



**ALAGAPPA UNIVERSITY**  
(Accredited with 'A+' Grade by NAAC (with CGPA: 3.64) in the Third Cycle and Graded  
As category - I University by MHRD-UGC)  
(A State University Established by the Government of Tamilnadu)



**KARAIKUDI – 630 003**

**DIRECTORATE OF DISTANCE EDUCATION**

**B.Sc. (PSYCHOLOGY)**

**Second Year – Third Semester**

**11933 –PSYCHOLOGICAL STATISTICS**

Copy Right Reserved

For Private Use only

**Author:**

**Dr. M. Ramya Maheswari**

Professor & Head

Department of Psychology,

Ethiraj College for women,

Chennai-8.

“The Copyright shall be vested with Alagappa University”

All rights reserved. No part of this publication which is material protected by this copyright notice may be reproduced or transmitted or utilized or stored in any form or by any means now known or hereinafter invented, electronic, digital or mechanical, including photocopying, scanning, recording or by any information storage or retrieval system, without prior written permission from the Alagappa University, Karaikudi, Tamil Nadu.

## SYLLABI BOOK MAPPING TABLE

### PSYCHOLOGICAL STATISTICS

Syllabi	Mapping in Book
<b>BLOCK I: STATISTICAL MEASURES</b>	<b>1-13</b>
<b>UNIT 1: Statistics</b> Definitions of statistics Importance of Statics in Psychological and Research Collections of Data Primary and Secondary-classification of Data Formation of Frequency distribution	
<b>UNIT 2: Diagrammatic and Graphical</b> Diagrammatic and Graphical representation of Statistical Data Simple Multiple sub divided, percentage Bar Diagram and Pie diagram – Histogram Frequency polygon Frequency Curve, Ogives	<b>14-24</b>
<b>UNIT 3: Measures of Central Tendency</b> Measures of Central Tendency Mean Median Mode Geometric Mean and Harmonic Mean	<b>25-34</b>
<b>BLOCK II: CONCEPTS OF DISPERSION, CORRELATION AND PROBABILITY</b>	<b>35-46</b>
<b>UNIT 4 : MEASURE OF DISPERSION</b> Measure of Dispersion range Quartile Deviation, Mean Deviation (about mean) Standard Deviation and Co efficient of variance –concept of Skeweness –Karl Pearson and Bowley’s coefficient of Skeweness – Simple Problems- Kurtosis 9concpet only	
<b>UNIT 5: CORRELATION</b> Correlation-Scatter diagram – Karl Pearson co efficient correlation Spearmen’s Rank correlation – regression equation –Properties of regressive coefficient simple problems	<b>47-60</b>
<b>UNIT 6: CONCEPT OF PROBABILITY</b> Concept of probability basic Definition- mathematical and statistical approach – Addition and multiplication Theorems(without proof) simple Problems	<b>61-69</b>
<b>BLOCK III : LEARNING, INTELLIGENCE AND PERSONALITY</b>	<b>70-82</b>
<b>UNIT 7 : NORMAL DISTRIBUTION</b> Concept of Probability, Meaning, Importance and Characteristics of Normal Probability Curve, Deviations from Normal Probability Curve	

<b>UNIT 8: BINOMIAL DISTRIBUTIONS</b> Skewness and Kurtosis, Applications of Normal Probability Curve, Binomial Distributions.	<b>83-93</b>
<b>UNIT 9: ANALYSIS OF VARIANCE</b>  Purpose and Assumptions of Analysis of Variance, One way and Two way Analysis of Variance	<b>94-110</b>
<b>UNIT 10: NON-PARAMETRIC TESTS</b> Nature and Assumptions, Distribution Free Statistics, Chi-square, Contingency Co- efficient, Median and Sign test, Friedman test	<b>111-121</b>
<b>BLOCK IV: EDUCATION FOR ALL</b>	<b>122-126</b>
<b>UNIT 11: Preparation of Data for Computer Analysis</b>  Familiarization with Software Packages for Statistical Analysis and their Applications.	
<b>UNIT 12: INFERENCE STATISTICS</b>  Introduction to Inferential Statistics and Hypothesis Testing about the Difference between Two Independent Means: The meaning of Statistical Inference and Hypothesis Testing; Hypothesis Testing about the difference between Two Independent means; Null and the Alternative Hypotheses;	<b>127-135</b>
<b>UNIT 13: THE RANDOM SAMPLING DISTRIBUTION</b>  The Random Sampling Distribution of the Difference between Two Sample Means; Properties of the Sampling Distribution of the Difference between Means; Choice of HA: One-Tailed and Two- Tailed Tests;	<b>136-151</b>
<b>UNIT 14: INTERPRETING THE RESULTS OF HYPOTHESIS TESTING</b>  A Statistically Significant Difference versus a Practically Important Difference; Errors in Hypothesis Testing; Power of a Test; Levels of Significance versus p-Values.	<b>152-158</b>
<b>Model Question Paper</b>	<b>159-160</b>

# CONTENTS

## UNIT 1: STATISTICS

1-13

- 1.1. Definitions of statistics:
- 1.2. Importance of statistics in psychology and research.
  - 1.2.1: Applications of Statistics
- 1.3. Branches of Statistical methods:
- 1.4. Variables and Measurement
  - 1.4.1: Types of Variables
  - 1.4.2: Scales of measurement.
    - 1.4.2.1: Nominal
    - 1.4.2.2: Ordinal
    - 1.4.2.3: Interval
    - 1.4.2.4: Ratio
- 1.5. Collection of data:
  - 1.5.1 Primary
    - 1.5.1.1 Observation method
    - 1.5.1.2 Interview method
    - 1.5.1.3 Questionnaires
    - 1.5.1.4 Schedules,
  - 1.5.2 Secondary
- 1.6. Classification of Data: Formation of Frequency distribution.
- 1.7. Let us Sum up
- 1.8. Unit – End exercises
- 1.9. Answer to check your progress
- 1.10. Suggested readings

## UNIT 2: DIAGRAMMATIC AND GRAPHICAL REPRESENTATION OF STATISTICAL DATA

14-24

### 2.1 Introduction

- 2.1.1 General principles of graphical representation of data
- 2.1.2 Advantages of graphical representation of data
- 2.1.3 Disadvantages of graphical representation of data
- 2.2 Graphs For Interval and Ratio Data
  - 1.2.1: Histogram
  - 1.2.2: Frequency Polygons
  - 1.2.3: Cumulative Percentage Curve or Ogive.
- 2.3 Graphs for Nominal and Ordinal Data
  - 2.3.1: Bar Diagram
    - 2.3.1.1: Simple bar diagram
    - 2.3.1.2: Multiple bar diagram
    - 2.3.1.3: Sub-divided bar diagram
    - 2.3.1.4: Percentage bar diagram
  - 2.3.2: Pie Charts
- 2.4 Let us Sum up
- 2.5 Unit – End exercises
- 2.6 Answer to check your progress
- 2.7 Suggested readings

## **UNIT 3: MEASURES OF CENTRAL TENDENCY**

**25-34**

- 3.1 Introduction
- 3.2 Arithmetic Mean
  - 3.2.1: Computation of mean from ungrouped data
  - 3.2.2: Computation of mean from grouped data
  - 3.2.3 : Properties of the mean
- 3.3 Median
  - 3.3.1: Computation of median from ungrouped data
  - 3.3.2: Computation of median from grouped data
  - 3.3.3 : Properties of the median
- 3.4 Mode
  - 3.4.1: Computation of mode from ungrouped data
  - 3.4.2: Computation of mode from grouped data
  - 3.4.3 Properties of the mode
- 3.5 Geometric Mean
- 3.6 Harmonic Mean.
- 3.7 Let us Sum up
- 3.8 Unit – End exercises
- 3.9 Answer to check your progress
- 3.10 Suggested readings

## **BLOCK II: CONCEPTS OF DISPERSION, CORRELATION AND PROBABILITY**

**35-46**

### **UNIT 4: MEASURES OF DISPERSION**

- 1.1 Introduction
- 1.2 Range
  - 1.2.1 Properties of the range
- 1.3 Quartile deviation
  - 4.3.1 Properties of the quartile deviation
- 1.4 Standard deviation and Variance for populations
  - 1.4.1 Computation of standard deviation from ungrouped and grouped data
  - 1.4.2 Properties of the Standard deviation
  - 1.4.3 Computation of variance
- 4.5 Standard deviation and Variance for samples
  - 4.5.4: Sampling variability and degrees of freedom
- 4.6 Concept of skewness and kurtosis
  - 4.6.1 Karl Pearson co-efficient of skewness
  - 4.6.2 Bowley's Co-efficient of skewness
  - 4.6.3 Kurtosis
- 4.7 Let us Sum up
- 4.8 Unit – End exercises
- 4.9 Answer to check your progress
- 4.10 Suggested readings

## **UNIT V: CORRELATION**

**47-60**

- 5.1:** Introduction to Correlation
  - 5.1.1: Characteristics of relationship
- 5.2 Scatter Diagram
- 5.3** Methods of computing co-efficient of correlation
  - 5.3.1 Karl Pearson co-efficient of correlation
  - 5.3.2 Spearman's rank correlation
- 5.4** Regression
  - 5.4.1: Regression Line
  - 5.4.2: Regression Equation
  - 5.4.3 Properties of regression co-efficient.
- 1.4 Let us Sum up
- 1.5 Unit – End exercises
- 1.6 Answer to check your progress
- 5.8 Suggested Readings

## **UNIT 6: CONCEPT OF PROBABILITY**

**61-69**

- 6.1 Introduction
- 6.2 Approaches to Probability
- 6.3 Random Sampling
- 6.4 Basic Rules of Probability
  - 6.4.1 Addition theorem
  - 6.4.2 Multiplication theorem for dependent and independent events
- 6.5 Let us Sum up
- 6.6 Unit – End exercises
- 6.7 Answer to check your progress
- 6.8 Suggested Readings

## **BLOCK III: DISTRIBUTIONS AND VARIANCE**

**70-82**

### **UNIT 7: NORMAL DISTRIBUTION**

- 7.1 Meaning and Importance
- 7.2 Characteristics of Normal Probability Curve
  - 7.2.1 Probability and Normal Distribution
- 7.3 Deviations of Normal Probability Curve
  - 7.3.1 Skewness
  - 7.3.2 Kurtosis
- 7.4 Applications of the Normal Probability Curve.
- 7.5 Let us Sum up
- 7.6 Unit – End exercises
- 7.7 Answer to check your progress
- 7.8 Suggested Readings

### **UNIT 8: BINOMIAL DISTRIBUTIONS**

**83-93**

- 8.1 Introduction
- 8.2 The Binomial Distribution

- 8.3 The Binomial Test
  - 8.3.1 Hypotheses for the binomial test
  - 8.3.2 The data for the binomial test
  - 8.3.3 The test statistic for the binomial test
- 8.4 Let us sum up
- 8.5 Unit End Exercises
- 8.6 Answers to check your progress
- 8.7 Suggested Readings

## **UNIT 9: ANALYSIS OF VARIANCE**

**94-110**

- 9.1 Purpose and Logic of Analysis of Variance
- 9.2 Assumptions of Analysis of Variance.
- 9.3 One Way Analysis of Variance.
  - 9.3.1 Illustration
- 9.4 Two Way Analysis of Variance
  - 9.4.1 Illustration
- 9.5 Let us Sum up
- 9.6 Unit – End exercises
- 9.7 Answer to check your progress
- 9.8 Suggested Readings

## **10 UNIT 10: NON- PARAMETRIC TESTS**

**111-121**

- 10.1. Nature and Assumptions
  - 10.1.1 Advantages of Parametric test
  - 10.1.1 Disadvantages of Parametric test
- 10.2 Chi-square
  - 10.2.1 Assumptions of the Chi-square test
  - 10.2.2 Null hypothesis for chisquare
  - 10.2.3 Steps to compute Chi-square.
  - 10.2.4 Uses of chi-square test
- 10.3 Contingency Co-efficient
- 10.4 Median test
  - 10.4.1 Steps in the computation of the median test
- 10.5 Sign test
- 10.6 Friedman test.
- 10.7 Let's sum up
- 10.8 Unit End Exercises
- 10.9 Answers to check your progress
- 10.10 Suggested Readings

## **BLOCK IV: ANALYSIS AND INTERPRETATION OF DATA**

### **UNIT 11: PREPARATION OF DATA FOR COMPUTER**

**122-126**

- 11.1 Introduction:
- 11.2 Software packages that are used for statistical analysis
  - 11.2.1. Programs for Quantitative Research.
  - 11.2.2 Programs for Qualitative Research
- 11.3 Let's sum up



- 11.4 Unit End Exercises
- 11.5 Answers to check your progress
- 11.6 Suggested Readings

## **UNIT 12: INFERENCE STATISTICS**

**127-135**

- 12.1 Introduction to Inferential Statistics
- 12.2 Types of Hypothesis
- 12.3 The Meaning of Statistical Inference
  - 12.3.1 Relationship between probability and Inferential Statistics
- 12.4 Methods of statistical Inference
  - 12.4.1 Statistical Estimation
  - 12.4.2 Statistical hypothesis testing
- 12.5 Let us sum up
- 12.6 Unit end Exercises
- 12.7 Answers to check your progress
- 12.8 Suggested Readings

## **UNIT 13: THE RANDOM SAMPLING DISTRIBUTION**

**136-151**

- 13.1 Population and Samples
- 13.2 The Distribution of sample means
- 13.3 Properties of the distribution of sample means
  - 13.3.1 The shape of the distribution of sample means
  - 13.3.2 The mean of the distribution of sample means: The expected value of M
  - 13.3.3 The standard error of M
- 13.4 Sampling distribution of differences between means
- 13.5 Introduction to t statistics
  - 13.5.1: Illustration of hypothesis testing with t statistics for Independent measures research design
  - 13.5.2 : Illustration of hypothesis testing with t statistics Repeated measures research design
- 13.6 Choice of alternative hypothesis: One tailed and two tailed tests of significance
- 13.7 Let us sum up
- 13.8 Unit end exercises
- 13.9 Answers to check your progress
- 13.10 Suggested Readings

**UNIT 14: INTERPRETING THE RESULTS OF HYPOTHESIS TESTING** **152-158**

14.1 A statistically significant difference versus practically important difference.

14.2 Errors in Hypothesis testing.

14.3 Power of the test.

14.4 Statistical decision making: Levels of Significance versus p –values

14.5 Let us sum up

14.6 Unit End exercises

14.7 Answers to check your progress

14.8 Suggested Readings

**Model Question Paper**

**159-160**

---

# BLOCK- I

---

Statistics

**NOTES**

---

## UNIT-I – STATISTICS

---

- 1.1. Definitions of statistics:
- 1.2. Importance of statistics in psychology and research.
  - 1.2.1: Applications of Statistics
- 1.3. Branches of Statistical methods:
- 1.4. Variables and Measurement
  - 1.4.1: Types of Variables
  - 1.4.2: Scales of measurement.
    - 1.4.2.1: Nominal
    - 1.4.2.2: Ordinal
    - 1.4.2.3: Interval
    - 1.4.2.4: Ratio
- 1.5. Collection of data:
  - 1.5.1 Primary
    - 1.5.1.1 Observation method
    - 1.5.1.2 Interview method
    - 1.5.1.3 Questionnaires
    - 1.5.1.4 Schedules,
  - 1.5.2 Secondary
- 1.6. Classification of Data: Formation of Frequency distribution.
- 1.7. Let us Sum up
- 1.8. Unit – End exercises
- 1.9. Answer to check your progress
- 1.10. Suggested readings

---

### 1.1 Definitions of statistics:

---

Statistics according to Horace Secrist is ‘Aggregates of facts, affected to marked extent by multiplicity of causes, estimated or enumerated according to reasonable standards of accuracy, collected in a systematic manner, for a pre-determined purpose and placed in relation to each other’. From the above data we understand that individual figures are not statistics.

Statistics refer to numerical facts. It also signifies the method or methods of dealing with numerical facts. Further statistics refers to the summarized figures of numerical facts such as percentages, averages, means, medians, modes etc. This indicates that qualitative data do not qualify as statistics.

**NOTES**

The term **Statistics** refers to a set of mathematical procedures for organising, summarizing and interpreting information.

---

## **1.2 Importance of statistics in psychology and research:**

---

Statistics renders valuable services in collection of evidences or facts. Statistics are used to organise and summarize the information so that the researcher can see what happened in the research study and can communicate the results to others. Statistics help the researcher to answer the questions that initiated the research by determining exactly what general conclusions are justified based on the specific results that were obtained.

Statistical procedures help to ensure that the information and observation are presented and presented and interpreted in an accurate and informative way. In addition statistics provide researchers with the set of standardised techniques that are recognised and understood throughout the scientific community. Thus, the statistical methods used by one researcher are familiar to other researchers, who can accurately interpret the statistical analyses with the full understanding of how analysis was done and what the results signify.

### **1.2.1: Applications of statistics**

#### **(1) Business**

Statistics plays an important role in business. Statistics helps businessmen to plan production according to the taste of the customers, and the quality of the products can also be checked more efficiently by using statistical methods. Thus, it can be seen that all business activities are based on statistical information. Businessmen can make correct decisions about the location of business, marketing of the products, financial resources, etc.

#### **(2) Economics**

Economics largely depends upon statistics. National income accounts are multipurpose indicators for economists and administrators, and statistical methods are used to prepare these accounts. In economics research, statistical methods are used to collect and analyze the data and test hypotheses. The relationship between supply and demand is studied by statistical methods; imports and exports, inflation rates, and per capita income are problems which require a good knowledge of statistics.

#### **(3) Mathematics**

Statistics plays a central role in almost all natural and social sciences. The methods used in natural sciences are the most reliable but conclusions drawn from them are only probable because they are based on incomplete

evidence. Statistics helps in describing these measurements more precisely. Statistics is a branch of applied mathematics. A large number of statistical methods like probability averages, dispersions, estimation, etc., is used in mathematics, and different techniques of pure mathematics like integration, differentiation and algebra are used in statistics.

#### **(4) Banking**

Statistics plays an important role in banking. Banks make use of statistics for a number of purposes. They work on the principle that everyone who deposits their money with the banks does not withdraw it at the same time. The bank earns profits out of these deposits by lending it to others on interest. Bankers use statistical approaches based on probability to estimate the number of deposits and their claims for a certain day.

#### **(5) State Management (Administration)**

Statistics is essential to a country. Different governmental policies are based on statistics. Statistical data are now widely used in making all administrative decisions. Suppose if the government wants to revise the pay scales of employees in view of an increase in the cost of living, and statistical methods will be used to determine the rise in the cost of living. The preparation of federal and provincial government budgets mainly depends upon statistics because it helps in estimating the expected expenditures and revenue from different sources. So statistics are the eyes of the administration of the state.

#### **(6) Natural and Social Sciences**

Statistics plays a vital role in almost all the natural and social sciences. Statistical methods are commonly used for analyzing experiments results, and testing their significance in biology, physics, chemistry, mathematics, meteorology, research, chambers of commerce, sociology, business, public administration, communications and information technology, etc.

#### **(7) Astronomy**

Astronomy is one of the oldest branches of statistical study; it deals with the measurement of distance, and sizes, masses and densities of heavenly bodies by means of observations. During these measurements errors are unavoidable, so the most probable measurements are found by using statistical methods. For example, the distance of the moon from the earth is measured. Since history, astronomers have been using statistical methods like method of least squares to find the movements of stars.

---

### **1.3 Branches of Statistical methods:**

---

Although researchers have developed a variety of different statistical procedures to organize and interpret data, these different procedures can be classified into two general categories.

- a) Descriptive Statistics: These are statistical procedures used to summarise, organize and simplify data.

**NOTES**

- b) Inferential Statistics: Consists of techniques that allow us to study samples and then make generalizations about the populations from which they are selected.

**Check your Progress -1**

1. Qualitative data are not \_\_\_\_\_.
2. What are the two categories of statistical techniques?

**1.4 Variables and Measurement:**

The scores that make up the data from a research study are the result of observing and measuring variables. The variables in a study can be characterized by the type of values that can be assigned to them

**1.4.1: Types of variables: Continuous and discrete variables**

**a) Continuous Variable:** For a **Continuous Variable**, there are an infinite number of possible values that fall between any two observed values. A continuous variable is divisible into an infinite number of fractional parts. Examples are height, weight etc all of which when measured can be broken down to an infinite number of possible points between any two points on a scale. **Two other factors apply for a continuous variable.**

1) When measuring a continuous variable, it should be very rare to obtain identical measurements for two different individuals, as it has an infinite number of possible values. If the data show a substantial number of tied scores, then you would suspect the measurement procedure is very crude or that the variable is not really continuous.

2) When measuring a continuous variable, each measurement category is actually an interval that must be defined by boundaries. For example, two people who both claim to weigh 150 pounds are probably not exactly the same weight. One person may weigh 149.6 and the other 150.3. Thus, a score of 150 is not a specific point on the scale but instead is an interval. To differentiate a score of 150 from a score of 149, we must set up boundaries on a scale of measurement. These boundaries are called **real limits** and are positioned exactly half way between adjacent scores. Thus, a score of  $X = 150$  pounds is actually an interval bounded by a lower real limit of 149.5 at the bottom and a real upper limit of 150.5 at the top. Any individual whose weight falls between these real limits will be assigned a score of  $X = 150$ .

**b) Discrete Variable:** Discrete variables consist of separate indivisible categories. No values can exist between two neighbouring categories. Discrete variables are commonly restricted to whole countable numbers- for example number of children in the family, number of students

attending class. If you observe class attendance from day to day, you may count 18 students one day and 19 students the next day. However it is impossible ever to observe a value between 18 and 19. A discrete variable may also consist of observations that differ qualitatively. For example people classified by gender (male/female), by occupation (nurse, teacher, etc)

#### **1.4.2 Scales of Measurement:**

Data collection requires that we make measurements of our observation. Measurement involves assigning individuals or events to categories. The categories can simply be names such as male/female, employed/unemployed, or they can be numerical values such as 68 inches or 175 pounds. The categories used to measure a variable make up scales of measurement, and the relationships between categories determine different types of scales. The distinction among the scales are important because they identify the limitations of certain types of measurements and because certain statistical procedures are appropriate for scores that have been measured on some scales but not others.. Stevens (1956) has recognized four type of scales - nominal, ordinal, interval and ratio.

##### **1.4.2.1 Nominal Scale:**

In nominal measurement numbers are used to name, identify a classify persons, objects, groups etc. For example a sample of persons being studied may be classified as (a) Hindu (b) Muslim (c) Sikh or the same sample may be classified on the basis of sex, rural - urban variable etc.

In nominal measurement, members of any two groups are never equivalent but all members of the same group are equivalent. In case of nominal measurement admissible statistical operations are counting a frequency, percentage proportion, mode and co-efficient of contingency. The drawback of nominal measurement is that it is most elementary and simple.

##### **1.4.2.2 Ordinal Scale:**

In ordinal measurement numbers denote the rank order of the objects or the individual. Ordinal measures reflect which person on object is larger or smaller, heavier or lighter, brighter or duller, harder or softer etc.

Persons may be grouped according to physical or psychological traits to convey a relationship like “greater than” or “less than”. Socio-economic status is a good example of ordinal measurement.

In ordinal measurement besides the relationship of equivalence a relationship of greater than or less than exists because all members of any particular subclass are equivalent to each other and at the same time greater or lesser than the members of other subclasses. The permissible statistical operations in ordinal measurement are median, percentiles and rank correlation co-efficient plus all those permissible for nominal measurement. The drawback of ordinal measurement is that ordinal

**NOTES**

measures are not absolute quantities, and do they convey that the distance between the different rank values is equal. This is because ordinal measurement are not equal-interval measurement nor do they incorporate absolute zero point.

**1.4.2.3 Interval or Equal-Interval Scale**

This measurement includes all the characteristics of the nominal and ordinal scales of measurement. The salient feature of interval measurement is that numerically equal distances on the scale indicate equal distance in the properties of the objects being measured. In other words, here the unit of measurement is constant & equal. Since the numbers are after equal intervals, they can legitimately be added and subtracted from each other ie on an interval measurement the intervals or distances (not the quantities or amounts) can be added. The difference or interval between the numbers on the scale reflects differences in magnitude.

Although it is true that the intervals can be added, it does not mean that the process of additivity can be carried out in the absolute sense. The process of additivity of intervals or distances on an interval measurement has only a limited value because in such measurement zero point is not true but rather arbitrary. Zero point, here, does not tell the real absence of the property being measured. The common statistics used in such measurement are arithmetic mean, standard deviation, pearson's r. and other statistics based on them. Statistic like the t test and F test which are widely used tests of significance, can also be legitimately applied. In psychology, sociology and education we frequently encounter interval measurement.

**1.4.2.4 Ratio Scale of Measurement**

It is the highest level of measurement and has all the properties of nominal. Ordinal and interval scales plus an absolute or true zero point. The salient feature of the ratio scale is that the ratio of two numbers is independent of the unit of measurement and therefore, it can be meaningfully equated.

Common examples of ratio scale are the measures of weight, width, length, loudness and so on. It is common among physical sciences than social sciences.

**Check your progress -2**

3. The number of eggs in a basket is an example of \_\_\_\_\_ variable.
4. The different television shows make up a \_\_\_\_\_ scale of measurement.



5. When measuring height to the nearest half inch, what are the real limits for the score of 68 inches?

---

### 1.5: Collection of Data:

---

While deciding about the method of data collection to be used for the study, the researcher should keep in mind two types of data viz., primary and secondary. The primary data are those which are collected afresh and for the first time, and thus happen to be original in character. The secondary data, on the other hand, are those which have already been collected by someone else and which have already been passed through the statistical process. The researcher would have to decide which sort of data he would be using (thus collecting) for his study and accordingly he will have to select one or the other method of data collection.

**1.5.1: Primary Data:** There are several ways in which primary data are collected especially in surveys and descriptive researches. We shall review the four most commonly employed ways of data collection.

