure on sales turnover. In the case of multiple relationship, more than two variables are involved. On this while one variable is a dependent variable the remaining variables are independent ones. For example, the turnover (y) may depend on advertising expenditure (x) and the income of the people (z). Then the functional relationship can be expressed as y = f (x,z).

b) Linear and Non-linear:

The linear relationships are based on straight-line trend, the equation of which has no-power higher than one. But, remember a linear relationship can be both simple and multiple. Normally a linear relationship is taken into account because besides its simplicity, it has a better predictive value; a linear trend can be easily projected into the future. In the case of non-linear relationship curved trend lines are derived. The equations of these are parabolic.

c) Total and Partial:

In the case of total relationships all the important variables are considered. Normally, they take the form of a multiple relationships because most economic and business phenomena are affected by multiplicity of cases. In the case of partial relationship one or more variables are considered, but not all, thus excluding the influence of those not found relevant for a given purpose.

Properties of Regression Co-efficient:

1. Both regression coefficients must have the same sign, i.e. either it will be positive or negative.
2. Correlation coefficient is the geometric mean of the regression coefficients i.e. $r = \pm\sqrt{b_1 b_2}$
3. The correlation coefficient will have the same sign as that of the regression coefficients.
4. If one regression coefficient is greater than unity, then other regression coefficient must be less than unity.
5. Regression coefficients are independent of origin but not of scale.
6. Arithmetic mean of b_1 and b_2 is equal to or greater than the coefficient of correlation. Symbolically $\frac{b1+b2}{2} \geq r$.
7. If r = 0, the variables are uncorrelated, the lines of regression become perpendicular to each other.
8. If r = ±1, the two lines of regression either coincide or parallel to each other
9. Angle between the two regression lines is $\theta = \tan^{-1} \frac{m1-m2}{1+m_1 m_2}$ where m_1 and, m_2 are the slopes of the regression lines X on Y and Y on X respectively.
10. The angle between the regression lines indicates the degree of dependence between the variables.

Difference between Correlation and Regression

Correlation	Regression
It is the relationship between two or more variables, which vary in sympathy with the other in the same or the opposite direction.	It means going back and it is a mathematical measure showing the average relationship between two variables.
Both the variables X and Y are random variables.	Here X is a random variable and Y is a fixed variable. Sometimes both the variables may be random variables.
It finds out the degree of relationship between two variables and not the cause and effect of the variables.	It indicates the causes and effect relationship between the variables and establishes functional relationship.
It is used for testing and verifying the relation between two variables and gives limited information.	Besides verification it is used for the prediction of one value, in relationship to the other given value.
The coefficient of correlation is a relative measure. The range of relationship lies between –1 and +1.	It is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable.
There may be spurious correlation between two	In regression there is no such spurious

variables.	regression.
It has limited application, because it is confined only to linear relationship between the variables.	It has wider application, as it studies linear and nonlinear relationship between the variables.
It is not very useful for further mathematical treatment.	It is widely used for further mathematical treatment.
If the coefficient of correlation is positive, then the two variables are positively correlated and vice-versa.	The regression coefficient explains that the decrease in one variable is associated with the increase in the other variable.

Linear Regression Equation:

If two variables have linear relationship then as the independent variable (X) changes, the dependent variable (Y) also changes. If the different values of X and Y are plotted, then the two straight lines of best fit can be made to pass through the plotted points. These two lines are known as regression lines. Again, these regression lines are based on two equations known as regression equations. These equations show best estimate of one variable for the known value of the other. The equations are linear.

Linear regression equation of Y on X is

$$Y = a + b\,X \ldots\ldots. (1)$$

And X on Y is

$$X = a + b\,Y \ldots\ldots. (2)$$

a, b are constants.

From (1) We can estimate Y for known value of X.

(2) We can estimate X for known value of Y.

The Regression Equation is the algebraic expression of the regression lines. It is used to predict the values of the dependent variable from the given values of independent variables. If we take two regression lines, say Y on X and X on Y, then there will be two regression equations:

- **Regression Equation of Y on X:** This is used to describe the variations in the value Y from the given changes in the values of X. It can be expressed as follows:

$$\mathbf{Y_c = a + bX}$$

Where Y_e is the dependent variable, X is the independent variable, and a & b are the two unknown constants that determine the position of the line. The parameter "a" tells about the level of the fitted line, i.e. the distance of a line above or below the origin and parameter "b" tells about the slope of the line, i.e. the change in the value of Y for one unit of change in X.

The values of „a‟ and „b‟ can be obtained by a method of least squares. According to which the line should be drawn connecting all the plotted points in such a manner that the sum of the squares of the vertical deviations of actual Y from the estimated values of Y is the least, or a best-fitted line is obtained when $\sum \mathbf{(Y-Y_e)^2}$ is the minimum.

The following algebraic equations can be solved simultaneously to obtain the values of parameter „a‟ and „b‟.

$$\sum \mathbf{Y = Na + b \sum X}$$

$$\sum \mathbf{XY = a \sum X + b \sum X^2}$$

- **Regression Equation of X on Y:** This is used to describe the variations in Y from the given changes in the value of X. It can be expressed as follows:

$$\mathbf{X_c = a + bY}$$

Where X_e is the dependent variable and Y is the independent variable. The parameters „a‟ and „b‟ are the two unknown constants. Again, „a‟ tells about the level of fitted line and „b‟ tells about the slope, i.e. the change in the value of X for a unit change in the value of Y.

The following are the two normal equations that can be solved simultaneously to obtain the values of both the parameters „a‟ and „b‟.

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Note: The line can be completely determined only if the values of the constant parameters „a‟ and „b‟ are obtained.

Regression Lines:

The Regression Line is the line that best fits the data, such that the overall distance from the line to the points (variable values) plotted on a graph is the smallest. In other words, a line used to minimize the squared deviations of predictions is called as the regression line.

There are as many numbers of regression lines as variables. Suppose we take two variables, say X and Y, then there will be two regression lines:

- Regression line of Y on X: This gives the most probable values of Y from the given values of X.
- Regression line of X on Y: This gives the most probable values of X from the given values of Y.

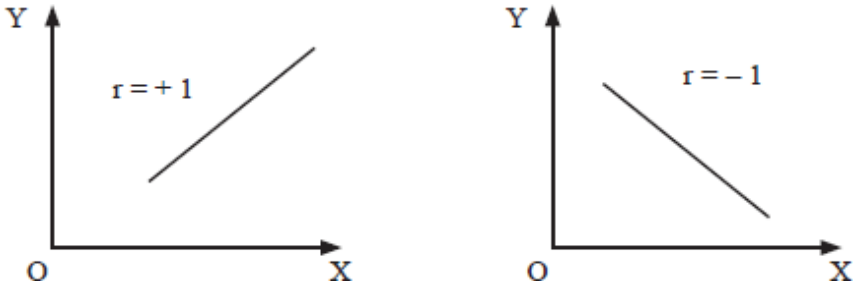
The algebraic expression of these regression lines is called as Regression Equations. There will be two regression equations for the two regression lines.

The correlation between the variables depend on the distance between these two regression lines, such as the nearer the regression lines to each other the higher is the degree of correlation, and the farther the regression lines to each other the lesser is the degree of correlation.

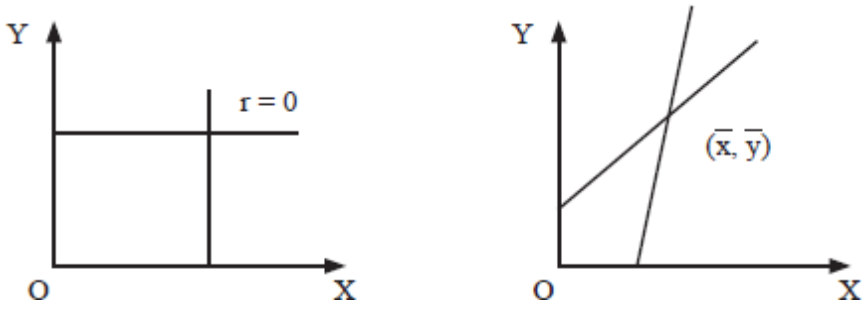
The correlation is said to be either perfect positive or perfect negative when the two regression lines coincide, i.e. only one line exists. In case, the variables are independent; then the correlation will be zero, and the lines of regression will be at right angles, i.e. parallel to the X axis and Y axis.

Note: The regression lines cut each other at the point of average of X and Y. This means, from the point where the lines intersect each other the perpendicular is drawn on the X axis we will get the mean value of X. Similarly, if the horizontal line is drawn on the Y axis we will get the mean value of Y.

For regression analysis of two variables there are two regression lines, namely Y on X and X on Y. The two regression lines show the average relationship between the two variables. For perfect correlation, positive or negative i.e., r = + 1, the two lines coincide i.e., we will find only one straight line. If r = 0, i.e., both the variables are independent then the two lines will cut each other at right angle. In this case the two lines will be parallel to X and Y-axes.



Lastly the two lines intersect at the point of means of X and Y. From this point of intersection, if a straight line is drawn on X-axis, it will touch at the mean value of x. Similarly, a perpendicular drawn from the point of intersection of two regression lines on Y-axis will touch the mean value of Y.



Principle of „Least Squares":

Regression shows an average relationship between two variables, which is expressed by a line of regression drawn by the method of "least squares". This line of regression can be derived graphically or algebraically. Before we discuss the various methods let us understand the meaning of least squares.

A line fitted by the method of least squares is known as the line of best fit. The line adapts to the following rules:

(i) The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero. i.e.,

$$\Sigma(X - X_c) = 0 \text{ or } \Sigma (Y - Y_c) = 0$$

Where Xc and Yc are the values obtained by regression analysis.

(ii) The sum of the squares of these deviations is less than the sum of squares of deviations from any other line. i.e.,

$$\Sigma (Y - Y_c)^2 < \Sigma (Y - A_i)^2$$

Where Ai = corresponding values of any other straight line.

(iii) The lines of regression (best fit) intersect at the mean values of the variables X and Y, i.e., intersecting point is $\bar{x}, \bar{y}$

.