**1.5.1.1 Observation method:** The observation method is the most commonly used method specially in studies relating to behavioural sciences. Observation becomes a scientific tool and the method of data collection for the researcher, when it serves a formulated research purpose, is systematically planned and recorded and is subjected to checks and controls on validity and reliability. Under the observation method, the information is sought by way of investigator's own direct observation without asking from the respondent. For instance, in a study relating to consumer behaviour, the investigator instead of asking the brand of wrist watch used by the respondent, may himself look at the watch. The main advantage of this method is that subjective bias is eliminated, if observation is done accurately. Secondly, the information obtained under this method relates to what is currently happening; it is not complicated by either the past behaviour or future intentions or attitudes. Thirdly, this method is independent of respondents' willingness to respond and as such is relatively less demanding of active cooperation on the part of respondents as happens to be the case in the interview or the questionnaire method. This method is particularly suitable in studies which deal with subjects (i.e., respondents) who are not capable of giving verbal reports of their feelings for one reason or the other. However, observation method has various limitations. Firstly, it is an expensive method. Secondly, the information provided by this method is very limited. Thirdly, sometimes unforeseen factors may interfere with the observational task. At times, the fact that some people are rarely accessible to direct observation creates obstacle for this method to collect data effectively.

**1.5.1.2 Interview method:** The interview method of collecting data involves presentation of oral-verbal stimuli and reply in terms of oral-verbal responses. This method can be used through personal interviews and, if

### NOTES

**NOTES**

possible, through telephone interviews. Personal interview method requires a person known as the interviewer asking questions generally in a face-to-face contact to the other person or persons. This method is particularly suitable for intensive investigations.

The method of collecting information through personal interviews is usually carried out in a structured way. Structured interviews involve the use of a set of predetermined questions and of highly standardised techniques of recording. As against it, the unstructured interviews are characterised by a flexibility of approach to questioning. In a non-structured interview, the interviewer is allowed much greater freedom to ask, in case of need, supplementary questions or at times he may omit certain questions if the situation so requires. But this sort of flexibility results in lack of comparability of one interview with another and the analysis of unstructured responses becomes much more difficult and time-consuming than that of the structured responses obtained in case of structured interviews. Unstructured interviews also demand deep knowledge and greater skill on the part of the interviewer. Unstructured interview, however, happens to be the central technique of collecting information in case of exploratory or formulative research studies. But in case of descriptive studies, we quite often use the technique of structured interview because of its being more economical, providing a safe basis for generalisation and requiring relatively lesser skill on the part of the interviewer.

**1.5.1.3 Questionnaires:** This method of data collection is quite popular, particularly in case of big enquiries. It is being adopted by private individuals, research workers, private and public organisations and even by governments. In this method a questionnaire is sent to the persons concerned with a request to answer the questions and return the questionnaire. A questionnaire consists of a number of questions printed or typed in a definite order on a form or set of forms. The questionnaire is mailed/or distributed individually to respondents who are expected to read and understand the questions and write down the reply in the space meant for the purpose in the questionnaire itself. The respondents have to answer the questions on their own. The method of collecting data by mailing the questionnaires to respondents is most extensively employed in various economic and business surveys. The following are the **merits** of the questionnaire method

1. There is low cost even when the universe is large and is widely spread geographically
2. It is free from the bias of the interviewer; answers are in respondents' own words.
3. Respondents have adequate time to give well thought out answers.

4. Respondents, who are not easily approachable, can also be reached conveniently.
5. Large samples can be made use of and thus the results can be made more dependable and reliable.

The main **demerits** of this system are listed below

1. Low rate of return of the duly filled in questionnaires; bias due to no-response is often indeterminate.
2. It can be used only when respondents are educated and cooperating.
3. The control over questionnaire may be lost once it is sent.
4. There is inbuilt inflexibility because of the difficulty of amending the approach once questionnaires have been despatched.
5. There is also the possibility of ambiguous replies or omission of replies altogether to certain questions; interpretation of omissions is difficult.
6. It is difficult to know whether willing respondents are truly representative.
7. This method is likely to be the slowest of all.

Before using this method, it is always advisable to conduct 'pilot study' for testing the questionnaires. In a big enquiry the significance of pilot survey is felt very much. Pilot survey is the rehearsal of the main survey. Such a survey, being conducted by experts, brings to the light the weaknesses (if any) of the questionnaires and also of the survey techniques. From the experience gained in this way, improvement can be effected.

**1.5.1.4 Schedules:** This method of data collection is very much like the collection of data through questionnaire, with little difference which lies in the fact that schedules are being filled in by the enumerators who are specially appointed for the purpose. These enumerators along with schedules, go to respondents, put to them the questions from the proforma in the order the questions are listed and record the replies in the space meant for the same in the proforma. Enumerators explain the aims and objects of the investigation and also remove the difficulties which any respondent may feel in understanding the implications of a particular question or the definition or concept of difficult terms. This method requires the selection of enumerators for filling up schedules or assisting respondents to fill up schedules and as such enumerators should be very carefully selected. The enumerators should be trained to perform their job well and the nature and scope of the investigation should be explained to them thoroughly so that they may well understand the implications of different questions put in the schedule. This method of data collection is very useful in extensive enquiries and can lead to fairly reliable results. It is, however, very expensive and is usually adopted in investigations

**NOTES**

conducted by governmental agencies or by some big organisations. Population census all over the world is conducted through this method.

**1.5.2: Secondary data:** Secondary data may either be published data or unpublished data. Usually published data are available in: (a) various publications of the central, state or local governments; (b) various publications of foreign governments or of international bodies and their subsidiary organisations; (c) technical and trade journals; (d) books, magazines and newspapers; (e) reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.; (f) reports prepared by research scholars, universities, economists, etc. in different fields; and (g) public records and statistics, historical documents, and other sources of published information. The sources of unpublished data are many; they may be found in diaries, letters, unpublished biographies and autobiographies and also may be available with scholars and research workers, trade associations, labour bureaus and other public/ private individuals and organisations. Researcher must be very careful in using secondary data. He must make a minute scrutiny because it is just possible that the secondary data may be unsuitable or may be inadequate in the context of the problem which the researcher wants to study

**Check your Progress -3**

6. Population Census all over the world is conducted through \_\_\_\_\_.
7. Why is it essential to do a pilot survey?

---

**1.6 Classification of Data : Formation of frequency distribution:**

---

Tests, experiments and survey studies in education and psychology provide us valuable data, mostly in the shape of numerical stores. For understanding the meaning and deriving useful conclusion the data have to be organized or arranged in systematic manner. One such means of organizing the original data or computed statistics are called frequency distribution.

A frequency table is a systematic listing of the number of scores of each value in the group studied. It makes it easy for the investigator to see a pattern in a large group of scores.

For example a frequency table for a nominal variable with the following score will look like the following.

5, 7, 4, 5, 6, 5, 4

**Frequency Table**

Value	Frequency	Percentage
7	1	14
6	1	14
5	3	43
4	2	29

**NOTES**

Sometimes there are so many possible values that an ordinary frequency table is too awkward to give a simple picture of the scores. The solution is to make groupings of value that include all values within a certain range. This combined category is a range of values that includes many values and is called an interval. A frequency table that uses intervals is called a frequency distribution. This shows all the possible scores and their frequency of occurrence.

**Guidelines for constructing a frequency distribution**

A set of raw scores does not result in a unique set of grouped scores. A given set of score can be grouped in more than one way. Some widely accepted conventions that help us to group scores easily are as follows.

- Be sure that the class intervals are mutually exclusive
- Make all intervals of the same width
- Make the intervals continuous throughout the distribution
- Place the interval containing the highest score value at the top
- For most work, use 10 to 20 class intervals.
- Choose a convenient interval width
- When possible make the lower score limits multiples of the interval width

**Steps to construct a grouped frequency distribution**

- Finding the range** - It is done by subtracting the lowest score from the highest.
- Determining the class interval or grouping interval:** *There are two different means to get an idea of the size of the class interval, the range is divided by the number of classes desired. Class interval is usually denoted by the symbol  $i$ .*

**NOTES**

$$i = \frac{\text{Range}}{\text{No. of classes desired}}$$

To decide the No. of classes desired as a general rule, Tate (1955) writes.

If the series contains fewer than about 50 item, more than about 10 classes are not justified. If the series contain from about 50 to 100 items, 10 to 15 classes tend to be appropriate. If more than 100 items, 15 or more class tend to be appropriate. Ordinarily not fewer than 10 classes or more than 20 are used.

Class interval 'i' can be decided first and then the number of classes desired. A combined procedure that takes into account the range, the number of classes and the class interval while planning for a frequency distribution is must so that we are able to arrive at a frequency distributions that minimizes grouping error. The wider the class interval width the greater the potential for grouping error.

**c. Writing the contents of the frequency distribution**

The classes of the distribution written and the scores given in the data are taken one by one and tallied in their proper classes. The tallies all totaled and frequencies of each class interval is noted down.

Classes of Scores	Tallies	Frequencies
65 - 69	1	1
60 - 64	111	3
55 - 59	1111	4
50 - 54	1111 11	7
45 - 49	1111 1111	9
40 - 44	1111 1111	11
35 - 39	1111 111	8
30 - 34	1111	4
25 - 29	11	2
20 - 24	1	1
----		
	Total frequencies No.	50
		-----

---

### 1.7 Lets Sum up:

---

In this unit, you have been introduced to basic statistical concepts. The need and importance of statistics and its applications in the diverse field has been presented. This knowledge will equip you to formulate research questions in various fields of study. The nature of variables and the scales of measurement will help you to operationally conceptualise the variable that you choose for your research study, The section on methods of data collection will provide insight on how to go about collecting data for the chosen research study.

---

### 1.8 Unit – End exercises:

---

1. List the difference between schedules and questionnaires
2. Give examples of variables that can be measured using the various scales of measurement.
3. List applications of statistics in day to day life.

---

### 1.9 Answer to check your progress

---

1. Qualitative data are not Statistics.
2. The two categories of statistical techniques are Descriptive and inferential.
3. The number of eggs in a basket is an example of discrete variable.
4. The different television shows make up a nominal scale of measurement.
5. 67.5 - 68.5 inches.
6. Population Census all over the world are conducted through schedules
7. Pilot study, is a small scale preliminary study conducted in order to evaluate feasibility, time, cost, adverse events, and improve upon the study design prior to performance of a full-scale research project.

---

### 1.10 Suggested readings

---

1. S.P. Gupta, Statistical Methods, Sultan Chand and Sons, New Delhi, 2002.
2. Henry E., Garret's Statistics in Psychology and Education, Feffer and Simans Private Limited, 1969, p.329.

---

## **UNIT 2: DIAGRAMMATIC AND GRAPHICAL**

---

### **2.1 Introduction**

**2.1.1 General principles of graphical representation of data**

**2.1.2 Advantages of graphical representation of data**

**2.1.3 Disadvantages of graphical representation of data**

### **2.2 Graphs For Interval and Ratio Data**

**1.2.1: Histogram**

**1.2.2: Frequency Polygons**

**1.2.3: Cumulative Percentage Curve or Ogive.**

### **2.3 Graphs for Nominal and Ordinal Data**

**2.3.1: Bar Diagram**

**2.3.1:1: Simple bar diagram**

**2.3.1:2: Multiple bar diagram**

**2.3.1:3: Sub-divided bar diagram**

**2.3.1: 4: Percentage bar diagram**

**2.3.2: Pie Charts**

**2.4 Let us Sum up**

**2.5 Unit – End exercises**

**2.6 Answer to check your progress**

**2.7 Suggested readings**

---

### **2.1 Introduction**

---

A graph is another good way to make a large group of scores easy to understand “A picture is worth a thousand words” - and sometimes a thousand numbers. A graph is a sort of chart through which *statistical data* are represented in the form of lines or curves drawn across the coordinated points plotted. In other words, it is a geometrical image of a set of data.

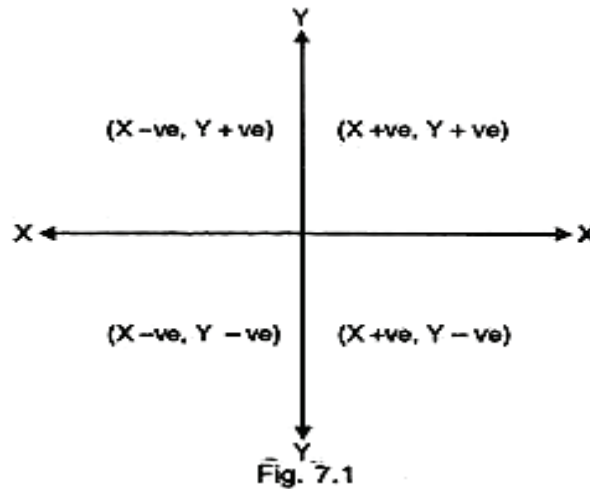
#### **General Principles of Graphic Representation:**

There are some algebraic principles which apply to all types of graphic representation of data. In a graph there are two lines called coordinate axes. One is vertical known as Y axis and the other is horizontal called X axis. These two lines are perpendicular to each other. Where these two lines intersect each other is called ‘0’ or the Origin. On the X axis the distances



right to the origin have positive value (see fig. 7.1) and distances left to the origin have negative value. On the Y axis distances above the origin have a positive value and below the origin have a negative value.

**NOTES**



**2.1.1 Advantages of graphical representation of data**

1. **Acceptability:** Such report is acceptable to the busy people because it easily highlights the theme of the report. This helps to avoid wastage of time.
2. **Comparative Analysis:** Information can be compared easily. Such comparative analysis helps for quick understanding and attention.
3. **Decision Making:** Business executives can view the graphs at a glance and can make decision very quickly which is hardly possible through descriptive report.
4. **Helpful for less literate Audience:** Less literate or illiterate people can understand graphical representation easily because it does not involve going through line by line of any descriptive report.
5. **A complete Idea:** Such representation creates clear and complete idea in the mind of audience. Reading hundred pages may not give any scope to make decision. But an instant view or looking at a glance obviously makes an impression.
6. **Use in the Notice Board:** Such representation can be hanged in the notice board to quickly raise the attention of employees in any organization.

**2.1.2 Disadvantages of Graphical Representation of Data**

1. **Lack of Secrecy:** Graphical representation makes full presentation of information which may hamper the objective to keep something secret.

**NOTES**

2. **More time:** Normal report involves less time to represent but graphical representation involves more time as it requires graphs and figures .
3. **Errors and Mistakes:** Since graphical representations are complex, there are chance of errors and mistake. This causes problems for better understanding to lay people.

**Check your progress- 1**

1. What is the chief need for representing data graphically?
2. The point of intersection between two co-ordinate axis in a graph is known as \_\_\_\_\_

---

**2.2 Graphs For Interval and Ratio Data:**

---

When the data consist of numerical scores that have been measured on an interval or ratio scale, histograms , polygons and frequency curves can be constructed

**2.2.1: Histogram :** To construct a histogram

- i. Make a frequency distribution
- ii. Put the values of real lower limits of the class interval along the bottom of the page.
- iii. Make a scale of frequencies along the Y axis that goes from 0 at the bottom to the highest frequency of any value.
- iv. Make a bar above each value with a height for the frequency of that class interval. Each class or interval with its specific frequency is represented by a separate triangle. The base of each rectangle is the width of the class interval (i) and the height is the respective frequency of that class or interval.

**NOTES**

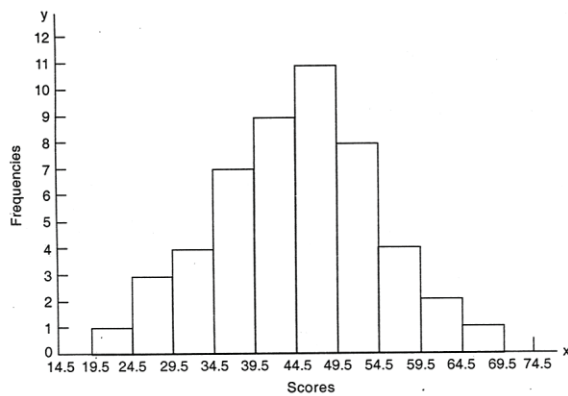


Figure 3.6 Histogram of frequency distribution.

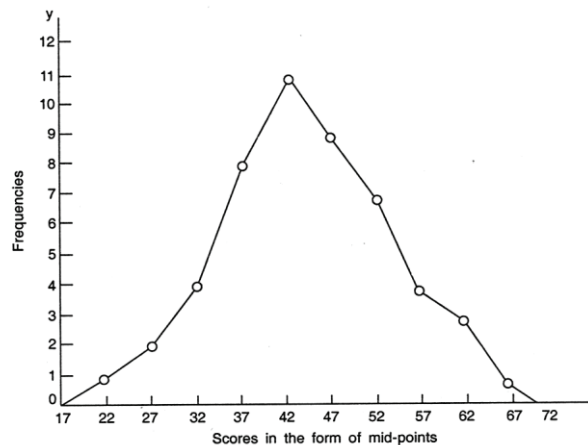
### 2.2.2 Frequency Polygons

Another way to graph a frequency table is to make a special kind of limit graph called a frequency polygon. A frequency polygon is essentially a line graph for graphical representation of the frequency distribution. We can get a frequency polygon from a histogram, if the midpoints of the upper bases of the rectangles are connected by straight lines. But it is not essential to plot a histogram first to draw a frequency polygon.

There are five steps for making a frequency polygon.

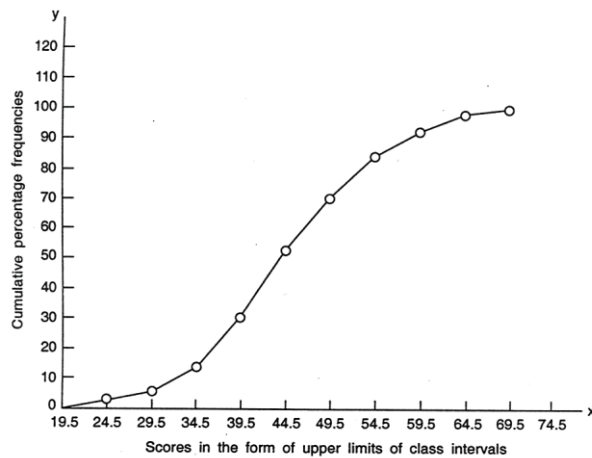
- a) Make the frequency table
- b) Put the midpoints of the class intervals along the bottom of the page, from left to right, starting one value below the lowest value and ending one value above the highest value. The extra values helps to close the figure.
- c) Make a scale of frequencies along the left edge of the page that goes from zero at the bottom to the highest frequency for any value.
- d) Make a point above each value with the height for the frequency of that value.
- e) Connect the point with lines.

**NOTES**



**2.2.3 Cumulative percentage curve/ Ogive:** When a population consists of numerical scores from an interval or a ratio scale, it is customary to draw the distribution with a smooth curve instead of the jagged, step-wise shapes that occur with histograms and polygons. The smooth curve indicates that you are not connecting a series of dots (real frequencies) but instead are showing the relative changes that occur from one score to the next. The first step in the construction of such graph is organising the data in the form of cumulative frequency distribution, which is then converted into cumulative percentages. Technically, a cumulative frequency distribution is the sum of the class and all classes below it in a frequency distribution. All that **means** is you're adding up a value and all of the values that came before it.

Cumulative percentage indicates the percentage of scores that lies below the upper real limit of the associated class interval. Therefore, in constructing cumulative percentage curves, the cumulative percentage is plotted at the upper real limit of the class interval. We then connect the points with straight lines and bring the curve down to zero at the lower end at the upper real limit of the next adjacent class interval. A cumulative percentage curve never has a negative slope. The result in a cumulative percentage curve with an S-shaped figure is called an Ogive. We may determine percentiles or are percentile ranks from the cumulative percentage curve.



### Check your progress -2

3. What is plotted on the x –axis when you construct a frequency polygon?
4. What is the advantage ogive over histogram and frequency polygon?
5. What do you understand by the term cumulative frequency?

---

### 2.3 Graphs for Nominal and Ordinal Data:

---

Bar graphs and Pie-charts are widely used for this purpose

**2.3.1: Bar Diagram:** The bar diagram is very similar to the histogram and is constructed in the same manner except that space appears between the rectangles, thus suggesting the essential discontinuity of several categories. However within categories, sub categories may be displayed as adjacent bars. Because qualitative categories on a nominal scale of measurement have no necessary order, we may arrange them in any order. However, for ordinal scale of measurement, the categories should be arranged in order of rank.

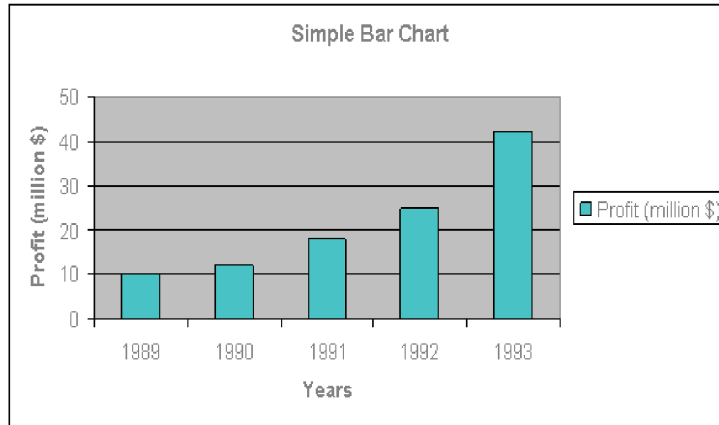
**2.3.1:1: Simple bar diagram:** A **simple bar chart** is used to represent data involving only one variable classified on a spatial, quantitative or temporal basis. In a **simple bar chart**, we make bars of equal width but variable length, i.e. the magnitude of a quantity is represented by the height or length of the bars.

The following steps are used to draw a simple bar diagram:

- Draw two perpendicular lines, one horizontally and the other vertically, at an appropriate place on the paper.

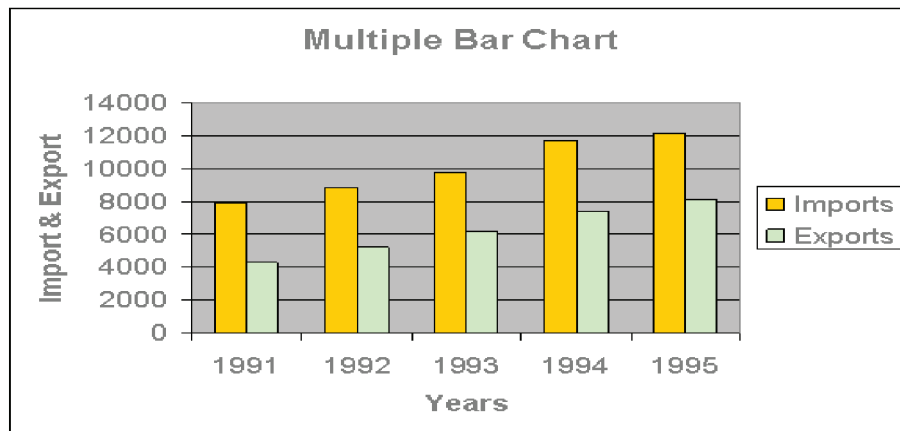
**NOTES**

- Take the basis of classification along the horizontal line (X-X- axis) and the observed variable along the vertical line (Y-Y- axis), or vice versa.
- Mark signs of equal breadth for each class and leave equal or not less than half a breadth between two classes.
- Finally mark the values of the given variable to prepare required bars.



The above is an illustration of a simple bar chart showing the profit of a particular bank over a five year period

**2.3.1:2: Multiple bar diagram:** In a multiple bars diagram two or more sets of inter-related data are represented (multiple bar diagram facilitates comparison between more than one phenomena). The technique of making a simple bar chart is used to draw this diagram but the difference is that we use different shades, colours, or dots to distinguish between different phenomena.



The above graph is an illustration of the multiple bar diagram showing import and export trade in a country over a 5 year time period.

**NOTES**

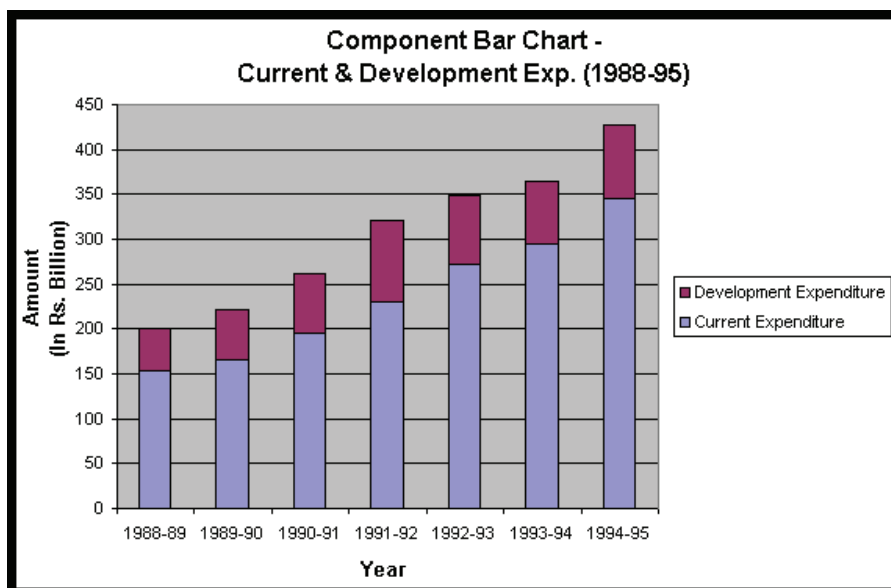
**2.3.1:3: Sub-divided bar diagram:** In this diagram, first we make simple bars for each class taking the total magnitude in that class and then divide these simple bars into parts in the ratio of various components.

1. This chart consists of bars which are sub-divided into two or more parts.
2. The length of the bars is proportional to the totals.
3. The component bars are shaded or coloured differently.

Below is a numerical table and graphical representation of the numerical facts in a sub-divided bar diagram

**Current and Development Expenditure – Pakistan (All figures in Rs. Billion)**

Years	Current Expenditure	Development Expenditure	Total Expenditure
1988-89	153	48	201
1989-90	166	56	222
1990-91	196	65	261
1991-92	230	91	321
1992-93	272	76	348
1993-94	294	71	365
1994-95	346	82	428

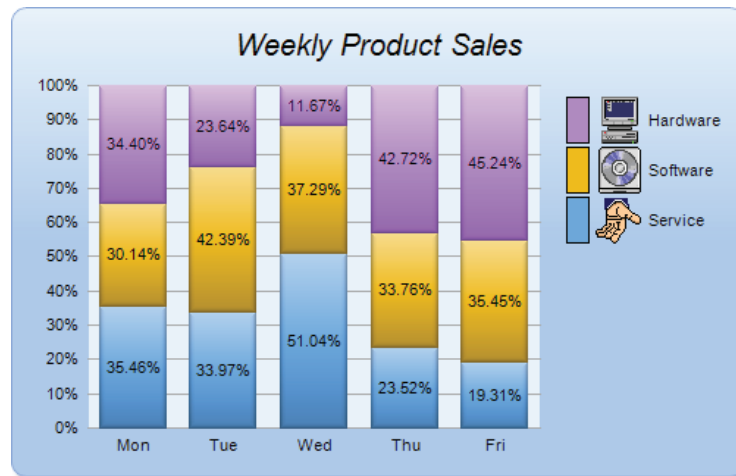


**2.3.1: 4: Percentage bar diagram:**

A sub-divided bar chart may be drawn on a percentage basis. To draw a sub-divided bar chart on a percentage basis, we express each component as the percentage of its respective total. In drawing a

**NOTES**

percentage bar chart, bars of length equal to 100 for each class are drawn in the first step and sub-divided into the proportion of the percentage of their component in the second step. The diagram so obtained is called a percentage component bar chart or percentage stacked bar chart. This type of chart is useful to make comparisons in components holding the difference of total constants.



**2.3.2: Pie Charts:** This data is presented in the form of a circle. There are segments and sectors into which a pie chart is being divided and each of these segments and sectors forms a certain portion of the total (in terms of percentage). In the pie-chart, the total of all the data is equal to 360 degrees. The degree of angles that are used to represent different items are calculated in the form of proportionality. In other words, the total frequencies or value is equated to 360 degrees, and then the angles corresponding to component parts are calculated (or the component parts are expressed as percentages of the total and then multiplied by 3.6). After determining these angles, the required sectors in the circle are drawn. In pie charts, the entire diagram looks like a pie and the components in it resembles the various slices cut from the pie. The pie-chart is thus used to show the break-up of one continuous variable into its components parts.

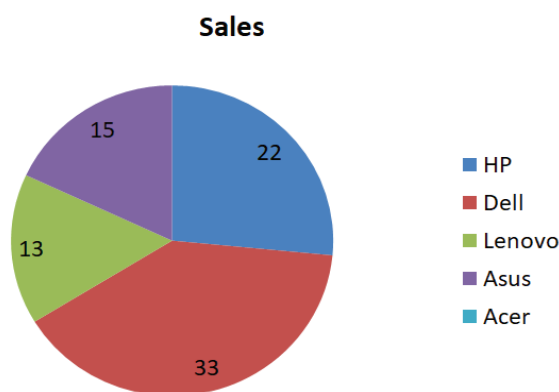
**Check your Progress -3**

- 3 When is simple bar diagram used?
- 4 When do we use a sub-divided bar diagram?

Shows the distribution of sales of the laptop industry between five companies



**NOTES**



**4.3 Let us Sum up:**

The statistical data may be presented in a more attractive form appealing to the eye with the help of graphic aids. Such presentation helps in communicating information to both well read and lay people. It helps viewer's to have an immediate and meaningful grasp of the large amount of data. For graphical representation of data measured on an interval and ratio scale, histogram, polygons and Ogives are used. For graphical representation of data measured on nominal and ordinal scale, bar charts and pie charts are used. The kind of bar diagram you choose depend on the nature of data.

**2.5 Unit – End exercises:**

1. What is the process for constructing histogram?
2. Draw a suitable bar diagram for the following data: The data shows the number of students (Both male and female) graduating from the year 2012-2017 of a city college.

Year of graduation	2012	2013	2014	2015	2016	2017
Gender						
Male	105	110	95	100	85	115
Female	115	125	105	85	150	200

3. State the advantages and disadvantages of graphical representation of data.

**2.6 Answer to check your progress:**

1. The value of graphs lies in the fact that it is an economic device for presentation, understanding and interpretation of collected statistical data.
2. Origin
3. Midpoints of the class –interval
4. From the ogive we can infer relative changes in the data. Percentiles and percentile ranks can be graphically computed from ogives.

**NOTES**

5. Cumulative frequency is the total of a frequency and all frequencies so far in a frequency distribution. It is the 'running total' of frequencies.
6. We use simple bar diagram when we have data in which one variable is classified on a spatial, qualitative or temporal basis.
7. A sub-divided or component bar chart is used when we have to represent data in which the total magnitude has to be divided into different components.

---

**2.7 Suggested readings:**

---

1. Garrett.H.E.(1981). *Statistics in Psychology and Education*. Vakils, Feffer and Simons Ltd.
2. Gupta, S.P. (2014). *Statistical Methods*. (44<sup>th</sup>e.d.) New Delhi :Roopak Printers.
3. Mangal S.K. (1987). *Statistics in Psychology and Education*. New Delhi: PHI Learning Private Ltd.
4. <https://byjus.com/maths/graphical-representation/>
5. <http://www.psychologydiscussion.net/educational-psychology/statistics/measures-of-central-tendency-and-how-it-helps-in-educational-psychology/2741>

---

# UNIT 3: MEASURES OF CENTRAL TENDENCY

---

- 3.1 Introduction
- 3.2 Arithmetic Mean
  - 3.2.1: Computation of mean from ungrouped data
  - 3.2.2: Computation of mean from grouped data
  - 3.2.3 : Properties of the mean
- 3.3 Median
  - 3.3.1: Computation of median from ungrouped data
  - 3.3.2: Computation of median from grouped data
  - 3.3.3 : Properties of the median
- 3.4 Mode
  - 3.4.1: Computation of mode from ungrouped data
  - 3.4.2: Computation of mode from grouped data
  - 3.4.3 Properties of the mode
- 3.5 Geometric Mean
- 3.6 Harmonic Mean.
- 3.7 Let us Sum up
- 3.8 Unit – End exercises
- 3.9 Answer to check your progress
- 3.10 Suggested readings

## NOTES

---

### 3.1 Introduction

---

The purpose of any measure of central tendency is to provide a single summary figure that best describes the central location of our entire distribution of observations. Such measures are useful in comparing the performance of a group with that of a standard reference group.

A measure of central tendency also helps simplify comparisons of two or more groups listed under different conditions. There are many measures of central tendency. The three most commonly used in education and behavioural sciences are:

- a) Arithmetic Mean
- b) Median
- c) Mode

---

### 3.2 Arithmetic Mean:

---

The arithmetic mean is the sum of all the scores in a distribution divided by the total number of scores.

#### 3.2.1: Computation of mean from ungrouped data

Let  $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$  be the scores obtained by 10 students on an achievement test. The arithmetic mean of the group of 10 students can be calculated as,

$$M = \frac{X_1 + X_2 + X_3 + X_4 + \dots + X_{10}}{10}$$

**NOTES**

The formula for calculating the mean of an ungrouped data is

$$M = \frac{\sum X}{N}$$

where X stands for the sum of score or values of the items and N for the total number of items in a series or group.

**3.2.2: i) Computation of mean from grouped data (Long Method)**

Scores	f	Mid points (X)	fX
65 - 69	1	67	67
60 - 64	3	62	186
55 - 59	4	57	228
50 - 54	7	52	364
45 - 49	9	47	423
40 - 44	11	42	462
35 - 39	8	37	296
30 - 34	4	32	128
25 - 29	2	27	54
20 - 24	1	22	22
			-----
-----			$\Sigma fX =$
2230			-----
-----			

In a frequency distribution where all the frequencies are greater than one, the mean is computed by the formula,

$$M = \frac{\sum fX}{N}$$

where X represents the midpoints of the interval, f its frequency and N the total of all frequency

$$M = \frac{\sum fX}{N} = \frac{2230}{50} = 44.6$$

**ii) Short cut method of computing the mean from grouped data (Assumed mean method)**

Mean is computed with the help of the formula

$$M = A + \frac{\sum fx^1}{N} \times i$$

where A = Assumed mean (The midpoint of the class interval with the highest frequency is usually taken as the assumed mean).

i = Class interval

f = Respective frequency of the mid-values of the class intervals

N = Total frequency

$$x' = \frac{X - A}{i}$$

The quotient obtained after division of the difference between the mid-value of the class interval and assumed mean by i, the width class intervals

**NOTES**

**Note: The population mean is denoted by the symbol  $\mu$**

Scores	f	X	$x'$ (X-A)	$fx'$
65 - 69	1	67	5	5
60 - 64	3	62	4	12
55 - 59	4	57	3	12
50 - 54	7	52	2	12
45 - 49	9	47	1	9
40 - 44	11	42	0	0
35 - 39	8	37	-1	-8
30 - 34	4	32	-2	-8
25 - 29	2	27	-3	-6
20 - 24	1	22	-4	-4
	-----			-----
--				
	N = 50			
$\Sigma fx' = 26$				
	-----			-----
---				

$$\mu = A + \frac{\sum fx^1}{N} \times i = 42 + \frac{26}{50} \times 5$$

**NOTES**

= 44.6

**3.2.3 : Properties of the mean**

- a) Mean is responsive to the exact position of each score on the distribution.
- b) The mean may be thought of as balance point of a distribution. This says that if we express the scores in terms of negative and positive deviation their sum is zero.
- c) The mean is more sensitive to the presence or absence of score at the extremes of the distribution. When a measure of central tendency should reflect the total of the scores the mean is the best choice.
- d) Changing the value of any score changes the mean. For example, a sample of quiz scores for a psychology lab section consists of 9, 8, 7, 5, and 1. Note that the sample consists of  $n = 5$  scores with  $\Sigma X = 30$ . The mean for this sample is 6. Now suppose that the score of  $X = 1$  is changed to  $X = 8$ . Note that we have added 7 points to this individual's score, which also adds 7 points to the total ( $\Sigma X$ ). After changing the score, the new distribution consists of 9, 8, 7, 5, 8, the mean becomes 7.4
- e) Adding a new score to a distribution, or removing an existing score, usually changes the mean. The exception is when the new score (or the removed score) is exactly equal to the mean.
- f) If a constant value is added to every score in a distribution, the same constant is added to the mean. Similarly, if you subtract a constant from every score, the same constant is subtracted from the mean.
- g) If every score in a distribution is multiplied by (or divided by) a constant value, the mean changes in the same way.

**Check your progress – 1**

1. Adding a new score to a distribution always changes the mean (True/ False).
2. A population has a mean of  $\mu = 60$ . If 5 points were added to every score, what would be the value for the new mean?
3. What do you understand by the term central tendency?

### 3.3 Median

The Median of a distribution is the point along the scale of possible score below which 50% of the scores fall. The median thus is the value that divides the distribution into halves. Its symbol is Mdn. For raw scores, we may think of the median as a middle score of a distribution based on score frequency. To determine the median therefore we need to know where all the scores are located.

### NOTES

#### 3.3.1: Computation of median from ungrouped data

To find the median we must first put the scores in rank order from lowest to highest. If N is odd, the median will be score that has an equal number of scores below and above it. For example, for the following scores.

1, 7, 8, 11, 15, 16, 20

The median is 11

When there is an even number of scores, there is no middle score, so the median is taken as the point halfway between the two scores that bracket the middle position. The example, for group of scores:

12, 14, 15, 18, 19, 20

the median is  $15 + (18 - 15) / 2 = 16.5$

This again shows that the position of each score in the series needs to be ascertained first and then median should be determined.

#### 3.3.2: Computation of median from grouped data

The formula for computing median from grouped data are as follows:

$$Mdn = L + \left[ \frac{(N/2) - F}{f} \right] \times i$$

L = Exact lower limit of the median classes

F = Total of all frequencies before the median class

f = frequency of the median class

i = Class interval width

N = Total of all the frequencies

**NOTES**

**Illustration: Compute the median for the frequency distribution given below:**

Scores	f
65 - 69	1
60 - 64	3
55 - 59	4
50 - 54	7
45 - 49	9
40 - 44	11
35 - 39	8
30 - 34	4
25 - 29	2
20 - 24	1
	-----
	N = 50
	-----

Since median is the central item, for this distribution it is likely to fall somewhere between the scores of 25th & 26th items. In a given frequency distribution table if we add frequencies either from above or below we may see that the class interval designated as 40-44 is to be labeled as a class where the score representing median will fall.

By applying the above formula we get

$$\begin{aligned} Mdn &= 395 + \left( \frac{50/2 - 15}{11} \right) \times 5 \\ &= 395 + \frac{10}{11} \times 5 = 44.05 \end{aligned}$$

**3.3.3 : Properties of the median**

The median responds to how many score lie below (or above) it but not to how far away the scores may be. A little below the median or way below, both count the same in determining its value. Thus, the median is less sensitive than the mean to the presence of a few extreme scores. Therefore in distribution that are strongly asymmetrical or have few very deviant scores the median may be the better choice for measuring the central tendency, if we wish to represent the bulk of the scores and not give undue weight to the relatively few deviant once.



**NOTES**

In behavioural studies, there are occasion when the researcher cannot record the exact value of scores at the upper end of the distribution. Distributions like this are called open-ended. In open-ended distribution, we cannot calculate the mean without making assumption. But we can find the median. It is also used when there are undetermined (Infinite) scores that make it impossible to compute a mean

Of the three measures of central tendency, the median stands second to the mean in ability to resist the influence of sampling fluctuations in ordinary circumstances. For large sample taken from a normal distribution, the median varies about one quarter more from sample to sample than does the mean. For small samples, the median is relatively better.

---

### **3.4 Mode:**

---

The final measure of central tendency that we consider is called the mode. In its common usage, the word mode means “the customary fashion” or “a popular style.” The statistical definition is similar in that the mode is the most common observation among a group of scores. In a frequency distribution, the mode is the score or category that has the greatest frequency.

#### **3.4.1 : Computation of mode from ungrouped data**

In ungrouped distribution, the mode is the score that occurs with the greatest frequency. In grouped data, it is taken as the midpoint of the class interval that contains the greatest number of scores. The symbol for mode to  $M_o$ .

The mode is a useful measure of central tendency because it can be used to determine the typical or average value for any scale of measurement, inducing a nominal scale . Consider, for example, the data given below

Restaurant	Frequency
College Grill	15
George & Harry’s	16
Luigi’s	42
Oasis Diner	18
Roxbury Inn	7
Sutter’s Mill	12

These data were obtained by asking a sample of 100 students to name their favourite restaurants in the city. For these data, the mode is Luigi’s, the restaurant (score) that was named most frequently as a favourite place. Although we can identify a modal response for these data, you should notice that it would be impossible to compute a mean or a median. For

**NOTES**

example, you cannot add the scores to determine a mean (How much is 5 College Grills plus 42 Luigi's?). Also, it is impossible to list the scores in order because the restaurants do not form any natural order. For example, the College Grill is not "more than" or "less than" the Oasis Diner, they are simply two different restaurants. Thus, it is impossible to obtain the median by finding the midpoint of the list. In general, the mode is the only measure of central tendency that can be used with data from a nominal scale of measurement.

**3.4.3 Properties of the Mode:**

The mode is easy to obtain, but it is not very stable from sample to sample further, when quantitative data are grouped, the mode may be strongly affected by the width and location of class interval. In addition, there may be more than one mode for a particular set of scores. The mode is the only measure that can be used for data that have the character of the nominal scale.

**Check your progress – 2**

4. If you have a score of 82 on an 80- exam point , then you scored definitely above the median(True/False).
5. During the month of October, an instructor recorded the number of absences for each student in a class of n= 20 and obtained the following distribution. Using the mode, What is the average number of absences for a class

No:of absences	6	4	3	1	2	0
F	1	2	7	5	3	2

---

**3.5 Geometric Mean:**

---

The Geometric Mean is a special type of average where we multiply the numbers together and then take a square root (for two numbers), cube root (for three numbers) .For **n** numbers: multiply them all together and then take the **nth root** (written  $\sqrt[n]{}$ )

More formally, the geometric mean of **n** numbers **a<sub>1</sub> to a<sub>n</sub>** is:

$$\sqrt[n]{(a_1 \times a_2 \times \dots \times a_n)}$$

Since it takes into account the compounding that occurs from period to period, in certain contexts the **geometric mean** is more accurate measure

than the arithmetic **mean**, such as in business where it is employed for describing proportional growth, both exponential growth (constant proportional growth) and varying growth

### 3.6 Harmonic Mean

Harmonic mean is a type of average generally used for numbers that represent a rate or ratio such as the precision and the recall in information retrieval. The harmonic mean can be described as the reciprocal of the arithmetic mean of the reciprocals of the data. It is calculated by dividing the number of observations by the sum of reciprocal of the observation. The formula is:

Formula: 
$$\text{Harmonic Mean} = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \frac{1}{a_4} + \dots + \frac{1}{a_n}}$$

Example: For the numbers 4 and 9,

$$\text{Harmonic Mean} = \frac{2}{\frac{1}{4} + \frac{1}{9}} = \frac{72}{13} = 5.54$$

**Harmonic means** are often **used** in averaging things like rates (e.g., the **average** travel speed given aduration of several trips)

### 3.7 Let us Sum up:

The purpose of the central tendency is to determine a single value that identifies the centre of the distribution. The three standard measures of central tendency are the mean, median and mode. The mean is the arithmetic average. It is computed by adding all of the scores and then dividing by the number of scores. Changing any score in the distribution causes the mean to be changed. When a constant value is added to (or subtracted from) every score in a distribution, the same constant value is added to (or subtracted from) the mean. If every score is multiplied by a constant, the mean is multiplied by the same constant. The median is the midpoint of a distribution of scores. The median is the preferred measure of central tendency when a distribution has a few extreme scores. The median also is used for open-ended distributions and when there are undetermined (infinite) scores that make it impossible to compute a mean. Finally, the median is the preferred measure of central tendency for data from an ordinal scale. The mode is the most frequently occurring score in a distribution. For data measured on a nominal scale, the mode is the appropriate measure of central tendency. It is possible for a distribution to have more than one mode. Geometric and Harmonic mean has unique properties but are rarely used in social sciences.

### NOTES

**NOTES**

---

**3.8 Unit – End exercises:**

---

- For the following sample, find the mean, median and mode. The scores are:  
5, 6, 8, 9, 11, 5, 12, 8, 9, 6, 9
- One question on a student survey asks: In a typical week, how many times do you eat at fast food restaurant? The following frequency distribution table summarizes the results for a sample of  $n=20$  students.

Number of times per week	5 or more	4	3	2	1	0
F	2	2	3	6	4	3

- Find the mode for this distribution
  - Find the median for this distribution
  - Explain why you cannot compute the mean using the data in the table.
- A sample of  $n=5$  scores has a mean  $=12$ . If one person with the score of  $X=8$  is removed from the sample, what is the value for the new mean?
  - Identify the circumstances in which the median rather than the mean is a preferred measure of central tendency.
  - Compute mean, median and mode from the following data:

Class intervals	20-	25-	30-	35-	40-	45-	50-	55-
	24	29	34-	39	45	49	54	59
F	<b>2</b>	<b>7</b>	<b>10</b>	<b>11</b>	<b>8</b>	<b>8</b>	<b>2</b>	<b>2</b>

---

**3.9 Answer to check your progress**

---

- False. If the score is equal to the mean, it does not change the mean..
- The new mean score would be 45.
- Central tendency is a statistical measure to determine a single score that defines the centre of a distribution. The goal of central tendency is to find a single score that is most typical or most representative of the entire group.
- False. The value of the median depends on where all of the scores are located. The mode is 3.

---

**3.10 Suggested Readings :**

---

- Gravetter, F.J., & Wallnau, L.B. (2013). Statistics for behavioural sciences (9<sup>th</sup> Ed.). Cengage Learning India Private Ltd.
- Gupta, S.P. (2014). *Statistical Methods*. (44<sup>th</sup> e.d.) New Delhi :RoopakPrinters.

---

## UNIT 4: MEASURES OF DISPERSION

---

### NOTES

- 4.1 Introduction
- 4.2 Range
  - 4.2.1 Properties of the range
- 4.3 Quartile deviation
  - 4.3.1 Properties of the quartile deviation
- 4.4 Standard deviation and Variance for populations
  - 4.4.1 Computation of standard deviation from ungrouped and grouped data
  - 4.4.2 Properties of the Standard deviation
  - 4.4.3 Computation of variance
- 4.5 Standard deviation and Variance for samples
  - 4.5.4: Sampling variability and degrees of freedom
- 4.6 Concept of skewness and kurtosis
  - 4.6.1 Karl Pearson co-efficient of skewness
  - 4.6.2 Bowley's Co-efficient of skewness
  - 4.6.3 Kurtosis
- 4.7 Let us Sum up
- 4.8 Unit – End exercises
- 4.9 Answer to check your progress
- 4.10 Suggested readings

---

### 4.5 Introduction

---

Variability expresses quantitatively the extent to which the scores in the distribution scatter about or cluster together. In general, a good measure of variability serves two purposes:

1. Variability describes the distribution. Specifically, it tells whether the scores are clustered close together or are spread out over a large distance. Usually, variability is defined in terms of distance. It tells how much distance to expect between one score and another, or how much distance to expect between an individual score and the mean. For example, we know that the heights for most adult males are clustered close together, within 5 or 6 inches of the average. Although more extreme heights exist, they are relatively rare.
2. Variability measures how well an individual score (or group of scores) represents the entire distribution. This aspect of variability is very important for inferential statistics, in which relatively small samples are used to answer questions about populations. For example, suppose that you selected a sample of one person to represent the entire population. Because most adult males have heights that are within a few inches of the population average (the distances are small), there is a very good chance that you would select someone whose height is within 6 inches of the

**NOTES**

population mean. On the other hand, the scores are much more spread out (greater distances) in the distribution of weights. In this case, you probably would not obtain someone whose weight was within 6 pounds of the population mean. Thus, variability provides information about how much error to expect if you are using a sample to represent a population. The various measures of variability are:

---

### **4.2 The Range:**

---

The simplest measure of variability is the range. The range is the difference between the highest and lowest score in the distribution. The range is the distance and not a location.

For example range for the following set of scores will be

3, 5, 5, 8, 15, 14, 18, 23

$$\text{Range} = 23 - 3 = 20$$

In a grouped frequency distribution, calculate the range as the difference between the value of the lowest raw score that could be included in the bottom class interval and that of the highest score that could be included in the upper most class interval.

#### **4.2.1 Properties of the Range:**

The range is easier to compute and is ideal for preliminary work or in other circumstances where precision is not an important requirement.

The range has some major shortcoming. The range is not sensitive to the total condition of the distribution. The range is also of little use beyond the descriptive level.

---

### **4.3 Quartile Deviation: (QD)**

---

The quartile deviation, also termed as the semi - inter quartile range, symbolized by the letter Q, is a more sophisticated measure that depends only on the relatively stable central portion of a distribution specifically, on the middle 50% of the score. It is defined as one-half the distance between the first and third quartile points

$$Q = \frac{Q_3 - Q_1}{2}$$

or

$$Q = \frac{P_{75} - P_{25}}{2}$$

The quartile points are the three score points that divide the distribution into four parts, each containing an equal number of cases.

**4.3.1: Properties of the Quartile deviation:**

QD or the semi-inter quartile range is closely, related to the median because both are defined in terms of percentage points of the distribution. The median is responsive to the number of scores lying below it rather than to their exact position and  $P_{25}$  and  $P_{75}$  are points defined in a similar way. We may therefore expect median and the semi inter quartile range to have properties in common.

The semi inter quartile range is less sensitive to the presence of few extreme scores than is the standard deviation. If a distribution is badly skewed or if it contains a few very extreme scores the semi inter quartiles range will respond to the presence of such scores, but it will not give them undue weight. With open-ended distribution the QD may be reasonable measure to compute.

**Check your progress**

1. What is the purpose of measuring variability? The purpose for measuring variability is to obtain an objective measure of how the scores are spread out in a distribution.
2. What is range? It is the distanced covered by the scores in the distribution from the smallest to the largest score
3. Semi interquartile deviation have similar properties to that of

**4.4: Standard deviation and Variance for populations:**

The standard deviation is the most commonly used and the most important measure of variability. Standard deviation uses the mean of the distribution as a reference point and measures variability by considering the distance between each score and the mean. The first step in finding the standard distance from the mean is to determine the *deviation*, or distance from the mean, for each individual score. By definition, the deviation for each score is the difference between the score and the mean. **Deviation** is distance from the mean:

$$\text{Deviation score} = X - \mu$$

For a distribution of scores with  $\mu$  (Population mean) = 50, if your score is  $X = 53$ , then your *deviation score* is  $X - \mu = 53 - 50 = 3$

If your score is  $X = 45$ , then your deviation score is  $X - \mu = 45 - 50 = -5$

Notice that there are two parts to a deviation score: the sign (+ or -) and the number. The sign tells the direction from the mean—that is, whether the score is located above (+) or below (-) the mean. The number gives the actual distance from the mean. For example, a deviation score of -6 corresponds to a score that is below the mean by a distance of 6 points.

Because our goal is to compute a measure of the standard distance from the mean, the

**NOTES**

**NOTES**

obvious next step is to calculate the mean of the deviation scores. To compute this mean, you first add up the deviation scores and then divide by  $N$ . When you do this you will find that sum of deviations around the mean will be zero, because of the positive and negative signs. The solution is to get rid of the signs (+ and -). The standard procedure for accomplishing this is to square each deviation score. Using the squared values, you then compute the *mean squared deviation*, which is called *variance*. **Population variance** equals the mean squared deviation. Variance is the average squared distance from the mean.

Our goal is to compute a measure of the standard distance from the mean. Variance, which measures the average squared distance from the mean, is not exactly what we want. The final step simply takes the square root of the variance to obtain the *standard deviation*, which measures the standard distance from the mean. **Standard deviation** is the square root of the variance and provides a measure of the standard, or average, distance from the mean. In other words, Standard deviation of a set of scores also defined as the square root of the average of the squares of the deviation of each score from the mean. This simply means standard deviation is the square root of the variance

$$SD = \sqrt{\frac{\sum(X - M)^2}{N}} = \sqrt{\frac{\sum x^2}{N}}$$

where

- X = is the individual score
- M = Mean of the given set of scores
- N = Total No. of scores
- x = Deviation of each score from the mean

Standard deviation is regarded as the most stable and reliable measure of variability as it employs the mean for its computation. It is often called root mean square deviation and is denoted by the greek letter sigma ( $\sigma$ ).

#### 4.4.1 Computation of standard deviation (SD) for ungrouped data.

Standard deviation can be computed from the ungrouped scores by the formula.