UNIT-5

SMALL SAMPLE TESTS

OBJECTIVE

To know various Sample Tests

SMALL SAMPLE TESTS

- t-distribution
- Properties and Applications
- One and Two Sample Mean Tests
- Paired t-Test

ANALYSIS OF VARIANCE

- One way Analysis
- Two way Analysis

CHI-SQUARE DISTRIBUTION

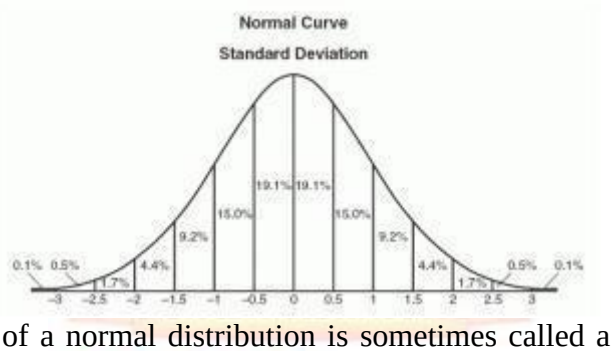
- Test for specified Variance
- Test for Independence of Attributes

STUDENT'S t-DISTRIBUTION (t-test)

Theoretical work on t-distribution was done by Sir W. S. Gosset (1876-1937). The t-distribution is a theoretical probability distribution. The t-distribution, also known as the student's t-distribution, is used when making inferences about a mean when the standard deviation is not known. It is symmetrical, bell-shaped, and similar to the standard normal curve. The higher the degrees of freedom, the closer that distribution will resemble a standard normal distribution with a mean of 0, and a standard deviation of 1. Thus, student distribution is the statistical measure that compares the observed data with the expected data obtained with a specific hypothesis. It complies with the central limit theorem which says that the distribution approaches the standard normal distribution as long as the sample size is large.

Normal Distribution:

A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme. Height is one simple example of something that follows a normal distribution pattern: Most people are of average height the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short. Here's an example of a normal distribution curve:

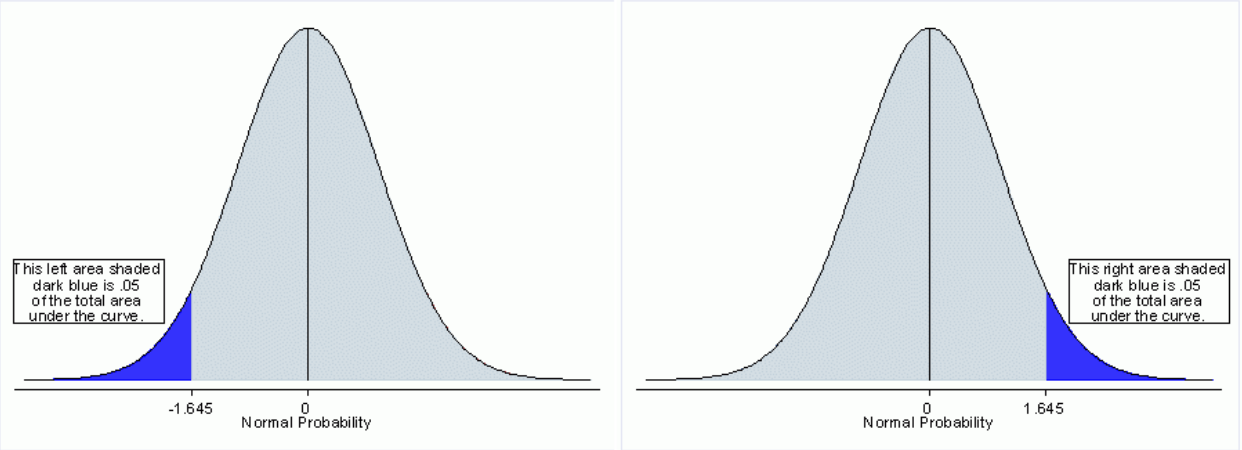


A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution the mean mode and median are all the same.

One-Tailed Test:

A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample that is being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis. One-tailed test is also known as a directional hypothesis or test.

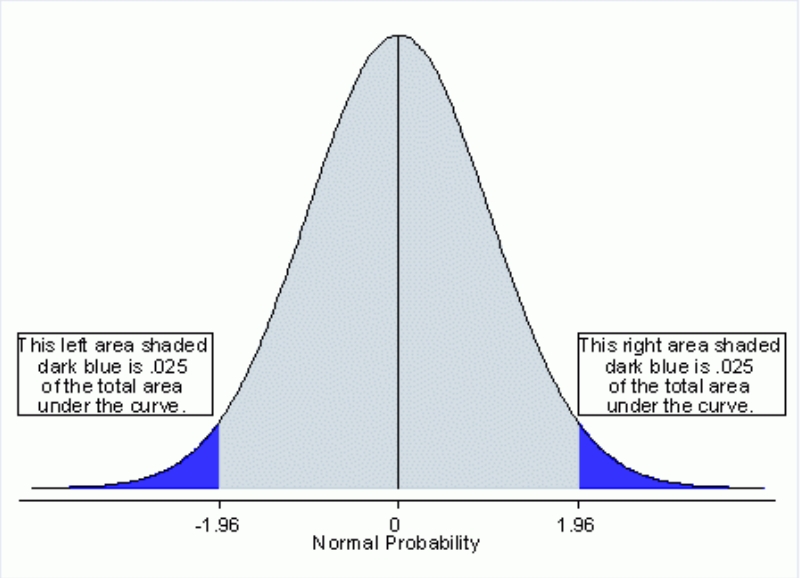
If you are using a significance level of .05, a one-tailed test allots your entire alpha to testing the statistical significance in the one direction of interest. This means that .05 is in one tail of the distribution of your test statistic. When using a one-tailed test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction. Let's return to our example comparing the mean of a sample to a given value x using a t-test. Our null hypothesis is that the mean is equal to x. A one-tailed test will test either if the mean is significantly greater than x or if the mean is significantly less than x, but not both. Then, depending on the chosen tail, the mean is significantly greater than or less than x if the test statistic is in the top 5% of its probability distribution or bottom 5% of its probability distribution, resulting in a p-value less than 0.05. The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction.



Because the one-tailed test provides more power to detect an effect, you may be tempted to use a one-tailed test whenever you have a hypothesis about the direction of an effect. Before doing so, consider the consequences of missing an effect in the other direction. Imagine you have developed a new drug that you believe is an improvement over an existing drug. You wish to maximize your ability to detect the improvement, so you opt for a one-tailed test. In doing so, you fail to test for the possibility that the new drug is less effective than the existing drug. The consequences in this example are extreme, but they illustrate a danger of inappropriate use of a one-tailed test.

Choosing a one-tailed test for the sole purpose of attaining significance is not appropriate. Choosing a one-tailed test after running a two-tailed test that failed to reject the null hypothesis is not appropriate, no matter how "close" to significant the two-tailed test was. Using statistical tests inappropriately can lead to invalid results that are not replicable and highly questionable–a steep price to pay for a significance star in results table.

Two-Tailed Test: A two-tailed test is a statistical test in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values. If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis. The two-tailed test gets its name from testing the area under both of the tails of a normal distribution, although the test can be used in other non-normal distributions. If you are using a significance level of 0.05, a two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction. This means that .025 is in each tail of the distribution of your test statistic. When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions. For example, we may wish to compare the mean of a sample to a given value x using a t-test. Our null hypothesis is that the mean is equal to x. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x. The mean is considered significantly different from x if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.



Difference between One-Tailed and Two-Tailed Test

Basis	ONE-TAILED TEST	TWO-TAILED TEST
Meaning	A Statistical hypothesis in which alternative hypothesis has only one end.	A significance in which alternative hypothesis has two ends.
Hypothesis	Directional	Non-directional
Region of Rejection	Either Left or Right	Both Left or Right
Determines	If there is a relationship between variables in single direction.	If there is a relationship between variables in either direction.
Result	Greater or Less than certain value.	Greater or Less than certain range of values.
Sign in alternative hypothesis	(>) or (<)	≠

Dependent and Independent Samples:

- Dependent samples are paired measurements for one set of items.
- Independent samples are measurements made on two different sets of items.
- When you conduct a hypothesis test using two random samples, you must choose the type of test based on whether the samples are dependent or independent. Therefore, it's important to know whether your samples are dependent or independent:
  - ✓ If the values in one sample affect the values in the other sample, then the samples are dependent.
  - ✓ If the values in one sample reveal no information about those of the other sample, then the samples are independent.

Example of collecting dependent samples and independent samples:

Consider a drug company that wants to test the effectiveness of a new drug in reducing blood pressure. They could collect data in two ways:

- Sample the blood pressures of the same people before and after they receive a dose. The two samples are dependent because they are taken from the same people. The people with the highest

blood pressure in the first sample will likely have the highest blood pressure in the second sample.

- Give one group of people an active drug and give a different group of people an inactive placebo, then compare the blood pressures between the groups. These two samples would likely be independent because the measurements are from different people. Knowing something about the distribution of values in the first sample doesn't inform you about the distribution of values in the second.

Assumptions of Students t-test:

- The sample is drawn from the Normal population
- The sample observations are independent
- The population standard deviation σ is unknown.

Properties of t-Distribution:

1. Like, standard normal distribution the shape of the student distribution is also bell-shaped and symmetrical with mean zero.
2. The student distribution ranges from $-\infty$ to $+\infty$ (infinity).
3. The shape of the t-distribution changes with the change in the degrees of freedom.
4. The variance is always greater than one and can be defined only when the degrees of freedom $\nu \geq$ **3** and is given as: Var (t) $= \frac{\nu}{\nu-2}$
5. It is less peaked at the center and higher in tails, thus it assumes platykurtic shape.
6. The t-distribution has a greater dispersion than the standard normal distribution. And as the sample size „n‟ increases, it assumes the normal distribution. Here the sample size is said to be large when $n \geq 30$.

Degrees of Freedom

It refers to the number of values involved in the calculations that have the freedom to vary. In other words, the degrees of freedom, in general, can be defined as the total number of observations minus the number of independent constraints imposed on the observations.

The degrees of freedom are calculated for the following statistical tests to check their validity:

1. t-Distribution
2. F- Distribution
3. Chi-Square Distribution

These tests are usually done to compare the observed data with the data that is expected to be obtained with a specific hypothesis.

It is usually denoted by a Greek symbol ν (mu) and is commonly abbreviated as, *df*. The statistical formula to compute the value of degrees of freedom is quite simple and is equal to the number of values in the data set minus one. Symbolically:

df= n-1

Where n is the number of values in the data set or the sample size. The concept of df can be further understood through an illustration given below:

Suppose there is a data set X that includes the values: 10, 20, 30, 40. First of all, we will calculate the mean of these values, which is equal to:

(10+20+30+40) /4 = 25.