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

Calculate SD for the following set of scores



52, 50, 56, 68, 65, 62, 57, 70

Mean of the given scores

$$M = 480 / 8 = 60$$

Scores X	Deviation from the mean	$x^2$
52	-8	64
60	-10	100
56	-4	16
68	8	64
65	5	25
62	2	4
57	-3	9
70	10	100
		$\Sigma x^2 = 382$

Applying the obtained values in the formula we get

$$\begin{aligned}\sigma &= \sqrt{\frac{382}{8}} \\ &= \sqrt{47.75} = 6.91\end{aligned}$$

#### Computation of SD from grouped data

Scores	F	X	M	X	$x^2$	$fx^2$
127 – 129	1	128	115	13	169	169
124 – 126	2	125	115	10	100	200
121 – 123	3	122	115	7	49	147
118 – 120	1	119	115	4	16	16
115 – 117	6	116	115	1	1	6
112 – 114	4	113	115	-2	4	16
109 – 111	3	110	115	-5	25	75
106 – 108	2	107	115	-8	64	128
103 – 105	1	104	115	-11	121	121
100 – 102	1	101	115	-14	196	196
	N = 24					$\Sigma fx^2 = 1074$

The formula for computing SD from grouped data is

$$SD = \sqrt{\frac{\Sigma fx^2}{N}} = \sqrt{\frac{1074}{24}} = 6.69$$

#### 4.4.2: Properties of Standard Deviation

The standard deviation, like the mean, is response to be exact position of every score in the distribution. Because it is calculated by

### NOTES

**NOTES**

taking deviations from the mean, if a score is shifted to a position more deviant from the mean, the standard deviation will increase. If the score is shifted to a position closer to the mean, the standard deviation decreases. Thus, the standard deviation is more sensitive to the condition of the distribution than is either the range or semi-inter quartile range.

The standard deviation is more sensitive than the semi - inter quartile range to the presence or absence of scores that lie at the extremes of the distribution. This should be obvious because the semi-inter quartile range is based only on the score points that make the border of the middle 50% of the scores. Because of this characteristic sensitivity, the standard deviation may not be the choice among measures of variability when the distribution contains a few very extreme score or when the distribution is badly skewed. For example, if we compare the variability of two distributions where only one of them contains several extreme score, the extreme scores will exert an influence on the standard deviation that is disproportionate to their relative number. Of course, if N is quite large and the extreme score are very few they will make little difference.

When we calculate deviation from the mean the sum of squares of their values is smaller than if they had been taken about any other point. In other words, it is the point about which the sum of squares of deviation scores is minimum.

One of the most important points favouring our use of standard deviation is its resistance to sampling variation. In repeated random samples drawn from population, of the type most frequently encountered in statistical works, the numerical value of the standard deviation tends to jump about less than would that of other means computed on the same samples. If our concern is in any way associated with inferring variation in a population from knowledge of variation in the sample, the property is clearly of worth.

Standard deviation appears explicitly or lies embedded in many procedures of both descriptive and inferential statistics.

**4.4.3 Computation of variance**

We will calculate the variance and standard deviation for the following population of N = 5 scores: 1, 9, 5, 8, 7

Score X	Deviation $X - \mu$	Squared Deviation $(X - \mu)^2$
1	-5	25
9	3	9
5	-1	1
8	2	4
7	1	1

		40 (Sum of squared deviation)
--	--	-------------------------------

For this set of  $N = 5$  scores, the squared deviations add up to 40. The mean of the squared deviations, the variance, is  $40/5 = 8$

Standard Deviation which is the square root of variance is a measure of dispersion like variance. But it is used more often than variance because the unit in which it is measured is the same as that of mean, a measure of central tendency. The advantage of variance is that it treats all deviations from the mean the same regardless of direction; as a result, the squared deviations cannot sum to zero.

## NOTES

### 4.5: Standard deviation and variance for samples:

The goal of inferential statistics is to use the limited information from samples to draw general conclusions about populations. The basic assumption of this process is that samples should be representative of the populations from which they come. This assumption poses a special problem for variability because samples consistently tend to be less variable than their populations. The fact that a sample tends to be less variable than its population means that sample variability gives a biased estimate of population variability. This bias is in the direction of underestimating the population value rather than being right on the mark. Fortunately, the bias in sample variability is consistent and predictable, which means it can be corrected.

The calculations of variance and standard deviation for a sample follow the same steps that were used to find population variance and standard deviation.

1. Find the deviation from the mean for each score: deviation =  $X - M$  ( $M$  represents sample mean while  $\mu$  represents population mean)
2. Square each deviation: squared deviation =  $(X - M)^2$
3. Add the squared deviations: Sum of squared deviation =  $\Sigma (X - M)^2$  or **SS (Sum of squares)**

$$\text{Sample variance } s^2 = \Sigma (X - M)^2 / n-1$$

Note that the denominator is  $n-1$ , instead of  $n$ . This is the adjustment that is necessary to correct for the bias in sample variability. The effect of the adjustment is to increase the value that you obtain. Dividing by a smaller number ( $n - 1$  instead of  $n$ ) produces a larger result and makes sample variance an accurate and unbiased estimator of population variance.

**4.5.1: Sample Variability and Degrees of Freedom:** Although the concept of a deviation score and the calculation of **Sum of squares (SS)** are almost exactly the same for samples and populations, the minor

**NOTES**

differences in notation are really very important. Specifically, with a population, you find the deviation for each score by measuring its distance from the population mean,  $\mu$ . With a sample, on the other hand, the value of  $\mu$  is unknown and you must measure distances from the sample mean. Because the value of the sample mean varies from one sample to another, you must first compute the sample mean before you can begin to compute deviations. However, calculating the value of  $M$  places a restriction on the variability of the scores in the sample. This restriction is demonstrated in the following example.

Suppose we select a sample of  $n = 3$  scores and compute a mean of  $M = 5$ . The first two scores in the sample have no restrictions; they are independent of each other and they can have any values. For this demonstration, we assume that we obtained  $X = 2$  for the first score and  $X = 9$  for the second. At this point, however, the third score in the sample is restricted

X	Sample of $n = 3$ scores with a mean of $M = 5$ .
2	
9	
?	What is the third score?

. For this example, the third score must be  $X = 4$ . The reason that the third score is restricted to  $X = 4$  is that the entire sample of  $n = 3$  scores has a mean of  $M = 5$ . For 3 scores to have a mean of 5, the scores must have a total of  $X = 15$ . Because the first two scores add up to 11 ( $9 + 2$ ), the third score must be  $X = 4$ .

In this the first two out of three scores were free to have any values, but the final score was dependent on the values chosen for the first two. In general, with a sample of  $n$  scores, the first  $n - 1$  scores are free to vary, but the final score is restricted. As a result, the sample is said to have  $n - 1$  **degrees of freedom**. For a sample of  $n$  scores, the degrees of freedom, or  $df$ , for the sample variance are defined as  $df = n - 1$ . The degrees of freedom determine the number of scores in the sample that are independent and free to vary.

**Check your progress 2**

4. What does standard deviation and variance measure?
5. The deviation scores are calculated for each individual in a population of  $N = 4$ . The first three individuals have deviations of 2, 4, and  $-1$ . What is the deviation for the fourth individual?
6. What is the standard deviation for the following set of  $N = 5$  scores: 10, 10, 10, 10, and 10?
7. Explain why the formula for sample variance divides  $SS$  by  $n - 1$  instead of dividing by  $n$ .

## 4.6 Concept of Skewness and Kurtosis:

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.

### NOTES

#### 4.6.1 Karl Pearson Co-efficient of Skewness

Karl Pearson developed two methods to find skewness in a sample.

1. Pearson's Coefficient of Skewness method 1 uses the mode. The formula is:

$$Sk_1 = \frac{\bar{X} - Mo}{s}$$

Where  $\bar{X}$  = the mean, Mo = the mode and s = the standard deviation for the sample.

2. Pearson's Coefficient of Skewness method 2 uses the median. The formula is:

$$Sk_2 = \frac{3(\bar{X} - Md)}{s}$$

Where  $\bar{X}$  = the mean, Mo = the mode and s = the standard deviation for the sample.

**Example :** Use Pearson's Coefficient #1 and #2 to find the skewness for data with the following characteristics:

- Mean = 70.5.
- Median = 80.
- Mode = 85.
- Standard deviation = 19.33.

**Pearson's Coefficient of Skewness #1 (Mode):**

Step 1: Subtract the mode from the mean:  $70.5 - 85 = -14.5$ .

Step 2: Divide by the standard deviation:  $-14.5 / 19.33 = -0.75$ .

**Pearson's Coefficient of Skewness #2 (Median):**

Step 1: Subtract the median from the mean:  $70.5 - 80 = -9.5$ .

Step 2: Multiply Step 1 by 3:  $-9.5(3) = -28.5$

Step 2: Divide by the standard deviation:  $-28.5 / 19.33 = -1.47$ .

NOTES

**Caution:** Pearson’s first coefficient of skewness uses the mode. Therefore, if the mode is made up of too few pieces of data it won’t be a stable measure of central tendency.

**Interpretation**

- The direction of skewness is given by the sign.
- The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the distribution differs from a normal distribution.
- A value of zero means no skewness at all.
- A large negative value means the distribution is negatively skewed (happens when there are more individuals at the right side (positive end) of the distribution).
- A large positive value means the distribution is positively skewed (happens when there are more individuals at the right side (negative end) of the distribution).

**4.6.2 Bowley’s Co-efficient of skewness :**

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 values is ignored. Coefficient of skewness lies within the limit  $\pm 1$ . This method is quite convenient for determining skewness where one has already calculated quartiles.

**4.6.3 Kurtosis**

When there are very few individuals whose scores are near to the average score for their group (too few cases in the central area of the curve) the curve representing such a distribution becomes ‘flattened in the middle. On the other hand, when there are too many cases in central area, the distribution curve becomes too ‘peaked in comparison to normal’. Both these characteristics of being flat (Platykurtic) or peaked (Leptokurtic), are used to describe the term Kurtosis.

**Check your progress 3**

8. When the frequency distribution graph is flattened at the top it is termed as \_\_\_\_\_
9. In an easy test, many individuals score higher than normal. When the scores are represented graphically, it is likely to result in \_\_\_\_\_.

---

**4.7 Let us Sum up:**

The purpose of variability is to measure and describe the degree to which the scores in a distribution are spread out or clustered together.

**NOTES**

In this unit, we covered four measures of variability: the range, the quartile deviation, the variance, and the standard deviation. The range is the distance covered by the set of scores, from the smallest score to the largest score. The range is completely determined by the two extreme scores and is considered to be a relatively crude measure of variability.

Standard deviation and variance are the most commonly used measures of variability. Both of these measures are based on the idea that each score can be described in terms of its deviation, or distance, from the mean. The variance is the mean of the squared deviations. The standard deviation is the square root of the variance and provides a measure of the standard distance from the mean. We also studied how to describe distributions in which the scores are erratically scattered through the concept of skewness and kurtosis.

---

#### **4.8 Unit – End exercises**

1. Calculate the variance for the following population of  $N = 5$  scores: 4, 0, 7, 1, 3.
2. For the following set of scores: 1, 5, 7, 3, 4. Assume that this is a sample of  $n = 5$  scores and compute SS and variance for the sample.
3. Write the properties of quartile deviation.
4. Explain the concept of degrees of freedom.

---

#### **4.9 Answer to check your progress**

1. The purpose for measuring variability is to obtain an objective measure of how the scores are spread out in a distribution.
2. It is the distance covered by the scores in the distribution from the smallest to the largest score
3. Median
4. Standard deviation measures the standard distance from the mean and variance measures the average squared distance from the mean.
5. The deviation scores for the entire set must add up to zero. The first four deviations add to 5 so the fifth deviation must be  $-5$ .
6. Because there is no variability (the scores are all the same), the standard deviation is zero.
7. Without some correction, sample variability consistently underestimates the population variability. Dividing by a smaller number ( $n - 1$  instead of  $n$ ) increases the value of the sample variance and makes it an unbiased estimate of the population variance.
8. Platykurtic
9. Negatively skewed.

**NOTES**

---

**4.10 Suggested readings:**

---

6. Gupta, S.P. (2014). *Statistical Methods*. (44<sup>th</sup>e.d.) New Delhi :Roopak Printers.
7. Garrett.H.E.(1981). *Statistics in Psychology and Education*. Vakils, Feffer and Simons Ltd.
8. [https://us.sagepub.com/sites/default/files/upm-binaries/66956\\_Tokunaga\\_Chapter\\_4.pdf](https://us.sagepub.com/sites/default/files/upm-binaries/66956_Tokunaga_Chapter_4.pdf)



---

## UNIT 5: CORRELATION

---

### 5.1: Introduction to Correlation

5.1.1: Characteristics of relationship

### 5.2 Scatter Diagram

### 5.3 Methods of computing co-efficient of correlation

5.3.1 Karl Pearson co-efficient of correlation

i. Spearman's rank correlation

### 5.4 Regression

5.4.1: Regression Line

5.4.2: Regression Equation

5.6.1 Properties of regression co-efficient.

- b. Let us Sum up
- c. Unit – End exercises
- d. Answer to check your progress

### 5.8 Suggested Readings

---

### 5.1 Introduction to Correlation

---

**Correlation** : Correlation is a statistical technique that is used to measure and describe a relationship between two variables. Usually the two variables are simply observed as they exist naturally in the environment - there is no attempt to control or manipulate the variables.

For example, a researcher interested in the relationship between nutrition and IQ could observe and record dietary patterns for a group of pre-school children and then measures IQ scores for the same group. Notice that the researcher is not trying to manipulate the children's diet or IQ but is simply observing what occurs naturally. Correlation requires two scores for each individual (one score from each of the two variables). These scores are normally identified as X and Y. The pairs of scores can be tested in a table or can be presented graphically in a scatter plot.

In the scatter plot, the X values are placed on the vertical axis. Each individual is then identified by a single point on the graph so that the co-ordinates of the point (the X & Y values) match the individual's X score and Y score.

#### 5.1.1 The characteristics of a relationship

A correlation measures three characteristics of the relationship between X and Y. The three characteristics are as follows

### NOTES

**NOTES**

- 1. The direction of the relationship:** Correlations can be classified into two basic categories positive and negative.

In positive correlation, the two variables tend to move in the same direction. When the X variable increases, the Y variable also increases, if the X variable decreases, the Y variable also decreases.

In negative correlation, the two variables tend to go in opposite direction. As the X variable increases, the Y variable decreases that is, it is an inverse relationship.

The direction of a relationship is identified by the sign of the correlation. A positive value (+) indicates a positive relationship; a negative value (-) indicates a negative relationship.

- 2. The Form of Relationship**

The most common use of correlation is to measure straight - line relationships. However, others forms of relationships do exist and that there are special correlations used to measure them. For example the relationship between reaction time and age. Reaction time improves with age until the late teens, when it reaches a peak, after that reaction time starts to get worse. This shows a curved relationship.

- 3. The Degree or the strength of Relationship**

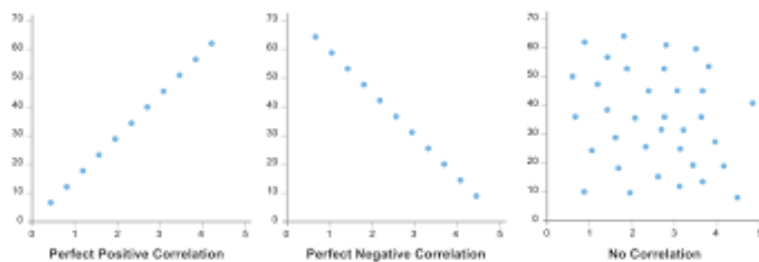
Finally, a correlation measures how well the data fit the specific form being considered. For example, a linear correlation measures how well the data points fit on a straight line. A perfect correlation always is identified by a correlation of 1.00 and indicates a perfect fit. At the other extreme, a correlation of 0 indicates no fit at all. Intermediate values represent the degree to which the data points approximate the perfect fit. The numerical value of the correlation also reflects the degree to which there is a consistent, predictable relationship between the two variables. Again, a correlation of 1.00 (or - 1.00) indicates a perfect consistent relationship.

---

**5.2: The scatter diagram:**

---

A Scatter (XY) Plot has points that show the relationship between two sets of data. Pairs of numerical data, with one variable on each axis are plotted to look for a relationship between them. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the tighter the points will hug the line. Below are examples of scatter plots showing positive , negative and zero correlation.

**NOTES**


---

### 5.3 Methods of computing Co-efficient of correlation

---

For expressing the degree of relationship quantitatively between two sets of measures or variables, we usually take the help of an index that is known as co-efficient of correlation. It is a ratio which expresses the extent to which changes in one variable are accompanied by change in the other variable. It involves no units and varies from - 1 to +1. In case the co-efficient of correlation is zero, it indicates zero correlation between two sets of measures.

#### Check your Progress -1

1. For each of the following, indicate whether you would expect a positive or a negative correlation.
  - a. Model year and price for a used Honda
  - b. IQ and grade point average for high school students
  - c. Daily high temperature and daily energy consumption for 30 winter days in New York City
  
2. The data points would be clustered more closely around a straight line for a correlation of  $-0.80$  than for a correlation of  $0.05$ . (True or false?)

#### 5.3.1 Pearson Product moment correlation

The Pearson's correlation measures the degree and the direction of linear relationship between two variables.

The Pearson's correlation is identified by the letter  $r$  and computed by the formula.

$$r = \frac{\text{degree to which } x \text{ and } y \text{ vary together}}{\text{degree to which } x \text{ and } y \text{ vary separately}}$$

When there is a perfect linear relationship, every change in the X variable is accompanied by a corresponding change in the Y variable.

To calculate the Pearson's correlation, it is necessary to compute the sum of products of deviations. The sum of products provides a parallel

**NOTES**

procedure for measuring the amount of co - variability between two variables. The Pearson's correlation consists of a ratio comparing the co-variability of X and Y (the numerator) with the variability of X and Y separately (the denominator).

**Computation correlation of Pearson Product moment**

Individuals	X	Y	X	Y	Xy	x <sup>2</sup>	y <sup>2</sup>
A	15	60	-10	10	-100	100	100
B	25	70	0	20	0	0	400
C	20	40	-5	-10	50	25	100
D	30	50	5	0	0	25	0
E	35	30	10	-20	-200	100	400

$$\Sigma xy = 250 \quad \Sigma x^2 = 250$$

$$\Sigma y^2 = 1000$$

Mean of series x,  $\mu_x = 25$

Mean of series y,  $\mu_y = 50$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 y^2}} = \frac{-250}{\sqrt{250 \times 1000}}$$

$$= 0.5$$

xy- sum of product of deviations from the mean.

**Understanding and interpreting the Pearson correlation.**

There are three additional consideration that we need to keep in mind while dealing with Pearson correlation.

1. Correlation simply describes a relationship between two variables. It does not explain why the two variables are related. The correlation cannot be interpreted as proof of a cause and effect relationship between two variables.
2. The value of a correlation can be affected greatly by the range of scores represented in the data.
3. When judging "how good" a relationship is, it is tempting to focus on the numerical value of the correlation for example, a correlation of +.5 is halfway between 0 and 1.00 and therefore appears to represent a moderate degree of relationship. However correlation should not be interpreted as a proportion. Although a correlation of 1.00 does mean that there is a 100% perfectly predictable

**NOTES**

relationship between X and Y, a correlation of 0.5 does not mean that you can make predictions with 50% accuracy. To describe how accurately one variable predicts the other, you must square the correlation. Thus, a correlation 0.5 provides only  $r^2 = 0.5^2 = 0.25$  or 25% accuracy. The value of  $r^2$  is called as the **co-efficient of determination** because it measures the proportion of variability in one variable that can be determined from the relationship with the other variable. Computed correlation co-efficient may be understood by summarizing it as follows.

<b>The range of computed correlation co-efficient</b>	<b>Interpretation</b>
1. 0 (zero value)	Zero correlation, absolutely no relationship
2. From 0.00 to $\pm 0.20$	Slight, almost negligible relationship
3. From $\pm 0.21$ to $\pm 0.40$	Low correlation, definite but small relationship
4. From $\pm 0.41$ to $\pm 0.70$	Moderate correlation, substantial but small relationship
5. From $\pm 0.71$ to $\pm 0.90$	High correlation, marked relationship
6. From $\pm 0.91$ to 0.99	Very high correlations, quite dependable relationship
7. $\pm 1$	Perfect correlation, almost identical and opposite relationship

The use of the product moment method for the computation of correlation co-efficient between two variable is based on the following **assumptions**.

1. Linearity of relationship: The relationship should be linear (described by a straight line).
2. Homoscedasticity: The standard deviation of the scores in the different columns and rows should be equal and fairly homogenous.
3. Continuity of the variables: The two variables for which correlation is to be computed should be continuous variables in terms of measurement.
4. Normality of the distribution: Distributions in two variables should be fairly symmetrical and unimodal. In other words it should not be badly skewed.

**NOTES****5.3.2 Spearman's Rank-Order correlation**

For computing the co-efficient of correlation between two sets of scores achieved by individuals with the help of this method, we require rank; ie position of merit of these individuals in possession of certain characteristics. The co-efficient of correlation computed by this method, is known as the rank correlation co-efficient as it considers only the ranks of the individuals in the characteristics A and B. It is designated by the Greek letter  $\rho$  (rho). Sometimes, it is also known as Spearman's co-efficient of correlation after the name of its inventor.

It we do not have scores and have to work with data in which differences between the individuals can be expressed by ranks, rank correlation co-efficient is the only way. But this does not mean that it cannot be computed from the usual data given in raw scores. In case the data contains scores of individuals, we can compute  $\rho$  by converting individual scores into ranks.

**Computation of Rank correlation co-efficiencies:**

Individuals	Marks in History (X)	Marks in civics (Y)	Ranks in History (R <sub>1</sub> )	Ranks in Civics (R <sub>2</sub> )	Difference ranks, irrespective of +ve or - ve signs R <sub>1</sub> - R <sub>2</sub> =  d	Difference squared (d <sup>2</sup> )
A	80	82	2	3	1	1
B	45	86	11	2	9	81
C	55	50	10	10	0	0
D	56	48	9	11	2	4
E	58	60	8	9	1	1
F	60	62	7	8	1	1
G	65	64	6	7	1	1
H	68	65	5	6	1	1
I	70	70	4	5	1	1
J	75	74	3	4	1	1
K	85	90	1	1	0	0
						$\Sigma d^2=92$

$$\rho = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$

$$= 1 - \frac{6 \times 92}{11(112 - 1)}$$

$$= 1 - \frac{23}{55} = 1 - 0.42 = 0.58$$

### Check your progress – 2

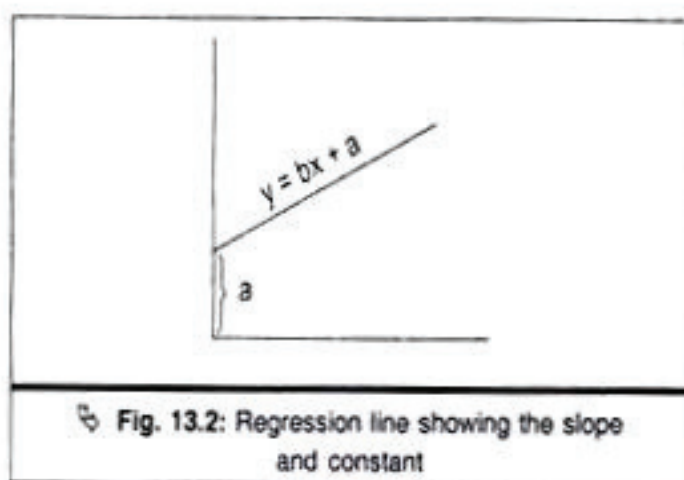
3. When is a spearman's rho used?
4. What is measured by pearson correlation?
5. How is co-efficient of correlation different from coefficient of determination?

### 5.4 Regression:

The statistical technique for finding the best-fitting straight line for a set of data is called **regression**, and the resulting straight line is called the **regression line**. Regression is used to denote the estimation and prediction of the average value of one variable for a specified value of the other variable. This estimation is done by deriving a suitable equation on the basis of available bivariate data. This equation is called Regression equation and its geometrical representation is called Regression curve. The regression equation requires the Regression coefficient.

#### 5.4.1 Regression Line:

When the bivariate data are plotted on graph paper, the concentration of points shows certain pattern showing the relationship. When the trend points are found to be linear then by least square method we can obtain the regression line.



If two variables are linearly related then the relation can be expressed as  $y = bx + a$ , where 'b' is the slope of the line and 'a' is the intercept of that line. This line serves several purposes. 1. The line makes the relationship

### NOTES

**NOTES**

between the two variables easier to see. **2.** The line identifies the center, or *central tendency*, of the relationship, just as the mean describes central tendency for a set of scores. **3.** Finally, the line can be used for prediction.

**5.4.2 Regression Equation:**

In general, a *linear relationship* between two variables  $X$  and  $Y$  can be expressed by the equation

$$Y = bX + a \text{ (where } a \text{ and } b \text{ are fixed constants)}$$

For example, a local video store charges a membership fee of \$5 per month, which allows you to rent videos and games for \$2 each. With this information, the total cost for 1 month can be computed using a *linear equation* that describes the relationship between the total cost ( $Y$ ) and the number of videos and games rented ( $X$ ).

$$Y = 2X + 5$$

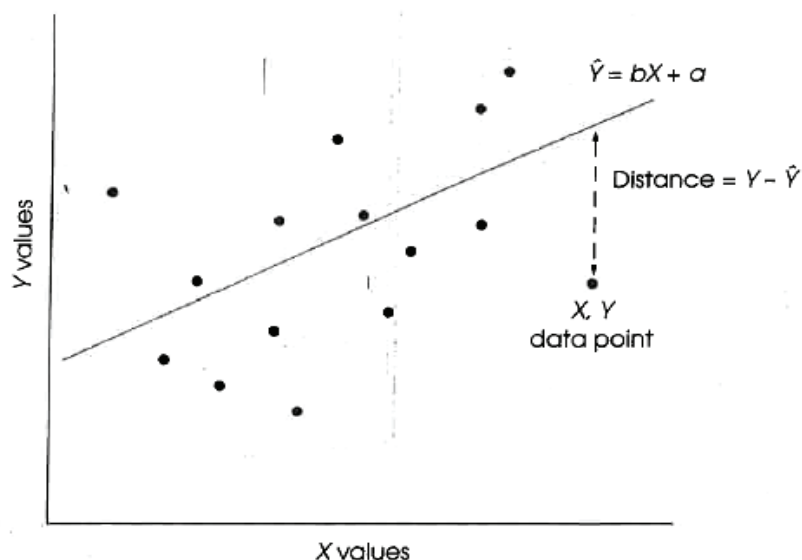
In the general linear equation, the value of  $b$  is called the *slope*. The slope determines how much the  $Y$  variable changes when  $X$  is increased by 1 point. For the video store example, the slope is  $b = 2$  and indicates that your total cost increases by \$2 for each video you rent. The value of  $a$  in the general equation is called the *Y-intercept* because it determines the value of  $Y$  when  $X = 0$ . (On a graph, the  $a$  value identifies the point where the line intercepts the  $Y$ -axis.) In other words, Constant  $a$  is distance between the point of origin and the point where the regression line touches  $Y$  axis. Constant  $b$  is the steepness or slope of line and is also called coefficient of regression. In the video store example,  $a = 5$ ; there is a \$5 membership charge even if you never rent a video.