Once the mean is calculated, apply the formula of degrees of freedom. As the number of values in the data set or sample size is 4, so,

df = 4-1=3.

Thus, this shows that there are three values in the data set that have the freedom to vary as long as the mean is 25.

The following are the important Applications of the t-distribution:

1. Test of the Hypothesis of the population mean.
2. Test of Hypothesis of the difference between the two means.
3. Test of Hypothesis of the difference between two means with dependent samples.
4. Test of Hypothesis about the coefficient of correlation.

APPLICATION-1
TEST OF HYPOTHESIS OF THE POPULATION MEAN

When the population is normally distributed, and the standard deviation „σ" is unknown, then "t" statistic is calculated as:

$t = \frac{\bar{X} - \mu}{S}\sqrt{n}$	$S = \sqrt{\sum \frac{(X - \bar{X})^2}{(n-1)}}$	$\bar{X}$ = Sample Mean μ = Population Mean n = Sample size S = SD of the sample

The null hypothesis is tested to check whether there is a significant difference between the $\bar{X}$ and μ. If the calculated value of „t" exceeds the table value of „t" at a specific significance level, then the null hypothesis is rejected considering the difference between the $\bar{X}$ and μ as significant. On the other hand, if the calculated value of „t" is less than the table value of „t", then the null hypothesis is accepted. It is to be noted that this test is based on the degrees of freedom, i.e. n-1.

Fiducial Limits of Population Mean:

Assuming that the sample is a random sample from a normal population of unknown mean the 95% and 99% fiducial limits of the population of the mean (μ) are

@ 95% Limit $= \bar{X} \pm \frac{S}{\sqrt{n}} t_{0.05}$	@ 99% Limit $= \bar{X} \pm \frac{S}{\sqrt{n}} t_{0.01}$

APPLICATION-2
TEST OF HYPOTHESIS OF THE DIFFERENCE BETWEEN TWO MEANS

In Testing hypothesis about the difference between two means drawn from the two systematic population whose variance is unknown, then t-test can be calculated in two ways:

Variances are equal:

When the population variances, though unknown are taken as equal, then the t- statistic to be used is:

$t = \frac{\bar{X} - \bar{X}_2}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$	$S = \sqrt{\frac{\sum(X_1 - \bar{X})^2 + \sum(X_2 - \bar{X}_2)^2}{(n_1 + n_2 - 2)}}$

$$S = \sqrt{\frac{\sum(X_1 - A_1)^2 + \sum(X_2 - A_2)^2 - n_1(\bar{X}_1 - A_1)^2 - n_2(\bar{X}_2 - A_2)^2}{(n_1 + n_2 - 2)}}$$

Where,
$\bar{X}_1$ and $\bar{X}_2$ are the sample means of sample 1 of size n_1 and sample 2 of size n_2.
S is the common standard deviation obtained by pooling the data from both the samples.
The null hypothesis is that there is no difference between two means and" is accepted when the calculated value of „t" at a specified significance level is less than the table value of „t" and is rejected when the calculated value exceeds the table value.

Variances are Unequal:
When the population variances are not equal, then we use the unbiased estimators S_1^2 and S_2^2. In this case, the sampling has the huge variability than the population variability and statistic to be used is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

$$d.f. = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}\right]}$$

Where,
μ_1 and μ_2 are the two population means.
This statistic may not strictly follow t-distributions, but however it can be approximated by t-distribution with the modified value for the degrees of freedom given above:

APPLICATION 3
TEST OF HYPOTHESIS OF THE DIFFERENCE BETWEEN TWO MEANS WITH DEPENDENT SAMPLES

In several situations, it is possible that the samples are drawn from the two populations that are dependent on each other. Thus, the samples are said to be dependent, as each observation included in sample one is associated with the particular observation in the second sample. Hence, due to this property the t-test that will be used here is called the paired t-test.
This test is applied in the situations when before and after experiments are to be compared. Usually, two methods are adopted that are related to each other. The following statistic is used when the means of both the methods applied is equal i.e. $\mu_1 = \mu_2$

$$t = \frac{\bar{d}\sqrt{n}}{S}$$

This statistic follows t- distribution with (n-1) degrees of freedom, where d = mean of the differences calculated as:

$$\bar{d} = \frac{\sum d}{n}$$

S is the standard deviation of differences and is calculated by applying the following formula:

$$S = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{(n-1)}}$$

n = Number of paired observations.

APPLICATION 4
TEST OF HYPOTHESIS ABOUT THE COEFFICIENT OF CORRELATION

There are three cases of testing the hypothesis about the coefficient of correlation. These are:

Case-1:
When the population coefficient of correlation is zero, i.e. ρ = 0. The coefficient of correlation measures the degree of relationship between the variables, and when ρ = 0, then there is no statistical relationship between the variables. To test the null hypothesis which assumes that there is no correlation between the populations, it is necessary that the sample coefficient of correlation „r" is known. The test statistic to be used is:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

Case -2:
When the Population Coefficient of Correlation is equal to some other value, other than zero, i.e. ρ≠0. In this case, the test based on t-distribution will not be correct and hence the hypothesis is tested using the Fisher"s z- transformation. Here the „r" is transformed into „z" by:

$$\mathbf{z = \frac{1}{2}\log_e \frac{1+r}{1-r}}$$

Here, $\log_e$ is a natural logarithm. The common logarithm can be shifted to a natural algorithm by multiplying it by the factor 2.3026. Thus, $\log_e X = 2.3026 \log_{10} X$, where X is the positive integer since, ½ X (2.3026) = 1.1513, then the following transformation formula is used:

$$\mathbf{z = 1.1513 \log_{10} \frac{1+r}{1-r}}$$

The following statistic is used to test the null hypotheses:

$$\mathbf{z = \frac{z - z_\rho}{\sigma_x}}$$

This follows the normal distribution and the test is said to be more appropriate as long as the sample size is large.

Case-3:
When the hypothesis is tested for the difference between two Independent Correlation Coefficients: To test the hypothesis of two correlations derived from the two separate samples, then the difference of the two corresponding values of z is to be compared with the standard error of the difference. The following statistic is used:

$$\mathbf{z = \frac{z_1 - z_2}{\sigma_{z_1} - z_2} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}}$$

ANALYSIS OF VARIANCE (ANOVA)

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis. Basically, you're testing groups to see if there's a difference between them.

Examples of when you might want to test different groups:

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- Students from different colleges take the same exam. You want to see if one college outperforms the other.

Assumptions of ANOVA:

1. **Normality of Errors**
   We assume in the ANOVA model that the error terms are normally-distributed with zero mean. If the data are not normally-distributed, but instead come from some other distribution (exponential or binomial, for example), then we may not be able to trust our *p*-values, which were built by assuming normality. "If I was to repeat my sample repeatedly and calculate the means, those means would be normally distributed."
2. **Equal Error Variance Across Treatments**
   The next assumption is that all error terms have the same variance σ^2. It is common to see that different treatments seem to result in different means AND ALSO different variances.
3. **Independence of Errors**
   We assume that each trial is independent of all other trials, except for the effect τi of the treatment on the mean. Statistical independence of two trials means that knowing the result of one trial doesn't change the distribution of the other trial. The most common causes of dependence in experimental data are confounding factors - something measured or unmeasured that affects the experiment. Randomization is a critical technique in experimental design because it can minimize the effect of any confounder. "Your samples have to come from a randomized or randomly sampled design."

Types of Tests

There are two main types: one-way and two-way. Two-way tests can be with or without replication.

1. One-way ANOVA between groups: used when you want to test **two groups** to see if there's a difference between them.
2. Two-way ANOVA without replication: used when you have **one group** and you're **double-testing** that same group. For example, you're testing one set of individuals before and after they take a medication to see if it works or not.
3. Two-way ANOVA with replication: **Two groups**, and the members of those groups are **doing more than one thing**. For example, two groups of patients from different hospitals trying two different therapies.

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 two or more independent (unrelated) groups (although you tend to only see it used when there are a minimum of three, rather than two groups). For example, you could use a one-way ANOVA to understand whether exam performance differed based on test anxiety levels amongst students, dividing students into three independent groups (e.g., low, medium and high-stressed students). Also, 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; it only tells you that at least two groups were different. Since you may have three, four, five or more groups in your study design, determining which of these groups differ from each other is important.

When to use a one way ANOVA?

Situation 1: You have a group of individuals randomly split into smaller groups and completing different tasks. For example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea.

Situation 2: Similar to situation 1, but in this case the individuals are split into groups based on an attribute they possess. For example, you might be studying leg strength of people according to weight. You could split participants into weight categories (obese, overweight and normal) and measure their leg strength on a weight machine.

Assumptions of One-way ANOVA

- Normal distribution of the population from which the samples are drawn.
- Measurement of the dependent variable is at interval or ratio level.
- Two or more than two categorical independent groups in an independent variable.
- Independence of samples
- Homogeneity of the variance of the population.

Common Uses

The One-Way ANOVA is often used to analyze data from the following types of studies:

- Field studies
- Experiments
- Quasi-experiments

The One-Way ANOVA is commonly used to test the following:

- Statistical differences among the means of two or more groups
- Statistical differences among the means of two or more interventions
- Statistical differences among the means of two or more change scores

Note: Both the One-Way ANOVA and the Independent Samples *t* Test can compare the means for two groups. However, only the One-Way ANOVA can compare the means across three or more groups.

Note: If the grouping variable has only two groups, then the results of a one-way ANOVA and the independent samples t test will be equivalent. In fact, if you run both an independent samples t test and a one-way ANOVA in this situation, you should be able to confirm that $t^2=F$.

Calculation and Decision Rules:

Hypotheses	The null hypothesis will be that all population means are equal, the alternative hypothesis is that at least one mean is different. In the following, lower case letters apply to the individual samples and capital letters apply to the entire set collectively. That is, n is one of many sample sizes, but N is the total sample size.
Grand Mean	The grand mean of a set of samples is the total of all the data values divided by the total sample size. This requires that you have all of the sample data available to you, which is usually the case, but not always. It turns out that all that is necessary to find perform a one-way analysis of variance are the number of samples, the sample means, the sample variances, and the sample sizes. Another way to find the grand mean is to find the weighted average of the sample means. The weight applied is the sample size.
Total Variation	The total variation (not variance) is comprised the sum of the squares of the differences of each mean with the grand mean. There is the between group variation and the within group variation. The whole idea behind the analysis of variance is to compare the ratio of between group variance to within group variance. If the variance caused by the interaction between the samples is much larger when compared to the variance that appears within each group, then it is because the means aren't the same.
Between Group Variation	The variation due to the interaction between the samples is denoted SS(B) for Sum of Squares Between groups. If the sample means are close to each other (and therefore the Grand Mean) this will be small. There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom. The variance due to the interaction between the samples is denoted MS(B) for Mean Square Between groups. This is the between group variation divided by its degrees of freedom.
Within Group Variation	The variation due to differences within individual samples, denoted SS(W) for Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. The degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: df = N - k. The variance due to the differences within individual samples is denoted MS(W) for Mean Square Within groups. This is the within group variation divided by its degrees of freedom. It is the weighted average of the variances (weighted with the degrees of freedom).
F test statistic	F variable is the ratio of two independent chi-square variables divided by their respective degrees of freedom. Also recall that the F test statistic is the ratio of two sample variances, well, it turns out that's exactly what we have here. The F test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the between group (k-1) and the degrees of freedom for the denominator are the degrees of freedom for the within group (N-k).
Decision Rule	The decision will be to reject the null hypothesis if the test statistic from the table is greater than the F critical value with k-1 numerator and N-k denominator degrees of freedom. If the decision is to reject the null, then at least one of the means is different. However, the ANOVA does not tell you where the difference lies.

ANOVA Table: One-Way Classification

Source of Variation	SS (Sum of Squares)	V (Degrees of freedom)	MS (Mean Square)	Variance Ratio of F
Between Samples	SSB	c-1	MSB	
Within Samples	SSW	n-c	MSW	MSB/MSW
Total	SST	n-1		

TWO WAY ANOVA

A Two Way ANOVA is an extension of the One Way ANOVA. With a One Way, you have one variable affecting a dependent variable. With a Two Way ANOVA, there are two independents. Use a two way ANOVA when you have one measurement variable (i.e. a quantitative variable) and two nominal variables. In other words, if your experiment has a quantitative outcome and you have two categorical explanatory variables, a two way ANOVA is appropriate.

For example, you might want to find out if there is an interaction between income and gender for anxiety level at job interviews. The anxiety level is the outcome, or the variable that can be measured. Gender and Income are the two categorical variables. These categorical variables are also the independent variables, which are called **factors** in a Two Way ANOVA.

The factors can be split into **levels**. In the above example, income level could be split into three levels: low, middle and high income. Gender could be split into three levels: male, female, and transgender. Treatment groups and all possible combinations of the factors. In this example there would be 3 x 3 = 9 treatment groups.

Assumptions of Two-Way ANOVA:

- Normal distribution of the population from which the samples are drawn.
- Measurement of dependent variable at continuous level.
- Two or more than two categorical independent groups in two factors.
- Categorical independent groups should have the same size.
- Independence of observations
- Homogeneity of the variance of the population.

Calculation and Decision Rules:

Hypotheses	There are three sets of hypothesis with the two-way ANOVA. The null hypotheses for each of the sets are given below. 1. The population means of the first factor are equal. This is like the one-way ANOVA for the row factor. 2. The population means of the second factor are equal. This is like the one-way ANOVA for the column factor. 3. There is no interaction between the two factors. This is similar to performing a test for independence with contingency tables.
Factors	The two independent variables in a two-way ANOVA are called factors. The idea is that there are two variables, factors, which affect the dependent variable. Each factor will have two or more levels within it, and the degrees of freedom for each factor is one less than the number of levels.
Treatment Groups	Treatment Groups are formed by making all possible combinations of the two factors. For example, if the first factor has 3 levels and the second factor has 2 levels, then there will be 3x2=6 different treatment groups.
Main Effect	The main effect involves the independent variables one at a time. The interaction is

	ignored for this part. Just the rows or just the columns are used, not mixed. This is the part which is similar to the one-way analysis of variance. Each of the variances calculated to analyze the main effects are like the between variances
Interaction Effect	The interaction effect is the effect that one factor has on the other factor. The degrees of freedom here is the product of the two degrees of freedom for each factor.
Within Variation	The Within variation is the sum of squares within each treatment group. You have one less than the sample size (remember all treatment groups must have the same sample size for a two-way ANOVA) for each treatment group. The total number of treatment groups is the product of the number of levels for each factor. The within variance is the within variation divided by its degrees of freedom. The within group is also called the error.
F-Tests	There is an F-test for each of the hypotheses, and the F-test is the mean square for each main effect and the interaction effect divided by the within variance. The numerator degrees of freedom come from each effect, and the denominator degrees of freedom is the degrees of freedom for the within variance in each case.

ANOVA Table: Two-Way Classification

Source of Variation	SS (Sum of Squares)	V (Degrees of freedom)	MS (Mean Sum of Squares)	Variance Ratio of F
Between Columns	SSC	c-1	MSC	MSC/MSE
Between Rows	SSR	r-1	MSR	MSR/MSE
Residual or error	SSE	(c-1) (r-1)	MSE	
Total	SST	n-1		

Difference between One-way and Two-way ANOVA

Basis for Comparison	One-way ANOVA	Two-way ANOVA
Meaning	It is a hypothesis test, used to test the equality of three of more population means simultaneously using variance.	It is a statistical technique wherein, the interaction between factors, influencing variable can be studied.
Independent Variable	One	Two
Compares	Three or more levels of one factor.	Effect of multiple levels of two factors.
Number of Observation	Need not to be same in each group.	Need to be equal in each group.
Design of experiments	Need to satisfy only two principles.	All three principles needs to be satisfied.