When $X = 3$	When $X = 8$
$Y = bX + a$	$Y = bX + a$
$= \$2(3) + \$5$	$= \$2(8) + \$5$
$= \$6 + \$5$	$= \$16 + \$5$
$= \$11$	$= \$21$

**Regression Equation through Least Square Method:**

Regression equation can be derived through least square method. Linear regression equations are the equations of those lines of best fit which are made on the basis of least square method. To determine how well a line fits the data points, the first step is to define mathematically the distance between the line and each data point.





**Figure: The X and Y data points and the regression line for the  $n = 8$  pairs of scores**

For every  $X$  value in the data, the linear equation determines a  $Y$  value on the line. This value is the predicted  $Y$  and is called  $\hat{Y}$  ("Y hat"). The distance between this predicted value and the actual  $Y$  value in the data is determined by

$$\text{distance} = Y - \hat{Y}$$

Note that we simply are measuring the vertical distance between the actual data point. This distance measures the error between the line and the actual data. Because some of these distances are positive and some are negative, the next step is to square each distance to obtain a uniformly positive measure of error. Finally, to determine the total error between the line and the data, we add the squared errors for all of the data points. The result is a measure of overall squared error between the line and the data:

$$\text{total squared error} = \sum(Y - \hat{Y})^2$$

Now we can define the *best-fitting* line as the one that has the smallest total squared - error. For obvious reasons, the resulting line is commonly called the *least-squared-error solution*. In symbols, we are looking for a linear equation of the form

$$\hat{Y} = bX + a$$

## NOTES

**NOTES**

For each value of  $X$  in the data, this equation determines the point on the line ( $Y$ ) that gives the best prediction of  $Y$ . The problem is to find the specific values for  $a$  and  $b$  that make this the best-fitting line.

The calculations that are needed to find this equation require calculus and some sophisticated algebra, so we do not present the details of the solution. The results, however, are relatively straightforward, and the solutions for  $b$  and  $a$  are as follows:

$$b = r \frac{SP}{SS_x}$$

where  $SP$  is the sum of products and  $SS_x$  is the sum of squares for the  $X$  scores.

A commonly used alternative formula for the slope is based on the standard deviations for  $X$  and  $Y$ . The alternative formula is

$$b = r \frac{s_y}{s_x}$$

where  $s_y$  is the standard deviation for the  $Y$  scores,  $s_x$  is the standard deviation for the  $X$  scores, and  $r$  is the Pearson correlation for  $X$  and  $Y$ . The value of the constant  $a$  in the equation is determined by

$$a = M_y - bM_x$$

Note that these formulas determine the linear equation that provides the best prediction of  $Y$  values. This equation is called the *regression equation for  $Y$*

$$\hat{Y} = bX + a$$

The scores in the following table are used to demonstrate the calculation and use of the regression equation for predicting  $Y$ .

X	Y	$X - M_x$	$Y - M_y$	$(X - M_x)^2$	$(Y - M_y)^2$	$(X - M_x)(Y - M_y)$
2	3	-2	-5	4	25	-10
6	11	2	3	4	9	6
0	6	-4	-2	16	4	8
4	6	0	-2	0	4	0
7	12	3	4	9	16	12
5	7	1	-1	1	1	-1
5	10	1	2	1	4	2
3	9	-1	1	1	1	-1
				$SS_x = 36$	$SS_y = 64$	$SP = 36$

**NOTES**

For these data,  $\sum X = 32$ . so  $M_x = 4$ . Also,  $\sum Y = 64$ , so  $M_y = 8$ . These values have been used to compute the deviation scores for each  $X$  and  $Y$  value. The final three columns show the squared deviations for  $X$  and for  $Y$ , and the products of the deviation scores.

Our goal is to find the values for  $b$  and  $a$  in the regression equation. Using Equations, the solutions for  $b$  and  $a$  are

$$b = \frac{SP}{SS_x} = \frac{36}{36} = 1.00$$

$$a = M_y - bM_x = 8 - 1(4) = 4.00$$

The resulting equation is

$$\hat{Y} = X + 4$$

**USING THE REGRESSION EQUATION FOR PREDICTION**

As we noted at the beginning of this section, one common use of regression equations is for prediction. For any given value of  $X$ , we can use the equation to compute a predicted value for  $Y$ . For the equation from

**NOTES**

Example, an individual with a score of  $X = 1$  would be predicted to have a  $Y$  score of

$$\hat{Y} = X + 4 = 1 + 4 = 5$$

Although regression equations can be used for prediction, a few cautions should be considered whenever you are interpreting the predicted values:

**USING THE REGRESSION EQUATION FOR PREDICTION**

1. The predicted value is not perfect (unless  $r = +1.00$  or  $-1.00$ ). Although the amount of error varies from point to point, on average the errors are directly related to the magnitude of the correlation. With a correlation near 1.00 (or - 1.00), the data points generally are clustered close to the line and the error is small. As the correlation gets nearer to zero, the points move away from the line and the magnitude of the error increases.
2. The regression equation should not be used to make predictions for  $X$  values that fall outside of the range of values covered by the original data.

---

**5.5 Properties of regression co-efficient:**


---

1. It is denoted by  $b$ .
2. It is expressed in terms of original unit of data.
3. Between two variables (say  $x$  and  $y$ ), two values of regression coefficient can be obtained. One will be obtained when we consider  $x$  as independent and  $y$  as dependent and the other when we consider  $y$  as independent and  $x$  as dependent. Both regression coefficients must have the same sign.
4. If one regression coefficient is greater than unity, then the other regression coefficient must be lesser than unity.

**Check your progress -3**

6..A local gym charges a \$25 monthly membership fee plus \$2 per hour for aerobics classes. What is the linear equation that describes the relationship between the total monthly cost ( $Y$ ) and the number of class hours each month ( $X$ )?

7. For the following linear equation, what happens to the value of  $Y$  each time  $X$  is increased by 1 point?

$$Y = 3X + 7$$

8. If the slope constant ( $b$ ) in a linear equation is positive, then a graph of the equation is a line tilted from lower left to upper right. (True or false?)

### 5.7 Let us Sum up:

A correlation measures the relationship between two variables,  $X$  and  $Y$ . The relationship is described by three characteristics: Direction, Form and strength of the relationship. The most commonly used correlation is the Pearson correlation, which measures the degree of linear relationship. The Pearson correlation is identified by the letter  $r$ . A correlation between two variables should not be interpreted as implying a causal relationship. Simply because  $X$  and  $Y$  are related does not mean that  $X$  causes  $Y$  or that  $Y$  causes  $X$ . To evaluate the strength of a relationship, you square the value of the correlation. The resulting value,  $r^2$ , is called the *coefficient of determination* because it measures the portion of the variability in one variable that can be predicted using the relationship with the second variable.

The Spearman correlation measures the consistency of direction in the relationship between  $X$  and  $Y$ —that is, the degree to which the relationship is one-directional, or monotonic. A correlation always has a value from +1.00 to -1.00. If you obtain a correlation outside this range, then you have made a computational error. When interpreting a correlation, do not confuse the sign (+ or -) with its numerical value. The sign indicates the direction of the relationship between  $X$  and  $Y$ . On the other hand, the numerical value reflects the strength of the relationship. We can sketch a scatter plot of the data and make an estimate of the correlation.

When there is a general linear relationship between two variables,  $X$  and  $Y$ , it is possible to construct a linear equation that allows you to predict the  $Y$  value corresponding to any known value of  $X$ . Predicted  $Y$  value =  $\hat{Y} = bX + a$ . The technique for determining this equation is called regression. By using a *least-squares* method to minimize the error between the predicted  $Y$  values and the actual  $Y$  values, the best-fitting line is achieved

### 5.7 Unit – End exercises

1. For the following data, compute the Pearson correlation.

X	2	1	3	2	4
Y	5	6	3	7	4

2. Rank the following scores and compute the Spearman correlation:

X	2	12	9	10	11
Y	7	35	6	19	8

## NOTES

**NOTES**

3. List assumptions that our data needs to satisfy to compute Pearson's product moment correlation.

4. Compute linear regression for the following set of scores

X	0	2	8	6	4
Y	4	1	10	9	6

---

### 5.8 Answer to check your progress

---

1. a. Positive: Higher model years tend to have higher prices.  
 b. Positive: More intelligent students tend to get higher grades.  
 c. Negative: Higher temperature tends to decrease the need for heating.
2. True. The numerical value indicates the strength of the relationship. The sign only indicates the direction of the relationship.
3. It is used when the researcher wants to establish relationship between two sets of scores measured on an ordinal scale.
4. The Pearson correlation measures the degree and direction of the linear relationship between two variables.
5. While coefficient of correlation tells whether changes in one variable are accompanied by changes in the other, co-efficient of determination measures the proportion of variability in one variable that can be determined from the relationship with the other variable.
6.  $Y = 2X + 25$
7. The slope is - 3, so  $Y$  decreases by 3 points each time  $X$  increases by 1 point.
8. True. A positive slope indicates that  $Y$  increases (goes up in the graph) when  $X$  increases

---

### 5.9 Suggested Readings

---

1. Garrett, H.E. (1981). *Statistics in Psychology and Education*. Vakils, Feffer and Simons Ltd.
2. Gravetter, F.J., & Wallnau, L.B. (2013). *Statistics for behavioural sciences* (9<sup>th</sup> Ed.). Cengage Learning India Private Ltd.
3. <https://www.simplypsychology.org/correlation.html>

## UNIT 6: CONCEPT OF PROBABILITY

### NOTES

#### 6.1 Introduction

#### 6.2 Approaches to Probability

#### 6.3 Random Sampling

#### 6.4 Basic Rules of Probability

##### 6.4.1 Addition theorem

##### 6.4.2 Multiplication theorem for dependent and independent events

#### 6.5 Let us Sum up

#### 6.6 Unit – End exercises

#### 6.7 Answer to check your progress

#### 6.8 Suggested Readings

### 6.1 Introduction:

The role of inferential statistics is to use the sample data as the basis for answering questions about the population. To accomplish this goal, inferential procedures are typically built around the concept of probability. Specifically, the relationships between samples and populations are usually defined in terms of probability. Suppose, for example, that you are selecting a single marble from a jar that contains 50 black and 50 white marbles. (In this example, the jar of marbles is the *population* and the single marble to be selected is the *sample*.) Although you cannot guarantee the exact outcome of your sample, it is possible to talk about the potential outcomes in terms of probabilities. In this case, you have a 50-50 chance of getting either colour. Now consider another jar (population) that has 90 black and only 10 white marbles. Again, you cannot predict the exact outcome of a sample, but now you know that the sample probably will be a black marble. By knowing the makeup of a population, we can determine the probability of obtaining specific samples. In this way, probability gives us a connection between populations and samples, and this connection is the foundation for the inferential statistics

A random experiment is an action or process that leads to one of many possible outcomes. Examples:

Experiment	Outcomes
Flip a coin	Heads , Tails
Roll a dice	1,2,3,4,5,6
Exam Marks	0-100
Course Grades	A,B,C,D,

The list of possible outcomes of a random experiment must be exhaustive and mutually exclusive. The set of all possible outcomes of an experiment is called the **sample space**. An individual outcome in the sample space is

## NOTES

called a simple event, while an event is a collection or set of one or more simple events in a sample space.

**Given a sample space  $S = \{O_1, O_2, \dots\}$ , the probabilities assigned to events must satisfy these requirements:**

1. The probability of any event must be nonnegative, e.g.,  $P(O_i) \geq 0$  for each  $i$ .
2. The probability of the entire sample space must be 1, i.e.,  $P(S) = 1$ .
3. For two disjoint events  $A$  and  $B$ , the probability of the union of  $A$  and  $B$  is equal to the sum of the probabilities of  $A$  and  $B$ , i.e.,  $P(A \cup B) = P(A) + P(B)$ .

---

### 6.2 Approaches to probability:

---

**a) Classical approach to assigning values to events:** If there are a finite number of possible outcomes of an experiment, all equally likely and mutually exclusive, then the **probability of an event is the number of outcomes favourable to the event, divided by the total number of possible outcomes**. Customary examples include tossing an unbiased coin or throwing a balanced dice.

**Example 1:** What are the chances of getting a 'Head' in tossing an unbiased coin? There are only two equally likely outcomes, namely head and tail. In our day to day language, we say that the coin has chance 1 in 2 of showing up a head. In technical language, we say that the probability of getting a head is  $1/2$ .

**Example 2:** Similarly, in the experiment of rolling a dice, there are six equally likely outcomes 1, 2, 3, 4, 5 or 6. The face with number '1' has chance 1 in 6 of appearing on the top. Thus, we say that the probability of getting 1 is  $1/6$ .

Probability can be defined as a proportion, or a part of the whole. This definition makes it possible to restate any probability problem as a proportion problem.

**Example 3:** The probability problem "What is the probability of selecting a king from a deck of cards?" can be restated as "What proportion of the whole deck consists of kings?" In each case, the answer is  $4/52$ , or "4 out of 52." This translation from probability to proportion may seem trivial now, but it is a great aid when the probability problems become more complex.

**b) Relative frequency Approach:** Most situations, however, do not involve equally likely outcomes. Nor does this definition explain what



**NOTES**

probability is, it just states how to assign a numeric value to this *primitive* idea in certain simple cases. Therefore there is another approach of assigning probable values. In this approach, the probability of an event denotes the relative frequency of occurrence of that event in the long run. This means that to estimate the probability the experiment has to be repeated indefinitely under identical conditions, atleast in principle.

**A relative frequency distribution** shows the proportion of the total number of observations associated with each value or class of values and is related to a probability **distribution**, which is extensively used in **statistics**.

Let A be an event of interest, and assume that you have performed the same experiment n times so that n is the number of times A could have occurred. Further, let nA be the number of times that A did occur. Now, consider the relative frequency  $n_A/n$ . Then, in this method, we “attempt” to define P (A) as:  $P(A) = n_A/n$ . The above can only be viewed as an attempt because it is not physically feasible to repeat an experiment an infinite number of times. Another important issue with this definition is that two sets of n experiments will typically result in two different ratios. However, we expect the discrepancy to converge to 0 for large n. Hence, for large n, the ratio  $n_A/n$  may be taken as a reasonable approximation for P (A).

**Example 1:** Roll of a Dice  $S = \{1, 2, \dots, 6\}$  Probabilities: Roll the given dice 100 times (say) and suppose the number of times the outcome 1 is observed is 15. Thus,  $A = \{1\}$ ,  $n_A = 15$ , and  $n = 100$ . Therefore, we say that P(A) is approximately equal to  $15/100 = 0.15$ .

**Example 2:** Computer Sales A computer store tracks the daily sales of desktop computers in the past 30 days. The resulting data is:

Desktop Sold	No: of Days
0	1
1	2
2	10
3	12
4	5
5 or more	0

Therefore the approximate possibilities are:

Desktop Sold	No: of Days	Probability
0	1	$1/30 = 0.03$
1	2	$2/30 = 0.07$
2	10	$10/30 = 0.33$
3	12	$12/30 = 0.40$
4	5	$5/30 = 0.17$
5 or more	0	0

**NOTES**

Thus, for example, there is a 40% chance that the store will sell 3 desktops on any given day.

---

**6.3 Random Sampling :**

---

For the definition of probability to be accurate, it is necessary that the outcomes be obtained by a process called *random sampling*. A **random sample** requires that each individual in the population has an *equalchance* of being selected. A second requirement, necessary for many statistical formulas, states that if more than one individual is being selected, the probabilities must *stay constant* from one selection to the next. Adding this second requirement produces what is called *independent randomsampling*. The term *independent* refers to the fact that the probability of selecting any particular individual is independent of those individuals who have already been selected for the sample. For example, the probability that you will be selected is constant and does not change even when other individuals are selected before you are.

Each of the two requirements for random sampling has some interesting consequences. The first assures that there is no bias in the selection process. For a population with  $N$  individuals, each individual must have the same probability,  $p = 1 \div N$ , of being selected. This means, for example, that you would not get a random sample of people in your city by selecting names from a yacht-club membership list. Similarly, you would not get a random sample of college students by selecting individuals from your psychology classes. You also should note that the first requirement of random sampling prohibits you from applying the definition of probability to situations in which the possible outcomes are not equally likely. Consider, for example, the question of whether you will win a million dollars in the lottery tomorrow. There are only two possible alternatives.

1. You will win.
2. You will not win.

According to our simple definition, the probability of winning would be one out of two. However, the two alternatives are not equally likely, so the simple definition of probability does not apply. The second requirement also is more interesting than may be apparent at first glance. Consider, for example, the selection of  $n = 2$  cards from a complete deck. For the first draw, the probability of obtaining the jack of diamonds is  $p$  ( jack of diamonds) =  $1 \div 52$ . After selecting one card for the sample, you are ready to draw the second card. What is the probability of obtaining the jack of diamonds this time? Assuming that you still are holding the first card, there are two possibilities:

$$p(\text{jack of diamonds}) = \frac{1}{51} \text{ (if the first card was not the jack of diamonds)}$$

or

**NOTES**

$p(\text{jack of diamonds}) = 0$  ( if the first card was the jack of diamonds)

In either case, the probability is different from its value for the first draw.

This contradicts

the requirement for random sampling, which says that the probability must stay constant. To keep the probabilities from changing from one selection to the next, it is necessary to return each individual to the population before you make the next selection. This process is called *sampling with replacement*.

**Check your progress -1**

1. A survey of the students in a psychology class revealed that there were 19 females and 8 males. Of the 19 females, only 4 had no brothers or sisters, and 3 of the males were also the only child in the household. If a student is randomly selected from this class,

- a. What is the probability of obtaining a male?
- b. What is the probability of selecting a student who has at least one brother or sister?
- c. What is the probability of selecting a female who has no siblings?

2. A jar contains 10 red marbles and 30 blue marbles.

a. If you randomly select 1 marble from the jar, what is the probability of obtaining a red marble?

3. Which of these number cannot be a probability?

- a) 0.00001
- b) 0.5
- c) 0
- e) 1

---

## 6.4 Basic Rules of probability:

---

All definitions of probability must follow the same rules.

**6.4.1: Union/Addition:**

Let A and B be two events. Then,  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ . The subtraction of  $P(A \cap B)$  is necessary because A and B may “overlap.” If A and B are mutually exclusive, i.e.,  $A \cap B = \emptyset$ , then  $P(A \cup B) = P(A) + P(B)$ .

**Example: Roll of a Dice**

$$P(\text{even}) = 3/6 \text{ and } P(\text{odd}) = 3/6$$

$$P(\text{even and odd}) = P(\{2\}) = 1/6$$

$$P(\text{even or odd}) = 3/6 + 3/6 - 1/6 = 5/6$$

$$P(\{1\} \text{ or } \{6\}) = 1/6 + 1/6 - 0 = 2/6$$

**NOTES**

**Conditional Probability:**

Let A and B be two events. Then, the conditional probability of A given that B has occurred,  $P(A | B)$ , is defined as:  $P(A | B) = P(A \cap B) / P(B)$ .

(1) The reasoning behind this definition is that if B has occurred, then only the “portion” of A that is contained in B, i.e.,  $A \cap B$ , could occur; moreover, the original probability of  $A \cap B$  must be recalculated to reflect the fact that the “new” sample space is B

**Example: Pick a Card from a Deck**

Suppose a card is drawn randomly from a deck and found to be an Ace. What is the conditional probability for this card to be Spade Ace?

A = Spade Ace

B = an Ace

$A \cap B$  = Spade Ace

$P(A) = 1/52$

$P(B) = 4/52$ ; and  $P(A \cap B) = 1/52$

Hence,  $P(A | B) = 1/52$  divided by  $4/52 = 1/4$

Note that probability is defined as a proportion, or a part of the whole. This definition makes it possible to restate any probability problem as a proportion problem. For example, the probability problem “What is the probability of selecting a king from a deck of cards?” can be restated as “What proportion of the whole deck consists of kings?” In each case, the answer is  $4 \div 52$ , or “4 out of 52.” This translation from probability to proportion may seem trivial now, but it is a great aid when the probability problems become more complex

**Check your progress -2**

4. Find the probability of rolling a ‘3 with a dice.’

5. Draw a random card from a pack of cards. What is the probability that the card drawn is a face card?

**6.4.2 Multiplication** The multiplication rule is used to calculate the joint probability of two events. It is simply a rearrangement of the conditional probability formula. When we know that a particular event B has occurred, then instead of Sample space, we concentrate on B for calculating the probability of occurrence of event A given B

Formally,  $P(A \cap B) = P(A | B)P(B)$  ; or,  $P(A \cap B) = P(B | A)P(A)$  .

**NOTES****Example 1: Drawing a Spade Ace**

A = an Ace

B = a Spade

$A \cap B$  = the Spade Ace

$P(B) = 13/52$ ;  $P(A | B) = 1/13$  Hence,  $P(A \cap B) = P(A | B)P(B) = 1/13 \times 13/52 = 1/52$

**Example 2: Selecting Students**

A statistics course has seven male and three female students. The professor wants to select two students at random to help her conduct a research project. What is the probability that the two students chosen are female?

A = the first student selected is female

B = the second student selected is female

$A \cap B$  = both chosen students are female

$P(A) = 3/10$ ;  $P(B | A) = 2/9$

Hence,  $P(A \cap B) = P(B | A)P(A) = 2/9 \times 3/10 = 1/15$

**Multiplication Theorem for Independent Events**

The multiplication theorem on probability for dependent events can be extended for the independent events. From the theorem, we have,  $P(A \cap B) = P(A) P(B | A)$ . If the events A and B are independent, then,  $P(B | A) = P(B)$ . The above theorem reduces to

$$P(A \cap B) = P(A) P(B).$$

This shows that the probability that both of these occur simultaneously is the product of their respective probabilities.

Problem: A box contains 5 black, 7 red and 6 green balls. Three balls are drawn from this box one after the other without replacement. What is the probability that the three balls are

1. all black balls
2. of different colors
3. two black and one green ball.

Solution: The total number of the balls in the box is  $5 + 7 + 6 = 18$ . Let events

**NOTES**

- B: drawing black balls.
- R: drawing red balls.
- G: drawing green balls.

The balls are drawn without replacement. For the first draw, there are 18 balls to choose from. The number of balls gets lessened by 1 for the second draw i.e.,  $18 - 1 = 17$  and 16 for the third draw.

1. Probability that the three balls are all black =  $P(B_1) P(B_2 | B_1) P(B_3 | B_1 \cap B_2) = 5/18 \times 4/17 \times 3/16 = 5/408$ .
2. The probability that the three balls are all different in color =  $P(B_1) P(R_1 | B_1) P(G_1 | B_1 \cap R_1) = 5/18 \times 7/17 \times 6/16 = 35/816$ .
3. Probability that two black and one green balls are drawn =  $P(B_1) P(B_2 | B_1) P(G_1 | B_1 \cap B_2) = 5/18 \times 4/17 \times 6/16 = 5/204$ .

It does not matter which colour ball is drawn first.

**6.5 Let us Sum up:**

**Probability** is a branch of mathematics that deals with the occurrence of a random event. This basic theory is also used in the probability distribution. The **Addition rule** states that when two events, A and B, are mutually exclusive, the probability that A or B will occur is the sum of the probability of each event, minus the probability of the overlap.  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ . The **Multiplication Rule** of probability states that, If events A and B come from the same sample space, the probability that both A and B occur is equal to the probability the event A occurs times the probability that B occurs, given that A has occurred.

**6.6 Unit – End exercises:**

1. State the addition and multiplication rule of probability.
2. Tickets numbered 1 to 20 are mixed up and then a ticket is drawn at random. What is the probability that the ticket drawn has a number which is a multiple of 3 or 5?
3. A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?
4. State why it is important to sample with replacement.

**6.7 Answer to check your progress**

1. a.  $p = 8/27$

**NOTES**

b.  $p = 20/27$

c.  $p = 4/27$

7  $p = 10/40 = 0.25$

8 a)-0.00001 (A probability is always greater than or equal to 0 and less than or equal to 1)

9 Sample Space = {1, 2, 3, 4, 5, 6}

Number of favourable event = 1

Total number of outcomes = 6

Thus, Probability,  $P = 1/6$

5. A standard deck has 52 cards.

Total number of outcomes = 52

Number of favourable events =  $4 \times 3 = 12$  (considered Jack, Queen and King only)

Probability,  $P = \text{Number of Favourable Outcome} / \text{Total Number of Outcomes}$

$$= 12/52 = 3/13.$$

---

### 6.8 Suggested Readings:

---

1. Gravetter, F.J., & Wallnau, L.B. (2013). Statistics for behavioural sciences (9<sup>th</sup> Ed.). Cengage Learning India Private Ltd.
2. <https://www.toppr.com/guides/maths/probability/multiplication-theorem-on-probability/>

---

## UNIT 7: NORMAL DISTRIBUTION

---

**NOTES****7.1 Meaning and Importance****7.2 Characteristics of Normal Probability Curve****7.8.1 Probability and Normal Distribution****7.3 Deviations of Normal Probability Curve****7.3.1 Skewness****i. Kurtosis****7.4 Applications of the Normal Probability Curve.****7.5 Let us Sum up****7.6 Unit – End exercises****7.7 Answer to check your progress****7.8 7.8 Suggested Readings**


---

### **7.1 Meaning and Importance of Normal Probability Curve:**

---

The literal meaning of the term normal is average. The data from a certain coin or a dice throwing experiment involving a chance success or probability, if plotted on a graph paper, give a frequency curve which closely resembles the normal curve. It is because of this reason and because of its origin from a game of chance the normal curve is called normal probability curve. The normal curve takes into account the law which states that the greater a deviation from the mean or average value in a series the less frequently it occurs. This is satisfactorily used for describing many distributions which arise in the fields of education, psychology and sociology. The normal probability curve is a mathematical distribution.

Researchers often compare the actual distribution of variables they are studying to the normal curve. They do not expect the distributions of their variables to match the normal curve (perfectly) but often check whether the variables approximately follow the normal curve. It is also called a Gaussian distribution, however its discovery is attributed to Abraham De Moivre.