CHI-SQUARE DISTRIBUTION

A Chi-square distribution is the distribution of the sum of squares of k independent standard normal random variable with k degree of freedom. It is a statistical hypothesis where the null hypothesis that the distribution of the test statistic is a chi-square distribution, is true. While it was first introduced by German Statistician Robert Helmert, and was used by Karl person in 1900. The most popular chi-square test is Pearson's chi-square test and is also called 'chi-squared' test and denoted by χ^2. A classical example of chi-square test is the test for fairness of a die where we test the hypothesis that all six possible outcomes are equally likely.

Definitions

Chi-square distribution	A distribution obtained from the multiplying the ratio of sample variance to population variance by the degrees of freedom when random samples are selected from a normally distributed population.
Contingency Table	Data arranged in table form for the chi-square independence test
Expected Frequency	The frequencies obtained by calculation.
Goodness-of-fit Test	A test to see if a sample comes from a population with the given distribution.
Independence Test	A test to see if the row and column variables are independent.
Observed Frequency	The frequencies obtained by observation. These are the sample frequencies.

Properties of the Chi-Square

- Chi-square is non-negative. Is the ratio of two non-negative values, therefore must be non-negative itself.
- Chi-square is non-symmetric.
- There are many different chi-square distributions, one for each degree of freedom.
- The degrees of freedom when working with a single population variance is n-1.

Uses

- 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).

Chi-square tests deals with three types of tests. They are:

1. Tests of hypothesis about contingency tables, called independence and homogeneity tests.
2. Tests of hypothesis for experiments with more than two categories, called Goodness-of-fit tests.
3. Tests of hypothesis about the variance and standard deviation of a single population.

Test for Independence

In the test for independence, the claim is that the row and column variables are independent of each other. This is the null hypothesis. The multiplication rule said that if two events were independent, then the probability of both occurring was the product of the probabilities of each occurring. This is key to working the test for independence. If you end up rejecting the null hypothesis, then the assumption must have been wrong and the row and column variable are dependent. Remember, all hypothesis testing is done under the assumption the null hypothesis is true. The test statistic used is the same as the chi-square goodness-of-fit test. The principle behind the test for independence is the same as the principle behind the goodness-of-fit test. The **test for independence is always a right tail test**. In fact, you can think of the test for independence as a goodness-of-fit test where the data is arranged into table form. This table is called a contingency table.

The test statistic has a chi-square distribution when the following assumptions are met

- The data are obtained from a random sample
- The expected frequency of each category must be at least 5.

The following are properties of the test for independence

- The data are the observed frequencies.
- The data is arranged into a contingency table.
- The degrees of freedom are the degrees of freedom for the row variable times the degrees of freedom for the column variable. It is not one less than the sample size, it is the product of the two degrees of freedom.
- It is always a right tail test.
- It has a chi-square distribution.
- The expected value is computed by taking the row total times the column total and dividing by the grand total
- The value of the test statistic doesn't change if the order of the rows or columns are switched.
- The value of the test statistic doesn't change if the rows and columns are interchanged (transpose of the matrix)

Test for Single Population Variance

The variable $3^2 = \frac{df.s^2}{\sigma^2}$ has a chi-square distribution if the population variance has a normal distribution. The degrees of freedom are n-1. We can use this to test the population variance under certain conditions

Conditions for testing

- The population has a normal distribution
- The data is from a random sample
- The observations must be independent of each other
- The test statistic has a chi-square distribution with n-1 degrees of freedom

Test for Goodness-of-fit:

The idea behind the chi-square goodness-of-fit test is to see if the sample comes from the population with the claimed distribution. Another way of looking at that is to ask if the frequency distribution fits a specific pattern. Two values are involved, an observed value, which is the frequency of a category from a sample, and the expected frequency, which is calculated based upon the claimed distribution. The idea is that if the observed frequency is really close to the claimed (expected) frequency, then the square of the deviations will be small. The square of the deviation is divided by

the expected frequency to weight frequencies. A difference of 10 may be very significant if 12 was the expected frequency, but a difference of 10 isn't very significant at all if the expected frequency was 1200. If the sum of these weighted squared deviations is small, the observed frequencies are close to the expected frequencies and there would be no reason to reject the claim that it came from that distribution. Only when the sum is large is a reason to question the distribution. Therefore, the **chi-square goodness-of-fit test is always a right tail test**.

The test statistic has a chi-square distribution when the following assumptions are met

- The data are obtained from a random sample.
- The expected frequency of each category must be at least 5. This goes back to the requirement that the data be normally distributed. You're simulating a multinomial experiment (using a discrete distribution) with the goodness-of-fit test (and a continuous distribution), and if each expected frequency is at least five then you can use the normal distribution to approximate (much like the binomial). If the expected

The following are properties of the goodness-of-fit test

- The data are the observed frequencies. This means that there is only one data value for each category.
- The degree of freedom is one less than the number of categories, not one less than the sample size.
- It is always a right tail test.
- It has a chi-square distribution.
- The value of the test statistic doesn't change if the order.