---

### **7.2 Characteristics of a normal curve:**

---

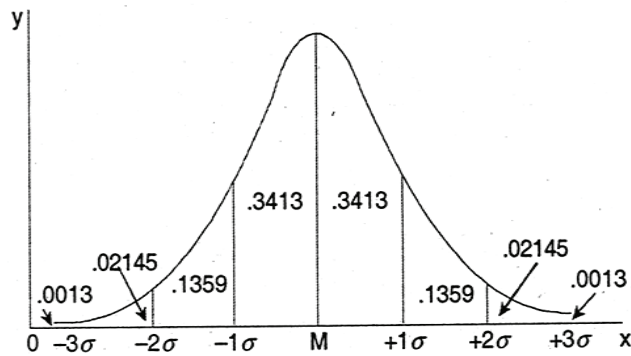
1. For this curve, mean, median and mode are the same.
2. The curve is perfectly symmetrical. In other words it is not skewed. The value of measure of skewness computed for this curve is zero.
3. The normal curve serves as a model for describing the peakedness or flatness of a curve through the measure of kurtosis. For the normal curve the value of kurtosis is 0.263. If for a distribution the value of kurtosis is more than 0.263, the distribution is said to be



- more flat at the top than the normal curve. But in case the value of kurtosis is less than 0.263, the distribution is said to be more peaked than the normal.
4. The curve is asymptotic. It approaches but never touches the baseline at the extremes because of the possibility of locating in the population a case which score still higher than our highest score or lower than our lowest score. Therefore, theoretically it extends from minus infinity to plus infinity.
  5. As the curve does not touch the baseline, the mean is used as a starting point for working with the normal curve.
  6. The curve has its maximum height or ordinate at the starting point i.e, the mean of the distribution. In a unit normal curve, the value of this ordinate is equal to .3989.
  7. To find the deviation from the point of departure standard deviation of the distribution is used as a unit of measurement.
  8. The curve extends on both sides  $-3\sigma$  distance on the left to  $+3\sigma$  distance on the right.
  9. The points of inflection of the curve occur at  $\pm 1$  standard deviation unit above or below the mean. Thus the curve changes from convex to concave in relation to the horizontal axis at these points.
  10. The total area under the curve extending from  $-3\sigma$  to  $+3\sigma$  is taken arbitrarily to be 10,000 because of the greater ease in the computation of the fractional parts of the total area found for the mean and the ordinates erected at various distances from the mean.
  11. We may find that 3413 cases out of 10,000 or 34.13% of the entire area of the curve lie between the mean and  $+1\sigma$  on the baseline of the normal curve. Similarly, another 34.13% cases lie between the mean and  $-1\sigma$  on the baseline. Consequently, 68.26% of the area of the curve falls within the limits  $\pm 1$  standard deviation unit from the mean. Going further it may be found that 95.44% cases lie from  $-2\sigma$  to  $+2\sigma$  and 99.74% cases the from  $-3\sigma$  to  $+3\sigma$ . Consequently only 26 cases in 10,000 (10000 - 9974) should be expected to the beyond the range  $\pm 3\sigma$  in a large sample. Since it's a mathematical curve these proportions are worked out based on the principles of probability.

**NOTES**

**NOTES**



**Figure 8.8** Normal curve showing areas at different distances from the mean.

12. In this curve, the limits of the distance  $1.96\sigma$  includes 95% and the limit  $+2.58\sigma$  includes 99% of the total area of curve, 5% and 1% of the area, respectively falling beyond these limits.

**7.2.1. Probabilty and normal distribution:**

As you have seen in the figure above the normal distribution is symmetrical, with the highest frequency in the middle and frequencies tapering off as you move toward either extreme. The exact shape can also be described by the proportions of area contained in each section of the distribution. Statisticians often identify sections of a normal distribution by using *z*-scores, which you see on the baseline of the curve.

The ***z* score** is a standard score that is used to identify and describe the location of each scores in a distribution. The original unchanged scores known as raw scores by itself does not provide much information about its position in the distribution, hence the transformation of *X* values(the raw scores) into *z* scores tell exactly where the original scores are located. The sign of the *z* score tells whether the score is located above (+) or below (-) the mean and the number tells the distance between the score and the mean in terms of number of standard deviation. In other words, *z*-scores measure positions in a distribution in terms of standard deviations from the mean. (Thus,  $z = +1$  is 1 standard deviation above the mean,  $z = +2$  is 2 standard deviations above the mean, and so on.). One major advantage of *z* score is that, even when different distribution have different means and standard deviations, the location identified by the *z* scores are the same for *all* distributions.

The normal curve table is worked by statisticians and it provides the percentage of scores between the mean and any other Z score. This provides psychologists more accurate information in many research and applied situation. The first column in the table tells the z score. The second column, labeled “% Mean to z” - gives the percentage of scores below the mean and that z score. The third column, labeled “z in Tail” gives the percentage of scores in the tail for that z score. The table lists only positive z scores. This is because the normal curve is perfectly symmetrical.

**NOTES****Check your Progress – 1**

1. The term normal in the normal distribution means \_\_\_\_\_
2. The base line of the normal curve is divided into \_\_\_\_\_ units.
3. What does the term “asymptotic” mean?
4. What does the normal curve table provide?
5. Describe the location of the distribution for each of the following z scores?
  - a)  $z = -1.50$  b)  $z = 0.25$
6. For a population with  $\mu = 30$  and  $\sigma = 8$ , Find the z score of a  $X = 32$

---

**7.3. Deviation from the normal curve:**


---

**7.3.1 Skewness:** The normal curve is symmetrical i.e the left half of the curve is equal to the right half of it. Also for this curve, mean, median and mode are the same.

In many distribution which deviate from normal, the value of mean, median mode are different and there is no symmetry. Lack of symmetry is referred as skewness. Such distributions are said to be skewed, being inclined more towards the left or the right to the centre of the curve.

The distribution is said to be skewed negatively when there are many individual in a group with their scores higher than the average score of the group. Similarly, the distributions are said to be skewed positively when there are more individuals in a group who score less than the average score for their group.

Skewness in any given distribution may be computed by the following formula:

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{SD}$$

$$Sk = \frac{3(M - Md)}{SD}$$

**NOTES**

In case when the percentiles are known, the value of skewness may be computed from the following formula:

$$S_k = \frac{P_{90} + P_{10} - P_{50}}{2}$$

**7.3.2 Kurtosis:**

When there are very few individuals whose scores are near to the average score for their group (too few cases in the central area of the curve) the curve representing such a distribution becomes 'flattened in the middle. On the other hand, when there are too many cases in central area, the distribution curve becomes too 'peaked in comparison to normal'. Both these characteristics of being flat on peaked, are used to describe the term Kurtosis.

Kurtosis is usually of three types.

**Platykurtic:** A frequency distribution is said to be platykurtic, when it is flatter than the normal.

**Leptokurtic:** A frequency distribution is said to be leptokurtic, when it is more peaked than the normal.

**Mesokurtic:** A frequency distribution is said to be mesokurtic, when it almost resemble the normal curve (neither flattened nor too peaked)

The value of kurtosis for a given curve may be computed through the following formula.

$$Kurtosis = \frac{\text{Quantile deviation}}{90th\ percentile - 10th\ percentile}$$

In the case of a normal curve, this value is equal to 0.263, the distribution is said to be platykurtic, if less than 0.263, the distribution is leptokurtic.

**Check your Progress- 2**

7. When do we say that the normal curve is skewed?
8. When do we observe a leptokurtic curve?

**7.4 Applications of the Normal Curve**

Normal curve has wide significance. Some of the main applications are:-

### 1. **Used as a Model**

Normal curve represents a model distribution. It can be used as a model to

- (a) Compare various distributions with it i.e. to say, whether the distribution is normal or not and if not, in what way it deviates from the normal.
- (b) Compare two or more distribution in terms of overlapping.
- (c) Evaluate students' performance from this score.

### 2. **Computing percentiles and percentile Ranks**

### 3. **Ability grouping**

A group of individuals may be conveniently grouped into certain categories A, B, C, D, E in terms of some that with the help of a normal curve.

### 4. **Transforming and combining qualitative data**

Under the assumptions of normality of the distributed variable, the sets of qualitative data such as ratings, letter grades and categorical ranks on a scale may be conveniently transformed and combined to provide an average rating for each individual.

### 5. **Converting Raw score into comparable Standard normalized scores**

With the help of the normal curve, we can convert the raw scores belonging to different tests into standardized normalized scores like the  $\sigma$  scores and T scores. For converting a given raw scores into a Z score, we subtract the mean of the scores of the distribution from the respective raw scores and divide it by the standard deviation of the distribution.

$$Z = \frac{X - M}{\sigma}$$

### 6. **Determining relative difficulty of test items**

Normal curve provides the simplest rational method of scaling test items for difficulty and therefore, may be conveniently employed for determining the relative difficulty of the test questions, problems and other test items.

## **NOTES**

**NOTES****Some illustrations of the applications of the normal curve****Case I:** Comparing scores on two different tests

A student obtains 80 marks in Maths and 50 in English. If the mean and SD for the scores in Maths are 70 and 20 and for the scores in English are 30 and 10 find out in which subject, Maths or English he did better.

**Solution:** Direct comparisons of marks obtained on both subjects are not possible as they do not belong to the same scale of measurement. Therefore to compare convert raw scores into z scores.

Raw scores in Maths ( $X_1$ ) = 80  $M_1 = 70$  and  $\sigma_1 = 20$

$$Z \text{ score in Maths} = \frac{X_1 - M_1}{\sigma_1} = \frac{80 - 70}{20} = 0.5$$

Raw scores in English  $X_2 = 50$ ,  $M_2 = 30$ ,  $\sigma_2 = 10$

Therefore

$$Z \text{ score in English} = \frac{X_2 - M_2}{\sigma_2} = \frac{50 - 30}{10} = 2$$

We can thus conclude that the student did better in English than in Maths.

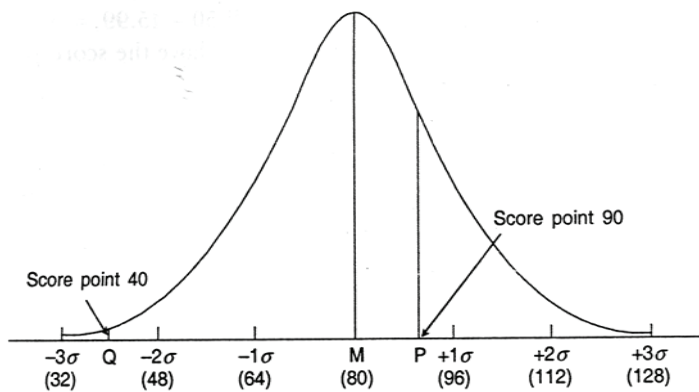
**Case II:** To determine percentile rank ie percentages of cases lying before a given score point.

Given a normal distribution  $N = 1000$ , mean 80 and  $SD = 16$  find the percentile equal of the individual scoring 90.

Solution (i) the percentile rank is essentially a rank or position of an individual (on a scale of 100) decided on the best of the individuals scores. In other words, we have to determine the percentage of cases lying below the score point 90. The first step is to convert the raw scores into z score.

$$z = \frac{X - M}{\sigma}$$

$$\frac{90 - 80}{10} = 0.625\sigma$$

**NOTES**

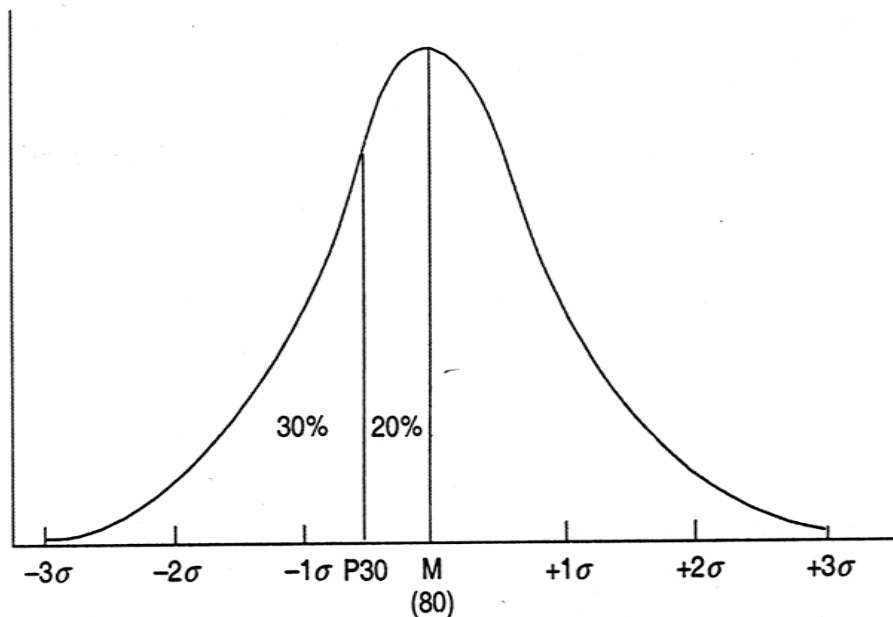
**Figure 8.12** Showing percentile rank of 90 and cases lying below 40.

From the table of normal curve we can say that 23.41% of cases lie between  $M \pm 0.625\sigma$  distance. But 50% of cases lie up to the mean. Therefore, it may be concluded that there are  $50 + 23.41 = 73.41\%$  of the individuals whose scores lie below the score point 90 or we may say the percentiles rank of the individual scoring 90 is 73.

**Case III:** To determine percentile points or limits in terms of score which include the lowest given percentage of cases.

Given a normal distribution of  $N=1000$ ,  $M = 80$  and  $\sigma = 16$ , determine the percentile  $P_{30}$ .

**NOTES**



In determining the percentile  $P_{30}$  we have to look for a score point on the scale of measurement below which 30% of the case lie. Such a score will have 20% of the cases lying on the left side of the mean. From the table of the normal curve, we find out the corresponding  $\sigma$  distances from the mean for 20% of the cases. We find that distance to be  $0.525\sigma$ . Therefore the required score here will be

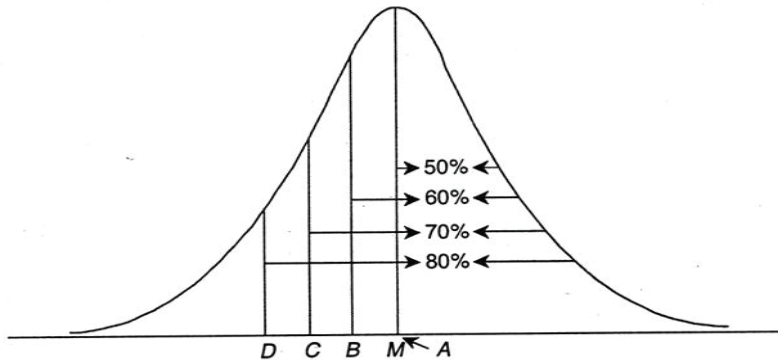
$$M - 0.525\sigma \text{ or}$$

$$80 - 0.525 \times 16 = 71.6 \text{ or } 72$$

**Case IV:** To determine the relative difficulty value of the items

Four problems A, B, C and D have been solved by 50%, 60%, 70% & 80% respectively of a large group. Compare the difference in difficulty between A & B with the difference between C & D.



**NOTES**

In the case of a large group, the assumption about normal distribution of the ability of a group in terms of the achievement on a test holds good. The percentage of students who are able to solve a particular problem are counted from extreme right. Therefore, while starting on the base line of the curve from the extreme right, up to the point M, we may cover 50% of the case who can solve the problem. For the rest 60, 70 and 80%. We have to proceed on the left side of the base line of the curve.

For problem A, we see that it has been solved by 50% of the group. It is also implied that 50% of the group has not been able to solve it. Therefore, we may say that it was an average problem having of zero difficulty value.

In the case of problem B, we see that it has been solve by 60% of the group. It has been a simple problem in comparison to A as 10% more individuals in the group are able to solve it. For determining the difficulty value of this problem, we find the  $\sigma$  distance from the mean of these 10% of individuals. From the normal curve table we see that 10% (1000 out of 10,000) cases fall at a sigma distance of 0.253 from the mean after interpolation. Therefore the difficulty value of problem B will be taken as  $-0.253\sigma$ .

Similarly, we may determine the difficulty value of problem C passed by 70% of the group (20% more individuals of the group than the average) and problem D by 80% (50% more individuals than the average).

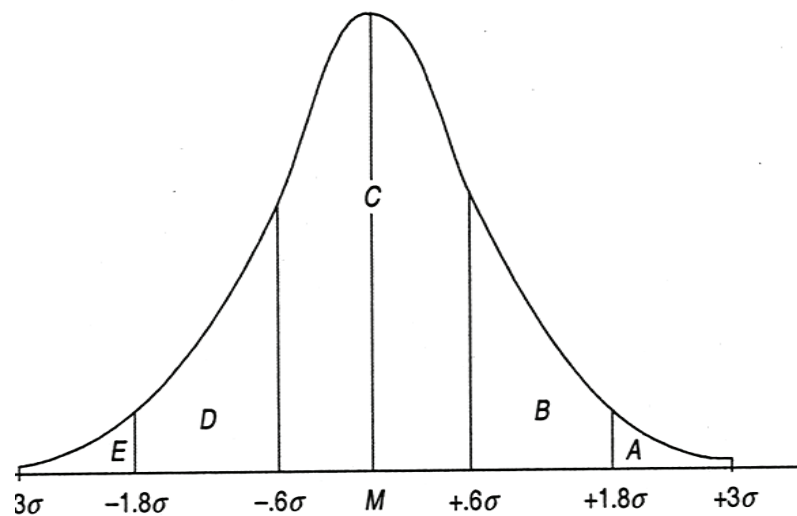
**NOTES**

From the table we know the sigma distance that corresponds to 20% & 30% are  $0.525\sigma$  &  $-0.84\sigma$ . Hence these are the difficulty value of these items.

**Case V:** To divide a group into categories according to an ability or trait assumed to be normally distributed.

**Example:** There is a group of 200 students who has to be classified into five categories. A, B, C, D & E, according to ability, the range of ability being equal in each category. If the trait counted under ability is normally distributed tell how many students should be placed in each category A, B, C, D & E.

Solution:



As the trait under measurement is normally distributed, the whole group is divided into five equal categories as shown in the above figure. It shows that the base line of the curve considered to extend from  $-3\sigma$  to  $+3\sigma$ , i.e. over a range of  $6\sigma$ , may be divided into five equal parts. It gives  $1.2\sigma$  as the portion of the base line to be allotted to each category. Here group A covers the upper  $1.2\sigma$  segment (falling between  $1.8\sigma$  and  $3\sigma$ ), group B the next  $1.2\sigma$ , group C lies  $0.6\sigma$  to the right and  $0.6\sigma$  to the left of the mean. Groups D and E cover the same relative position on the left side.

of the mean as covered by the groups B and A on the right side of the mean.

After the area of the curve covered by the respective categories is demarcated, the next step is to find out from the normal curve the percentage of cases lying within each of these area.

For example area A extends from  $1.8\sigma$  to  $3\sigma$ . The percentage of cases lying between mean and  $3\sigma$  is 49.86% and mean and  $1.8\sigma$  is 46.41%. The difference ( $49.86\% - 46.41\% = 3.45\%$ ) will yield the required percentage of the whole group belonging to category A.

Similarly, group B will cover the cases lying between  $0.6\sigma$  and  $1.8\sigma$ . We find from the table that the percentage of the whole group lying between mean and  $1.8\sigma$  is 46.41 and between M &  $0.6\sigma$  is 22.57 therefore group B may be said to comprise of  $46.41\% - 22.57\% = 23.84\%$  of the entire group.

Group C extends from  $-0.6\sigma$  to  $0.6\sigma$  on both sides of the mean. The normal curve table tells us that 22.57% cases to between M &  $0.6\sigma$  and a similar percentage of cases i.e, 22.57, lies between  $M_2 - 0.6\sigma$ . Therefore, group C may be said to compute  $22.57 \times 2 = 45.14$  or 45% of the entire group.

Group D & E are identical to groups B & A. Therefore, may be found to consist of the same percentage of cases as covered in groups B & A respectively.

---

### 7.5 Let us Sum up:

---

The literal meaning of the term normal is average. Though normal distribution is a mathematical distribution and not the law of nature, we do find it as a convenient model to explain many physical and psychological attributes such as intelligence, ability, weight, aptitude etc. When we study each of the above mentioned attributes we do find that there are a number of people who deviate from average but only very few who markedly differ from the average. If we graphically plot the scores that were obtained by measuring these attributes, we are likely to get a typical curve resembling the vertical cross section of the bell. This bell shaped curve is called as the normal curve and has several interesting properties. Probability forms a direct link between samples and the populations from which they come. This link is the foundation for the inferential statistics in future chapters. Deviations from the normal curve are measured in terms of skewness and Kurtosis. The normal curve as seen in the text above has wide significance and applicability.

### NOTES

**NOTES****7.6 Unit – End exercises**

1. Given the following data regarding two distributions

	Mathematics	Physics
M	60	33
S.D	8	9
Achievement scores of students	70	67

Find out whether the student did better in Maths or Physics

4. Bring out the utility of the normal curve in the field of psychology.
5. Given a normal distribution with the mean of 50 and standard deviation of 15
  - a) What percentage of the cases lie between the scores of 47 and 60?
  - b) What percentage of cases will lie between 40 and 47?
  - c) What percentage of group is expected to have scores greater than 6?

4. What is a z score?

**7.7 Answer to check your progress**

1. Average

2.  $6\sigma$  units.

3. The term “asymptotic” in the normal curve refers to the fact that the curve never touches the baseline due to the possibility of locating in a population a case which score either higher than the highest scores or lower than the lowest score.

4. It shows the exact percentages between a score and the mean, which is worked out by statisticians based probability

5. a) Below the mean  $1\frac{1}{2}$  standard deviation
- b) Above the mean  $\frac{1}{4}$  standard deviation

6.  $z = 0.25$

7. We state that the normal curve is skewed when the tail on one side is longer than the others. Such asymmetry is observed when there are too many individuals in either the right or left of the distribution

8. We observe a leptokurtic curve when there are too many cases in central area, thus making the distribution curve too ‘peaked in comparison to normal’

**7.8 Suggested Readings:**

1. Mangal S.K. (1987). *Statistics in Psychology and Education*. New Delhi: PHI Learning Private Ltd.
2. Haslam, S.M., & McGarty, C. (2003). *Research Methods and statistics in Psychology*. Sage Publications Ltd.

---

## UNIT 8: BINOMIAL DISTRIBUTIONS

---

- 8.1 Introduction
- 8.2 The Binomial Distribution
- 8.3 The Binomial Test
  - 8.3.1 Hypotheses for the binomial test
  - 8.3.2 The data for the binomial test
  - 8.3.3 The test statistic for the binomial test
- 8.4 Let us sum up
- 8.5 Unit End Exercises
- 8.6 Answers to check your progress

### 8.7 Suggested Readings

### NOTES

---

#### 8.1 Introduction

---

When a variable is measured on a scale consisting of exactly two categories, the resulting data are called binomial. The term *binomial* can be loosely translated as "two names," referring to the two categories in the measurement scale.

Binomial data can occur when a variable naturally exists with only two categories. For example, people can be classified as male or female, and a coin toss results in either heads or tails. It also is common for a researcher to simplify data by collapsing the scores into two categories. For example, a psychologist may use personality scores to classify people as either high or low in aggression.

In binomial situations, the researcher often knows the probabilities associated with each of the two categories. With a balanced coin, for example,  $p(\text{heads}) = p(\text{tails}) = \frac{1}{2}$ . The question of interest is the number of times each category occurs in a series of trials or in a sample of individuals. For example: What is the probability of obtaining 15 heads in 20 tosses of a balanced coin? What is the probability of obtaining more than 40 introverts in a sampling of 50 college freshmen?

As we shall see, the normal distribution serves as an excellent model for computing probabilities with binomial data.

## 8.2 The Binomial Distribution

### NOTES

To answer probability questions about binomial data, we must examine the binomial distribution. To define and describe this distribution, we first introduce some notation.

1. The two categories are identified as  $A$  and  $B$ .
2. The probabilities (or proportions) associated with each category are identified as

$$p = p(A) = \text{the probability of } A$$

$$q = p(B) = \text{the probability of } B$$

Notice that  $p + q = 1.00$  because  $A$  and  $B$  are the only two possible outcomes,

3. The number of individuals or observations in the sample is identified by  $n$ .

4. The variable  $X$  refers to the number of times category  $A$  occurs in the sample.

Notice that  $X$  can have any value from 0 (none of the sample is in category  $A$ ) to  $n$  (all of the sample is in category  $A$ ).

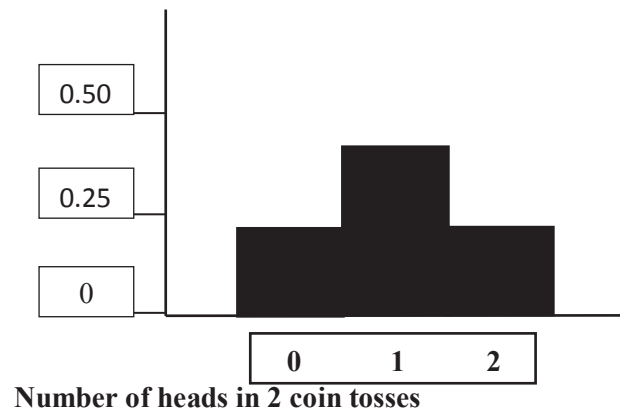
Using the notation presented here, the **binomial distribution** shows the probability associated with each value of  $X$  from  $X = 0$  to  $X = n$ .

A simple example of a binomial distribution is presented next.

Figure below shows the binomial distribution for the number of heads obtained in 2 tosses of a balanced coin. This distribution shows that it is possible to obtain as many as 2 heads or as few as 0 heads in 2 tosses. The most likely outcome (highest probability) is to obtain exactly 1 head in 2 tosses. For this example, the event we are considering is a coin toss. There are two possible outcomes, heads and tails. We assume the coin is balanced, so

$$P = p(\text{heads}) = \frac{1}{2}$$

**Figure The binomial distribution showing the probability for the number of head sin 2 tosses of a balanced coin**



$$q = p(\text{tails}) = \frac{1}{2}$$

We are looking at a sample of  $n = 2$  tosses, and the variable of interest is

$X$  = the number of heads;

To construct the binomial distribution, we look at all of the possible outcomes from tossing a coin 2 times. The complete set of 4 outcomes is listed in the following table.

1st Toss	2nd Toss	
Heads	Heads	(Both heads)
Heads	Tails	(Each sequence has exactly 1 head)
Tails	Heads	
Tails	Tails	(No heads)

Notice that there are 4 possible outcomes when you toss a coin 2 times. Only 1 of the 4 outcomes has 2 heads, so the probability of obtaining 2 heads is  $p = \frac{1}{4}$ . Similarly, 2 of the 4 outcomes have exactly 1 head, so the probability of 1 head is  $p = \frac{2}{4} = \frac{1}{2}$ . Finally, the probability of no heads ( $X = 0$ ) is  $p = \frac{1}{4}$ . These are the probabilities shown in Figure

## NOTES

**NOTES**

Note that this binomial distribution can be used to answer probability questions. For example, what is the probability of obtaining at least 1 head in 2 tosses? According to the distribution, the answer is  $\frac{3}{4}$

**Check your Progress -1**

1. What is a binomial data? When a variable is measured on a scale consisting of exactly two categories, the resulting data are called binomial
2. What does the binominal distribution give? The binomial distribution gives the probability for each value of  $X$ , where  $X$  equals the number of occurrences of category  $A$  in a sample of  $n$  events.

---

**8.3 The Binomial Test:**


---

In this chapter, we examine the statistical process of using binomial data for testing hypotheses about the values of  $p$  and  $q$  for the population. This type of hypothesis test is called a *binomial test*. A **binomial test** uses sample data to evaluate hypotheses about the values of  $p$  and  $q$  for a population consisting of binomial data.

Consider the following two situations:

1. In a sample of  $n = 34$  colour-blind students, 30 are male, and only 4 are female. Does this sample indicate that colour blindness is significantly more common for males in the general population?
2. In 2005, only 10% of American families had incomes below the poverty level. This year, in a sample of 100 families, 19 were below the poverty level. Does this sample indicate that there has been a significant change in the population proportions?

Notice that both of these examples have binomial data (exactly two categories). Although the data are relatively simple, we are asking a statistical question about significance that is appropriate



for a hypothesis test: Do the sample data provide sufficient evidence to make a conclusion about the population?

---

### 8.3.1 HYPOTHESES FOR THE BINOMIAL TEST I

---

In the binomial test, the null hypothesis specifies exact values for the population proportions  $p$  and  $q$ . Theoretically, you could choose any proportions for  $H_0$ , but usually there is a clear reason for the values that are selected. The null hypothesis typically falls into one of the following two categories:

1. **Just Chance.** Often the null hypothesis states that the two outcomes,  $A$  and  $B$ , occur in the population with the proportions that would be predicted simply by chance. If you were tossing a coin, for example, the null hypothesis might specify  $p(\text{heads}) = \frac{1}{2}$  and  $p(\text{tails}) = \frac{1}{2}$ . Notice that this hypothesis states the usual, chance proportions for a balanced coin. Also notice that it is not necessary to specify both proportions. Once the value of  $p$  is identified, the value of  $q$  is determined by  $1 - p$ . For the coin toss example, the null hypothesis<sup>1</sup> would simply state

$$H_0: p = p(\text{heads}) = \frac{1}{2} \quad (\text{The coin is balanced.})$$

Similarly, if you were selecting cards from a deck and trying to predict the suit on each draw, the probability of predicting correctly would be  $p = \frac{1}{4}$  for any given trial. (With four suits, you have a 1-out-of-4 chance of guessing correctly.) In this case, the null hypothesis would state

$$H_0: P = p(\text{guessing correctly}) = \frac{1}{4} \quad (\text{The outcome is simply the result of chance.})$$

In each case, the null hypothesis states that there is nothing unusual about the proportions of the population; that is, the outcomes are occurring by chance.

### NOTES

**NOTES**

2. **No Change or No Difference:** Often you may know the proportions for one population and want to determine whether the same proportions apply to a different population. In this case, the null hypothesis would simply specify that there is no difference between the two populations. Suppose that national statistics indicate that 1 out of 12 drivers will be involved in a traffic accident during the next year. Does this same proportion apply to 16-year-olds who are driving for the first time? According to the null hypothesis,

$H_0$ : For 16-year-olds,  $p = p(\text{accident}) = \frac{1}{12}$  (Not different from the general population)

Similarly, suppose that last year, 30% of the freshman class failed the college writing test. This year, the college is requiring all freshmen to take a writing course. Will the course have any effect on the number who fail the test? According to the null hypothesis,

$H_0$ : For this year,  $p = p(\text{fail}) = 30\%$  (Not different from last year's class)

**8.3.2 THE DATA FOR THE BINOMIAL TEST**

For the binomial test, a sample of  $n$  individuals is obtained and you simply count how many are classified in category  $A$  and how many are classified in category  $B$ . We focus attention on category  $A$  and use the symbol  $X$  to stand for the number of individuals classified in category  $A$ .  $X$  can have any value from 0 to  $n$  and that each value of  $X$  has a specific probability. The distribution of probabilities for each value of  $X$  is called the *binomial distribution*.

**8.3.3 THE TEST STATISTIC FOR THE BINOMIAL TEST**

When the values  $pn$  and  $qn$  are both equal to or greater than 10, the binomial distribution approximates a normal distribution. This fact is important because it allows us to compute z-scores and use the unit normal table to answer probability questions about binomial

events. In particular, when  $pn$  and  $qn$  are both at least 10, the binomial distribution has the following properties:

1. The shape of the distribution is approximately normal.
2. The mean of the distribution is  $\mu = pn$ .
3. The standard deviation of the distribution is

$$\sigma = \sqrt{npq}$$

With these parameters in mind, it is possible to compute a z-score corresponding to each value of  $X$  in the binomial distribution.

$$z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}}$$

This is the basic z-score formula that is used for the binomial test. However, the formula can be modified slightly to make it more compatible with the logic of the binomial hypothesis test. The modification consists of dividing both the numerator and the denominator of the z-score by  $n$ . (You should realize that dividing both the numerator and the denominator by the same value does not change the value of the z-score.) The resulting equation is

$$z = \frac{X/n - p}{\sqrt{npq}}$$

The logic underlying the binomial test is exactly the same as with the original z-score hypothesis. The hypothesis test involves comparing the sample data with the hypothesis. If the data are consistent with the hypothesis, then we conclude that the hypothesis is reasonable. But if there is a big discrepancy between the data and the hypothesis, then we reject the hypothesis. The value of the standard error provides a benchmark for determining whether the discrepancy between the data and the hypothesis is more than would be expected by chance. The alpha level for the test

## NOTES

provides a criterion for deciding whether the discrepancy is significant. The hypothesis-testing procedure is demonstrated in the following section.

## NOTES

STEP 1 *State the hypotheses.* In the binomial test, the null hypothesis specifies values for the population proportions  $p$  and  $q$ . Typically,  $H_0$  specifies a value only for  $p$ , the proportion associated with category  $A$ . The value of  $q$  is directly determined from  $p$  by the relationship  $q = 1 - p$ . Finally, you should realize that the hypothesis, as always, addresses the probabilities or proportions for the *population*. Although we use a sample to test the hypothesis, the hypothesis itself always concerns a population.

STEP 2 *Locate the critical region.* When both values for  $pn$  and  $qn$  are greater than or equal to 10, then the scores form an approximately normal distribution. Thus, the unit normal table can be used to find the boundaries for the critical region. With  $\alpha = .05$ , for example, you may recall that the critical region is defined as  $z$ -score values greater than  $+1.96$  or less than  $-1.96$ .

STEP 3 *Compute the test statistic (z-score).* At this time, you obtain a sample of  $n$  individuals (or events) and count the number of times category  $A$  occurs in the sample. The number of occurrences of  $A$  in the sample is the  $X$  value for Equation given below. Because the two  $z$ -score equations are equivalent, you may use either one for the hypothesis test.

STEP 4 *Make a decision.* If the  $z$ -score for the sample data is in the critical region, then you reject  $H_0$  and conclude that the discrepancy between the sample proportions and the hypothesized population proportions is significantly greater than chance. That is, the data are not consistent with the null hypothesis, so  $H_0$  must be wrong. On the other hand, if the  $z$ -score is not in the critical region, then you fail to reject  $H_0$ . The following example demonstrates a complete binomial test.

**EXAMPLE:** A *visual cliff* experiment was designed to examine depth perception in infants. An infant is placed on a wide board that appears to have a deep drop on one side and a relatively shallow drop on the other. An infant who is able to perceive depth should avoid the deep side and move toward the shallow side. Without depth perception, the infant should show no preference between the two sides. Of the 27 infants in the experiment, 24 stayed exclusively on the shallow side and only 3 moved onto the deep side. The purpose of the hypothesis test is to determine whether these data demonstrate that infants have a significant preference for the shallow side.

This is a binomial hypothesis-testing situation? The two categories are

$A$  = move onto the deep side

$B$  = move onto the shallow side

**STEP1** The null hypothesis states that, for the general population of infants, there is no preference between the deep and the shallow sides; the direction of movement is determined by chance. In symbols,

$$H_0: p = p(\text{deep side}) = \frac{1}{2} \left( \text{and } q = \frac{1}{2} \right)$$

$$H^1: p \neq \frac{1}{2} (\text{There is a preference.})$$

We use  $\alpha = .05$

**STEP2** With a sample of  $n = 27$ ,  $pn = 13.5$  and  $qn = 13.5$ . Both values are greater than 10, so the distribution of  $z$  - scores is approximately normal. With  $\alpha = .05$ , the critical region is determined by boundaries of  $z = \pm 1.96$ .

**STEP3** For this experiment, *the data consist of*  $X = 3$  out of  $n = 27$ . Using Equation given below, these data produce a  $z$ -score value of

## NOTES

$$z = \frac{X - np}{\sqrt{npq}} = \frac{3 - 13.5}{\sqrt{27 \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{-10.5}{2.60} = -4.04$$

**NOTES**

To use the other Equation , you first compute the sample proportion,  $X/n = 3/27 = 0.11$ . The  $z$  - score is then

$$z = \frac{X/n - p}{\sqrt{pq/n}} = \frac{0.111 - 0.5}{\sqrt{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)/27}} = \frac{-0.389}{0.096} = -4.05$$

Within rounding error, the two equations produce the same result.

**STEP 4** Because the data are in the critical region, our decision is to reject  $H_0$ . These data do provide sufficient evidence to conclude that there is a significant preference for the shallow side. Gibson and Walk (1960) interpreted these data as convincing evidence that depth perception is innate.

**Check your Progress -2**

3. For the research study using a visual cliff given above State the null hypothesis for this study in words and as a probability value ( $p$ ) that an infant will crawl off the deep side.
4. If the visual cliff study had used a sample of  $n = 15$  infants, would it be appropriate to use the normal approximation to the binomial distribution? Explain why or why not.
5. For a binomial test, the null hypothesis always states that  $p = 1/2$ . (True or false?)

**8.4 Let us sum up:**

The binomial test is used with dichotomous data—that is, when each individual in the sample is classified in one of two categories. The two categories are identified as  $A$  and  $B$ . The binomial test uses sample data to test hypotheses about the binomial proportions,  $p$  and  $q$ , for a population. The null hypothesis specifies  $p$  and  $q$ , and the binomial distribution (or the normal approximation) is used to determine the critical region.

**8.5 Unit End Exercises**

1. To investigate the phenomenon of “home team advantage,” a researcher recorded the outcomes from 64 college football games on one Saturday in October. Of the 64 games, 42 were won by home teams. Does this result provide enough evidence to conclude

that home teams win significantly more than would be expected by chance? Use a two-tailed test with  $\alpha = .05$ .

2. A researcher would like to determine whether people really can tell the difference between bottled water and tap water. Participants are asked to taste two unlabelled glasses of water, one bottled and one tap, and identify the one they thought tasted better. Out of 40 people in the sample, 28 picked the bottled water. Was the bottled water selected significantly more often than would be expected by chance? Use a two-tailed test with  $\alpha = .05$ .

## NOTES

---

### 8.6 Answers to check your progress

---

1. When a variable is measured on a scale consisting of exactly two categories, the resulting data are called binomial
2. The binomial distribution gives the probability for each value of  $X$ , where  $X$  equals the number of occurrences of category  $A$  in a sample of  $n$  events.
3. The null hypothesis states that the probability of choosing between the deep side and the shallow side is just chance:  $p(\text{deep side}) = \frac{1}{2}$ .
4. The normal approximation to the binomial distribution requires that both  $pn$  and  $qn$  are at least 10. With  $n = 15$ ,  $pn = qn = 7.5$ . The normal approximation should not be used.
5. False.

---

### 8.7 Suggested Readings

---

1. Gravetter, F.J., & Wallnau, L.B. (2013). Statistics for behavioural sciences (9<sup>th</sup> Ed.). Cengage Learning India Private Ltd.

**NOTES**


---

## UNIT 9: ANALYSIS OF VARIANCE

---

- 9.1 Purpose and Logic of Analysis of Variance**
- 9.2 Assumptions of Analysis of Variance.**
- 9.3 One Way Analysis of Variance.**
  - 9.3.1 Illustration**
- 9.4 Two Way Analysis of Variance**
  - 9.4.1 Illustration**
- 9.5 Let us Sum up**
- 9.6 Unit – End exercises**
- 9.7 Answer to check your progress**
- 9.8 Suggested Readings**

---

### 9.1 PURPOSE FOR ANALYSIS OF VARIANCE (ANOVA) :

---

Analysis of Variance is carried out when we want to determine the significances of difference between more than two group means. In ANOVA, the variable (independent or quasi-independent) that designates the groups being compared is called a **factor**. In addition, the individual groups or treatment conditions that are used to make up a factor are called the *levels* of the factor. The technique of ANOVA as a single composite test of significance, for the difference between several group means demands derivation of two independent estimates of the population variance, one based on the variance of group means and other variance within groups.

### THE LOGIC OF ONE-WAY ANALYSIS OF VARIANCE

**Variation within and between groups:** Suppose if we have selected three samples of 10 cases each at random and have given them three different treatment conditions. If we wish to test our null hypothesis for possible difference in treatment conditions, we have to deal with two types of variation.

**Within group Variation:** Within each sample the individual vary about the sample mean. We call this within group variation. It is a direct reflection of the inherent variation among individual given the same treatment. We can present exactly the same stimulus to everyone in a group and still observe variation in reaction times.

**Between Group Variations:** The means of the samples vary among themselves. It is important to realize that between groups variation is also a reflection of the inherent variation among individuals. The greater the inherent variation among individuals, the greater the opportunity for chance to produce sample means that vary from one another.



**NOTES**

Now the question is what the variation in scores within and among groups has to do with testing hypothesis about means? The answer is that these two kinds of variation provides the basis for testing the null hypothesis.

For example, In a study, there were to forty boys in all belonging to four different locations. If we add the IQ score of these 40 boys and divide the sum by 40, we get the value of the grand or general mean which is the best estimate of the population mean. There are 4 groups of 10 boys each. The mean and IQ scores of 10 boys in each groups is called the group mean and in this way, there will be the 4 group means, which will vary considerably from each other.

Now the question arises as to how far does an IQ score of a particular boy belonging to a particular sample deviate from the grand mean of 40 boys. We may observe that the deviation of the score from the mean of that particular group and the deviation of the mean of the group from the grand mean. For deriving useful results we can use variance as a measure of dispersion in place of standard deviation. As variance of an individual's score from the grand mean may be broken into two parts i.e. within group variance and between group variance

If we wanted to compare the variance of the two variables one obvious way is to form a ratio. To do this all we need to do is to divide the variance of one variable by the variance of the other. It is convenient to divide the larger variance by a smaller one. A ratio of this form gives us the information about whether one variance is larger than another and by what amount. If the difference is big we would also like to know if it is big enough to be meaningfully different.

The related question in this case is: how plausible is it that the variances we have observed are really the estimates of the same population variance? i.e. can they be treated as if they were two different samples drawn in the same way from the same population. If the variances are both estimates of the same population variance then they should be pretty similar and any difference between them will be due to chance variations. In order to examine this issue further we need to know how this ratio based statistic behaves. In the case of computing variances, we need to know about the sampling distribution of the ratio of variance. As it happens, the behaviour of the sampling distribution is quite well understood and providing that certain assumptions hold, it follows a particular pattern. This pattern was called F-ratio denoted by

$$F = \frac{\text{between group variance}}{\text{within group variance}}$$

**NOTES**

is used as a critical ratio for determining the significance of difference between group means at a given level of significance. If we assume that these two variances are drawn from the same population after calculating  $F$  ratio, what we now need to do is check what proportions of the  $F$  distribution is cut off by this value. The critical value in the  $F$  table will suggest whether or not the obtained ratio is particularly unusual one to be obtained if the samples really were both drawn from the same populations with the same variance.

---

### 9.2 ASSUMPTIONS IN ANALYSIS OF VARIANCE

---

The following are the fundamental assumption for the use of analysis of variance technique:

1. The dependent variable which is measured should be normally distributed in the population.
2. The individuals being observed should be distributed randomly in the groups
3. Within-groups variances must be approximately equal.
4. The contributions to variance in the total sample must be additive.

#### Check your Progress – 1

1. ANOVA is a statistical procedure that compares two or more treatment conditions for differences in variance. (True or false?)

2. In ANOVA, the total variability is partitioned into two parts. What are these two variability components called, and how are they used in the  $F$ -ratio? The two components are between-treatments variability and within-treatments variability. Between-treatments variance is the numerator of the  $F$ -ratio, and within-treatments variance is the denominator.

3. What happens to the value of the  $F$ -ratio if differences between treatments are increased? What happens to the  $F$ -ratio if variability inside the treatments is increased? As differences between treatments increase, the  $F$ -ratio increases. As variability within treatments increases, the  $F$ -ratio decreases.

---

### 9.3 ONE WAY ANALYSIS OF VARIANCE

---

**The Null Hypothesis:** One way analysis of variance allows us to compare the means of two or more groups simultaneously. It is closely related to 't' test. In one-way analysis of variance, there may be two or more treatment conditions, often referred to as different levels of the independent variable. The Null Hypothesis in ANOVA states that if the different treatments have no different effect on the variable under observation, then we expect the

**NOTES**

population means to be equal. To inquire whether variation in treatment conditions makes a difference, we test the Null hypothesis. The Null hypothesis in ANOVA is often referred to as an omnibus hypothesis (i.e. covering many situations at once) and ANOVA itself as an omnibus test.

As already pointed out, the deviation of an individual's score belonging to a sample or a group of population from the grand mean can be divided into two parts: (i) deviation of the individual's score from his group mean; and (ii) deviation of the group mean from the grand mean. Consequently, the total variance of the scores of all individuals included in the study may be partitioned into within-group variance and between-groups variance. The formula used for the computation of variance ( $\sigma^2$ ) is  $\Sigma x^2/N$ , i.e., sum of the squared deviation from the mean value divided by total frequencies. Hence, by taking  $N$  as the common denominator, the total sum of the squared deviation of scores around the grand or general mean for all groups combined can be made equal to the sum of the two partitioned, between-groups and within-groups sum of squares. Mathematically,

Total sum of squares (around the general mean) = between-groups sum of squares + within-groups sum of squares

or

$$S_t^2 = S_b^2 + S_w^2$$

Hence the procedure for the analysis of variance involves the following main tasks:

- (i) Computation of total sum of squares ( $S_t^2$ )
- (ii) Computation of between-groups sum of squares ( $S_b^2$ )
- (iii) Computation of within-groups sum of squares ( $S_w^2$ )
- (iv) Computation of F-ratio.
- (v) Use of t test (if the need for further testing arises).

All these tasks may be carried out in a series of systematic steps. Let us try to understand these steps by adopting the following terminology:

$X$  = Raw score of any individual included in the study  
(any score entry in the given table)

$\Sigma X$  = Grand Sum

$\frac{\Sigma X}{N}$  = Grand or general mean

$X_1, X_2, \dots$  denote scores within first group, second group, ...

$n_1, n_2, n_3 =$  No. of individuals in first, second and third groups

$\frac{\Sigma X_1}{N}, \frac{\Sigma X_2}{N}, \dots$  denote means of the first group, second group, ...

## NOTES

$N =$  total No. of scores or frequencies

### Let us now outline the steps.

*Step 1. Arrangement of the given table and computation of some initial values.* In this step, the following values needed in computation are calculated from the experimental data arranged in proper tabular form:

(i) Sum of squares,  $\Sigma X_1, \Sigma X_2, \dots$  and the grand sum,  $\Sigma X$

(ii) Group means,  $\frac{\Sigma X_1}{N}, \frac{\Sigma X_2}{N}, \dots$ , and the grand mean  $\frac{\Sigma X}{N}$

(iii) Correction term  $C$  computed by the formula

$$C = \frac{(\Sigma X)^2}{N} = \frac{\text{Square of the grand sum}}{\text{Total No. of cases}}$$

*Step 2. Arrangement of the given table into squared-form table and calculation of some other values.* The given table is transformed into a squared-form table by squaring the values of each score given in the original table and then the following values are computed;

(i)  $\Sigma X_1^2, \Sigma X_2^2, \Sigma X_3^2, \dots$

(ii)  $\Sigma X^2$

*Step 3. Calculation of total sum of squares.* The total sum of squares around the general mean is calculated with the help of the following formula:

$$\begin{aligned} S_t^2 &= \Sigma X^2 - C \\ &= \Sigma X^2 - \frac{(\Sigma X)^2}{N} \end{aligned}$$

*Step 4. Calculation of between-group sum of squares.* The value of the between-groups sum of squares may be computed with the help of the following formula:

$$S_b^2 = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \dots + \frac{(\Sigma X_n)^2}{n_n} - C$$

*Step 5. Calculation of within-group sum of squares.* Between groups and within - groups sum of squares constitute the total sum of

**NOTES**

squares. Therefore, after steps 3 and 4, the value of within-groups sum of squares (sum of squares of deviation within each group about their respective groups means) may be calculated by subtracting the value of between-groups sum of squares from the total sum of squares. Its formula, therefore, goes thus :

$$S_w^2 = S_t^2 - S_b^2$$

*Step 6. Calculation of the number of degrees of freedom.* All these sums of squares calculated in step 3-5 possess different degrees of freedom given by

$$\text{Total sum of squares, } (S_t^2) = N - 1$$

$$\text{Between - groups sum of squares, } (S_b^2) = K - 1$$

$$\text{Within groups sum of squares, } (S_w^2) = (N - 1) - (K - 1) = N - K$$

where N represents the total number of observations, scores or frequencies and K, the number of groups in the research study.

*Step 7. Calculations of F-ratio.* The value of F-ratio furnishes a comprehensive or overall test of significance of the difference between means. For its computation, we have to arrange the data and computation work in the following manner.

Source of variance	Sum of squares	df	Means square variance
Between - groups	$S_b^2$ (computed in step 4)	K-1	$\frac{S_b^2}{K - 1}$
Within - groups	$S_w^2$ (computed in step 5)	N-1	$\frac{S_w^2}{N - K}$

$$F = \frac{\text{Mean square variance between - groups}}{\text{Mean square variance within - groups}}$$

*Step 8: Interpretation of F-ratio:* F-ratio are interpreted by the use of the critical value of F-ratios. This table has the number of degrees of freedom for the greater mean square variance across the top and the number of degrees of freedom in the smaller mean square variance on the left-hand side. If our computed value of F is equal to or greater than the critical tabled value of F at a given level of significance 0.05 or 0.01, it is assumed to be significant and consequently we reject the null hypothesis of no

**NOTES**

difference among these means at that level of significance. However, a significant F does not tell us which of the group means differ significantly; it merely tells us that at least one of the means is relatively different from some other. Consequently, there arises a need for further testing to determine which of the difference between means are significant.

In case our computed value of F is less than the critical tabled value of F at a given level of significance, it is taken as non - significant and consequently, the null hypothesis cannot be rejected. Then there is no reason for further testing (as none of the difference between means will be significant). In a summarized form, the above analysis may be represented as follows :

F → Significant → Null hypothesis rejected → Need for further testing

F → Non-significant → Null hypothesis not rejected → No need for further testing

Generally as and when we get the value of F as less than 1, we straightaway interpret it as non - significant resulting in the non - rejection of the null hypothesis.

*Step 9. Testing differences between means with the t test.* When F is found significant, the need for further testing arises. We take pairs of the group means one by one for testing the significance of differences. The t test provides an adequate procedure for testing the significance when we have means of only two samples or group at a time for consideration. Therefore, we make use of the t test to test the difference between pairs of means.

As we have seen in Chapter 10, the usual formula for computation of t value is

$$t = \frac{D}{\sigma_D} = \frac{\text{Difference between two means}}{\text{Standard error of the difference between the means}}$$

and  $\sigma_D$  is computed by the formula

$$\sigma_D = \sigma \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

Where

$\sigma$  = Pooled SD of the samples drawn from the same population.

$n_1, n_2$  = Total No. of cases in samples 1 and II, respectively.

In the analysis of variance technique, within groups means square variance provides us the value of  $\sigma^2$ , the square root of which can give us the required pooled SD of the samples or groups included in our study.

The degrees of freedom for within groups sum of squares are given by the formula  $N - K$ . With these degrees of freedom we can read the  $t$  values, at 0.05 and 0.01 levels of significance. If the computed value of  $t$  is found to be equal to or greater than the critical value of  $t$  at 0.05 or 0.01 levels, we can reject the null hypothesis at that level of significance.

**9.3.1 Illustration:** The aim of an experimental study was to determine the effect of three different technique of training on the learning of a particular skill. Three groups, each consisting of seven students of classes IX, assigned randomly, were given training through these different techniques. The scores obtained on a performance test were recorded as follows:

Group I	Group II	Group III
3	4	5
5	5	5
3	3	5
1	4	1
7	9	7
3	5	3
6	5	7

Test the difference between groups by adopting the analysis of variance technique.

Solutions

Step 1. Original table computation

Organization of Data

Rating of the coaching experts			Total
Group I ( $X_1$ )	Group II( $X_2$ )	Group III ( $X_3$ )	
3	4	5	12
5	5	5	15
3	3	5	11
1	4	1	6
7	9	7	23
3	5	3	11
6	5	7	18
$\Sigma X_1=28$	$\Sigma X_2=35$	$\Sigma X_3=33$	$\Sigma X=96$
			<b>Grand Total</b>

$$n_1 = n_2 = n_3 = 7, N = n_1 + n_2 + n_3 = 21$$

**NOTES**

$$\text{Group means} = \frac{\Sigma X_1}{n_1} = \frac{28}{7} = 4$$

$$\frac{\Sigma X_2}{n_2} = \frac{35}{7} = 5$$

$$\frac{\Sigma X_3}{n_3} = \frac{33}{7} = 4.71$$

$$\text{Correction term } C = \frac{(\Sigma X)^2}{N} = \frac{96 \times 96}{21} = \frac{9216}{21} = 438.85$$

Step 2: Squared - Table computation

$X_1^2$	$X_2^2$	$X_3^2$	<b>Total</b>
9	16	25	50
25	25	25	75
9	9	25	43
1	16	1	18
49	81	49	179
0	25	9	43
36	25	49	110
$\Sigma X_1^2 = \mathbf{138}$	$\Sigma X_2^2 = \mathbf{197}$	$\Sigma X_3^2 = \mathbf{193}$	$\Sigma X^2 = \mathbf{518}$

Step 3 Total sum of squares ( $S_t^2$ ):

$$S_t^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = \Sigma X^2 - C = 518 - 438.85 = 79.15$$

Step 4. Between - group sum of squares ( $S_b^2$ ):

$$\begin{aligned} S_b^2 &= \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} - C \\ &= \frac{28 \times 28}{7} + \frac{35 \times 35}{7} + \frac{33 \times 33}{7} - 438.85 \\ &= \frac{784 + 1225 + 1089}{7} - 438.85 \\ &= 442.57 - 438.85 = 3.72 \end{aligned}$$

Step 5: Within - group sum of squares ( $S_w^2$ ). This is obtained as



$$S_w^2 = S_1^2 - S_b^2 = 79.15 - 3.72 = 75.43$$

Step 6. Number of degrees of freedom

For total sum of squares

$$(S_1^2) = N - 1 = 21 - 1 = 20$$

For between - groups sum of squares.

$$(S_b^2) = K - 1 = 3 - 1 = 2$$

For within - groups sum of squares

$$(S_w^2) = N - K = 21 - 3 = 18$$

Step 7. Calculation of  $F$  - ratio

Source of variation	Sum of squares	df	Mean square variance
Between - groups	$S_b^2 = 3.72$	2	$3.72/2 = 1.86$
Within - groups	$S_w^2 = 75.43$	18	$75.43/18 = 4.19$

$$F = \frac{\text{Mean square variance between - group}}{\text{Mean square variance within - group}} = \frac{1.86}{4.19} = 0.444$$

Step 8. Interpretation of  $F$ -ratio. The  $F$ - ratio table is referred to for 2 degrees of freedom for smaller mean square variance on the left hand side, and for 18 degrees of freedom for greater mean square variance across the top. The critical values of  $F$  obtained by interpolation are as follows:

Critical value of  $F = 19.43$  at 0.05 level of significance

Critical value of  $F = 99.44$  at 0.01 level of significance

Our computed value of  $F (.444)$  is not significant at both the levels of significance and hence, the null hypothesis cannot be rejected and we may confidently say that the differences between means are not significant and therefore, there is no need for further testing with the help of  $t$  test.

### Check your Progress 2

4. A researcher uses an ANOVA to compare three treatment conditions with a sample of  $n = 8$  in each treatment. For this analysis, find  $df_{total}$ ,  $df_{between}$ , and  $df_{within}$ .

5. A researcher conducts an experiment comparing four treatment conditions with a separate sample of  $n = 6$  in each treatment. An ANOVA

### NOTES

is used to evaluate the data, and the results of the ANOVA are presented in the following table. Complete all missing values in the table. *Hint:* Begin with the values in the *df* column.

**NOTES**

Source	Sum of squares	MS	df	F Ratio
Between treatments	-	-	-	-
Within treatments	-	2	-	
Total	58			

**9.4 TWO-WAY ANALYSIS OF VARIANCE**

So far, we have dealt with one-way analysis of variance involving one experimental variable. However, experiments may be conducted in the fields of education and psychology for the simultaneous study of two experimental variables. Such experiments involve two-way classification based on two experimental variables. Let us make some distinction between the need for carrying out one-way and two-way analyses of variance through some illustrations.

Suppose that we want to study the effect of four different methods of teaching. Here, the method of teaching is the experimental variable (independent variable which is to be applied at four levels). We take four groups of students randomly selected from a class. These four groups are taught by the same teacher in the same school but by different methods. At the end of the session, all the groups are tested through an achievement test by the teacher. The mean scores of these four groups are computed. If we are interested in knowing the significance of the differences between the means of these groups, the best technique is the analysis of variance. Since only one experimental variable (effect of the method of teaching) is to be studied, we have to carry out one-way analysis of variance.

Let us assume, that there is one more experimental or independent variable in the study, in addition to the method of teaching. Let it be the school system at three levels which means that three school systems are chosen for the experiment. These systems can be: government school, government-aided school and public school. Now the experiment will involve the study of 4 x 3 groups. We have 4 groups each in the three types of schools (all randomly selected). The achievement scores of these groups can then be compared by the method of analysis of variance by establishing a null hypothesis that neither the school system nor the methods have anything to do with the achievement of pupils. In this way,

we have to simultaneously study the impact of two experimental variables, each having two or more levels, characteristics or classifications and hence we have to carry out the two-way analysis of variance.

In the two-way classification situation, an estimate of population variance, i.e. total variance is supposed to be broken into: (i) variance due to methods alone, (ii) variance due to school alone, and (iii) the residual variance in the groups called interaction variance ( $M \times S$ ;  $M$ =methods,  $S$  = schools) which may exist on account of the following factors:

1. Chance
2. Uncontrolled variables like lack of uniformity in teachers
3. Relative merits of the methods (which may differ from one school to another).

In other words, there may be interaction between methods and schools which means that although no method may be regarded as good or bad in general, yet it may be more suitable (or unsuitable) for a particular school system. Hence, the presence of interaction variance is a unique feature with all the experimental studies involving two or more experimental variables. In such problems, requiring two or more ways of analysis of variance, we will have to take care of the residual or interaction variance in estimating their population variance. If the null hypothesis is true, variance in terms of the method should not be significantly different from the interaction variance. Similarly, the variance due to schools may also be compared with the interaction variance. For all such purposes of comparison, the F-ratio test is used.

The design of the above two-way classification experiment can be further elaborated by having more experimental or independent variables, each considered at several levels. As the number of variables increases, the order of interaction also increases. According to the needs and requirements of the experiment, a research worker has to select a particular experimental design involving two or more independent or experimental variables.

9.4.2 Illustration :In a research study, there were two experimental or independent variable: a seven member group of player and three coaches who were asked to rate the players in terms of a particular trait on a ten - point scale. The data were recorded as under:

Rating by three coaches	Players						
	1	2	3	4	5	6	7
A	3	5	3	1	7	3	6
B	4	5	3	4	9	5	5

## NOTES

C	5	5	5	1	7	3	7
---	---	---	---	---	---	---	---

**NOTES**

Apply the technique of analysis of variance for analysing these data.

*Step 1 : Arrangement of the data in a proper table and computation of essential initial value*

Players	Rating of coaches			Total of rows	Square of the total of rows
	By A (X <sub>1</sub> )	By B (X <sub>2</sub> )	By C (X <sub>3</sub> )		
1	3	4	5	12	144
2	5	5	5	15	225
3	3	3	5	11	121
4	1	4	1	6	36
5	7	9	7	23	529
6	3	5	3	11	121
7	6	5	7	18	324
Total of columns	ΣX <sub>1</sub> = 28	ΣX <sub>2</sub> = 35	ΣX <sub>3</sub> = 33	ΣX = 96	(Grand Total)
Square of the totals of the column	(ΣX <sub>1</sub> ) <sup>2</sup> = 784	(ΣX <sub>2</sub> ) <sup>2</sup> = 1225	(ΣX <sub>3</sub> ) <sup>2</sup> = 1089	(ΣX) <sup>2</sup> = 9216	

Here,

$$N = n_1 + n_2 + n_3 = 7 + 7 + 7 = 21$$

$$C = \frac{(\Sigma X)^2}{N} = \frac{9216}{21} = 438.85$$

*Step 2: Arrangement of the table in square form and computation of some essential values.*

X <sub>1</sub> <sup>2</sup>	X <sub>2</sub> <sup>2</sup>	X <sub>3</sub> <sup>2</sup>	Total
9	16	25	50
25	25	25	75
9	9	25	43
1	16	1	18
49	81	49	179

9	25	9	43
36	25	49	110
$\Sigma X_1^2 = 138$	$\Sigma X_2^2 = 197$	$\Sigma X_3^2 = 183$	$\Sigma X^2 = 518$

**NOTES**

Step 3. Calculation of the total sum of squares ( $S_t^2$ ) (around grand mean).

$$S_t^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = \Sigma X^2 - C$$

$$= 518 - 438.85 = 79.15$$

Step 4 : Calculation of the sum of squares for rows ( $S_r^2$ ) (between the means of players to be rated by three coaches).

$$S_r^2 = \frac{144 + 225 + 121 + 36 + 529 + 121 + 324}{3} - C$$

$$= \frac{1500}{3} - 438.85 = 61.15$$

Step 5 : Calculation of the sum of squares for columns ( $S_c^2$ ) (between the means of coaches rating 7 players).

$$S_c^2 = \frac{784 \times 1225 \times 1089}{7} - C$$

$$= 442.57 - 438.85 = 3.72$$

Step 6: Calculation of the interaction or residual sum of squares ( $S_t^2$ ). The interaction sum of squares is given by

$$S_t^2 = S_t^2 - (S_r^2 + S_c^2)$$

$$= 79.15 - (61.15 + 3.72) = 14.28$$

Step 7. Computation of F - ratios.

Source of variation	Df	Sum of squares	Mean square variance
Rows (players)	(r-1)=6	61.15	$\frac{61.15}{6} = 10.19$
Columns (coaches)	(c-1)=2	3.72	$\frac{3.72}{2} = 1.86$

Interaction or residual	(r-1) =12	(c-1)	14.28	$\frac{14.28}{12} = 1.19$
-------------------------	--------------	-------	-------	---------------------------

**NOTES**

$$F(\text{for rows}) = \frac{\text{Mean square variance between rows (players)}}{\text{Mean square variance interms of interaction}}$$

$$= \frac{10.19}{1.19} = 8.56$$

$$F(\text{for columns}) = \frac{\text{Mean square variance between columns (coaches)}}{\text{Mean square variance interms of interaction}}$$

$$= \frac{1.86}{1.19} = 1.56$$

Step 8. Interpretation of F-ratios.

Kind of F	df for greater mean square variance	df for smaller means square variance	Critical values of F		Judgment about the significance of computed F	Conclusion
			at 0.05 level	At 0.01 level		
F (for rows)	2	12	3.88	6.93	Significant at both levels	Null hypothesis is replaced
F (for columns)	6	12	3.00	4.82	Not significant at both levels	Null hypothesis is not rejected

Hence, F (for rows) is highly significant and hence null hypothesis is rejected. It indicates that the coaches did discriminate among the players. The second F (for columns) is insignificant and hence, the null hypothesis cannot be rejected. It indicates that the coaches did not differ significantly among themselves in their ratings of the players. In other words, their ratings may be taken as trustworthy and reliable.

**Check your Progress 3**

6. When is Two –way analysis of variance a preferred statistical measure?

**9.5 Lets sum up:**

Analysis of variance (ANOVA) is a statistical technique that is used to test the significance of mean difference among two or more treatment

conditions. The null hypothesis for this test states that, in the general population, there are no mean differences among the treatments. The alternative states that at least one mean is different from another. The test statistic for ANOVA is a ratio of two variances called an  $F$ -ratio. The variances in the  $F$ -ratio are called mean squares, or  $MS$  values. Each  $MS$  is computed by Sum of Squares (SS)÷df. For the independent-measures ANOVA, the  $F$ -ratio is  $MS_{\text{between}} \div MS_{\text{within}}$

## NOTES

### 9.6 Unit End Exercises:

1. Explain the term factor and the levels of a factor with an example
2. Explain the logic of ANOVA
3. There is some research indicating that college students who use Facebook while studying tend to have lower grades than non-users. A representative study surveys students to determine the amount of Facebook use during the time they are studying or doing homework. Based on the amount of time spent on Facebook, students are classified into three groups and their grade point averages are recorded. The following data show typical pattern of results

Non – User	4	2	5	1	3	7	4	4	8	2
Rarely Use	5	7	3	4	8	6	2	7	3	5
Regularly Use	5	7	6	6	8	9	6	4	6	8

Use ANOVA to determine whether there are significant mean differences among the three groups.

### 9.7 Answers to check your progress:

1. False. Although ANOVA uses variance in the computations, the purpose of the test is to evaluate differences in *means* between treatments.
2. The two components are between-treatments variability and within-treatments variability. Between-treatments variance is the numerator of the  $F$ -ratio, and within-treatments variance is the denominator.
3. As differences between treatments increase, the  $F$ -ratio increases. As variability within treatments increases, the  $F$ -ratio decreases.
4. Ans :  $df_{\text{total}} = 23$ ;  $df_{\text{between}} = 2$ ;  $df_{\text{within}} = 21$ .
- 5.

Source	Sum of squares	$MS$	df	F Ratio
Between treatments	18	6	3	3.00
Within treatments	40	2	20	
Total	58		23	

**NOTES**

6. Two way analysis of variance is a preferred statistical measure when we wish to examine the influence of two different categorical independent variables on one continuous dependent variable. The two-way ANOVA not only aims at assessing the main effect of each independent variable but also if there is any interaction between them.

---

**9.8 Suggested Readings**

---

3. Mangal S.K. (1987). *Statistics in Psychology and Education*. New Delhi: PHI Learning Private Ltd.
4. Gravetter, F.J., & Wallnau, L.B. (1995). *Essentials of statistics*. (2<sup>nd</sup> Ed.). West Publishing Company.
5. King, B.M., & Bear, G. (2001). *Statistical Reasoning in Psychology and Education* (3<sup>rd</sup> Ed.). John Wiley & Sons Inc.



---

## UNIT 10: NON- PARAMETRIC TESTS

---

**10.1.Nature and Assumptions****10.1.1 Advantages of Parametric test****10.1.1 Disadvantages of Parametric test****10.2 Chi-square****10.2.1 Assumptions of the Chi-square test****10.2.2 Null hypothesis for chisquare****10.2.3 Steps to compute Chi-square.****10.2.4 Uses of chi-square test****10.3 Contingency Co-efficient****10.4 Median test****10.4.1 Steps in the computation of the median test****10.5 Sign test****10.6 Friedman test.****10.7 Let's sum up****10.8 Unit End Exercises****10.9 Answers to check your progress****10.10 Suggested Readings****NOTES**


---

### 10.1 Nature and Assumptions

---

A non-parametric statistical test is one which does not specify any conditions about the parameter of the population from which the sample is drawn. Since these statistical tests do not make any specified and precise assumption about the form of the distribution of the population, these are also known as **distribution free statistics**.

For a non - parametric statistical test, the variables under study should be continuous and the observation should be independent. But these condition are not too rigid & elaborate as in parametric statistics. A non - parametric tests should be used only in the following case.

1. The shape of the distribution of the population from which a sample is drawn is not known to be a normal one.
2. The variables have been quantified on the basis of nominal measures (or frequency counts).
3. The variables have been quantified on the basis of ordinal measures (or ranking)

As non-parametric statistical tests are based upon frequency counts or rankings rather than on the measured values, they are less precise, and are less likely to reject null hypothesis when it is fake. That is why, a non-parametrical statistical test is used only when the parametric assumptions cannot be met.

**NOTES**

Bradly (1968) has enumerated several advantages and disadvantages of non-parametric statistical tests in comparison to parametrical tests.

**10.1.1 Advantages of non-parametric tests.**

1. **Simplicity and facilitation in derivation:** Most of the non-parametric statistics can be derived by using simple computational formula.
2. **Wide scope of application:** Since non-parametric states are based upon fewer and less rigid and elaborate assumptions regarding the form of the population distributions they can be easily applied to much wide situations.
3. **Susceptibility to violation of assumption:** Since the assumptions in case of non-parametric statistics are fewer and less elaborate, they are less susceptible to violations.
4. **Speed of application:** When the size of the sample is small, calculation of non-parametric statistics is faster than parametric statistics.
5. **Type of measurement required:** Non parametric statistics require measurement based upon a nominal scale and ordinal scale.
6. **Impact of sample size:** When sample size is 10 or less than 10, non-parametric statistics are easier, quicker and more efficient than parametric statistics. However, as the sample size increases, non-parametric statistics become time consuming, laborious and less efficient than parametric statistics.
7. **Statistical efficiency:** Non-parametric tests are often more convenient than the parametric tests. If the data is such that it meets all assumption of non-parametric statistics but not of parametric statistics, then non-parametric statistics have statistical efficiency equal to parametric statistics. If both parametric and non-parametric statistics are applied to the data, which fulfills all assumptions of parametric tests, the distribution - free statistics become more efficient with a small sample size but they become less and less efficient as sample size increases.

**10.1.2 Disadvantages of non-parametric tests**

1. According to Moses (1952), the non-parametric statistics have lower statistical efficiency than parametric statistics when sample size is large, preferably above 30.

- If all the assumptions of parametric statistics are fulfilled, Siegel (1956), and Siegel & Castellan (1988) consider the use of non-parametric statistics as simply 'wasteful of data'.

### Check your Progress – 1

- Non parametric tests make assumptions about the population distribution-True/False
- Data measured on a ratio scale is suitable for non – parametric tests – True/False

### NOTES

---

## 10.2 Chi-square test

---

The chi-square is one of the most important non-parametric statistics, which is used for several purposes. For this reason, Guilford (1956) has called it general purpose statistics. The chi-squares applies only to discrete data. However, any continuous data can be reduced to the categories in such a way that they can be treated as discrete data and then the application of chi-square is justified. The formula for computing  $\chi^2$  is given below

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- $\chi^2$  = chi-square  
 $f_o$  = obtained or observed frequency  
 $f_e$  = expected theoretical frequency

There are two types of chi-square test. The chi-square test for goodness of fit involves testing the levels of a single nominal variable and the chi-square test of independence is used when there are two nominal variables, each with several categories.

### 10.2.1 Assumptions of chi-square statistic

Prior to using the chi square test there are certain requirements that must be met. These are :-

- The data must be in the form of frequencies counted in each of a set of categories.
- The total numbers observed must exceed 20.
- The expected frequency under the  $H_0$  hypothesis in any one fraction must not normally be less than 5.
- All the observations must be independent of each other. In other words, one observation must not have an influence upon another observation.

**NOTES****10.2.2.Null Hypothesis**

**Null hypothesis for Chi square goodness of fit :** Generally  $H_0$  falls into one of the following categories:

**1. No Preference, Equal Proportions.** The null hypothesis often states that there is no preference among the different categories. In this case,  $H_0$  states that the population is divided equally among the categories. For example, a hypothesis stating that there is no preference among the three leading brands of soft drinks, would mean that if 60 respondents were surveyed the expected frequencies will be set 20 for each category

**2.No Difference from a Known Population.** The null hypothesis can state that the proportions for one population are not different from the proportions that are known to exist for another population. For example, suppose it is known that 28% of the licensed drivers in the state are younger than 30 years old and 72% are 30 or older. A researcher might wonder whether this same proportion holds for the distribution of speeding tickets. The null hypothesis would state that tickets are handed out equally across the population of drivers, so there is no difference between the age distribution for drivers and the age distribution for speeding tickets.

**Null Hypothesis for Chi-Square test of Independence:** The chi-square test for independence uses the frequency data from a sample to evaluate the relationship between two variables in the population. Each individual in the sample is classified on both of the two variables, creating a two-dimensional frequency-distribution matrix. The frequency distribution for the sample is then used to test hypotheses about the corresponding frequency distribution for the population.

**10.2.3 Steps for computing the chi-square statistic:**

1. Determine the actual, observed frequency in each category. The observed frequency is likely to be a whole number as it is obtained by counting the number of people in the sample.
2. Determine the expected frequency in each category. (This can be in decimals). In case of hypothesis of equal probability, the expected frequencies are obtained by dividing  $N$  by the total number of categories.
3. In each category, take observed minus expected frequencies.
4. Square each of the differences.
5. Divide each squared difference by the expected frequency for its category.
6. Add up the results of step 5 for all the categories.

The numerator of the chi-square statistics is obtained by measuring the difference between the data (the  $f_o$  values) and the hypothesis (represented

by the  $f_e$  values) and squaring them. The expected frequencies for the goodness-of-fit test are determined by expected frequency =  $f_e = pn$  where  $p$  is the hypothesized proportion (according to  $H_0$ ) of observations falling into a category and  $n$  is the size of the sample. At the end, we add the values to obtain the total discrepancy between the data and the hypothesis. Thus, a large value for chi-square indicates that the data do not fit the hypothesis, and leads us to reject the null hypothesis.

In order to determine the significance of Chi-square degrees of freedom are computed. For Chi-square test of goodness of fit,  $df$  is obtained by the formula  $C-1$  where  $C$  is the number of categories in the variable. Degrees of freedom for the test for independence are computed by  $df = (R - 1)(C - 1)$  where  $R$  is the number of row categories and  $C$  is the number of column categories.

#### 10.2.4. Uses of chi-square.

1. The chi-square may be used as a test of equal probability hypothesis. Suppose for example, 100 students answer an item in an attitude scale the item has five response options. According to equal probability hypothesis, the expected frequency of responses given by 100 students would be 20 in each. The chi-square test would test whether or not the equal probability hypothesis is tenable.
2. The second use of the chi-square test is in testing the significance of the independence hypothesis. The chi-square is a measure of the degree of relationship in such a situation.
3. The third important use of chi-square is in testing a hypothesis regarding the normal shape of a frequency distribution. When chi-square is used in this connection, it is commonly referred to as a test of goodness of fit.
4. The fourth use of chi-square is in testing the significance of several statistics. In testing the significance of phi-co-efficient, co-efficient of concordance and co-efficient of contingency we convert the values obtained from these test into chi-square values. If the chi-square value appears to be a significant one, we also take their original values as significant.

#### Check your Progress =2

3. For a chi-square test, the observed frequencies are always whole numbers. (True or false?)
4. For a chi-square test, the expected frequencies are always whole numbers. (True or false?)

### NOTES

**NOTES**

5. A researcher has developed three different designs for a computer keyboard. A sample of  $n = 60$  participants is obtained, and each individual tests all three keyboards and identifies his or her favourite. The frequency distribution of preferences is as follows:

Design A	Design B	Design C
23	12	25

- What is the  $df$  value for the chi-square statistic?
- Assuming that the null hypothesis states that there are no preferences among the three designs, find the expected frequencies for the chi-square test.

---

### 10.3 Contingency Co-efficient:

The contingency coefficient is a coefficient of association that tells whether two variables or data sets are independent or dependent of each other. It is also known as *Pearson's Coefficient*. It is based on the chi-square statistic, and is defined by:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

in this formula:

- $\chi^2$  is the chi-square statistic,
- $N$  is the total number of cases or observations in our analysis/study,
- $C$  is the contingency coefficient.

The contingency coefficient helps us decide if variable  $b$  is 'contingent' on variable  $a$ . However, it is a rough measure and doesn't quantify the dependence exactly; It can be used as a rough guide:

- If  $C$  is **near zero** (or equal to zero) you can conclude that your variables are independent of each other; there is no association between them.
- If  $C$  is **away from zero** there is some relationship;  $C$  can only take on positive values.

---

### 10.4 Median test:

The *median test* provides a nonparametric alternative to the independent-measures  $t$  test (or ANOVA) to determine whether there are significant differences among two or more independent samples. The null hypothesis for the median test states that the different samples come from populations that share a common median (no differences). The alternative hypothesis

states that the samples come from populations that are different and do not share a common median.

#### 10.4.1 Steps in the computation of the median test

1. The first step in computation of a median test is to compute a common median for both distribution taken together.
2. Subsequently, a 2 x 2 contingency table is set. one for individuals with scores above the median and one for individuals with scores below the median. Finally, for each sample, count how many individuals scored above the combined median and how many scored below. These values are the observed frequencies that are entered in the matrix. The frequency-distribution matrix is evaluated using a chi-square test for independence.
3. The values are substituted in the following formula

$$x^2 = \frac{N[IAD - BCI]^2}{(A+B)(C+D)(A+C)(B+D)}$$

4. When frequencies are less than 5 Yate's correction need to be applied. Formula for Yates correction

$$x^2 = \frac{N[IAD - BCI - N/2]^2}{(A+B)(C+D)(A+C)(B+D)}$$

5. Testing the significance of the obtained  $X^2$  value with the help of the table

#### Illustration

Group A 16, 17, 18, 12, 14, 9, 7, 5, 20, 22, 4, 26, 27, 5, 10, 19

Group B 28, 30, 33, 40, 45, 47, 40, 38, 42, 50, 20, 18, 18, 19

For computing median, both distribution all put together

49 - 53 1

44 - 48 2

39 - 43 3

34 - 38 1  $Median = f + \frac{(N/2 - F)i}{f_m}$

29 - 33 2

24 - 28 3

#### NOTES

**NOTES**

$$19 - 23 \quad 5 \quad = 18.5 + \left(\frac{15-13}{5}\right) \times 5$$

$$14 - 18 \quad 5$$

$$9 - 13 \quad 3 \quad = 20.5$$

$$4 - 8 \quad 5$$

The 2 x 2 contingency table for the above data

	Above Mdn	Not above Mdn	Total
Group A	A	B	
	3	13	16
Group B	C	D	
	10	4	14
Total	13	17	30

$$R2 = \frac{30[3(4) - (13)(10)]^2}{(16)(14)(13)(17)}$$

$$= 8.44$$

$$df = (r-1)(c-1) = 1$$

The obtained chi-square value of 8.44 exceeds the value of 6.65 in the probability table for chi-square. Hence we reject the Null Hypothesis and conclude that the two samples have not been drawn from the same population on from population having equal medians.

**10.5 Sign test:**

The sign test may be used for testing the significance of differences between two correlated samples in which the data is available either in ordinal measurement or simply expressed in terms of positive and negative signs, showing the direction of differences existing between the observed scores of matched pairs. The null hypothesis test here is that the median change is zero, that is there are equal numbers of positive and negative signs. For example, a clinician may observe patients before therapy and after therapy and simply note whether each patient got better or worse. Note that there is no measurement of how much change occurred; the clinician is simply recording the direction of change. Also note that the direction of change is a binomial variable; that is, there are only two values. In this situation it is possible to use a binomial test to evaluate the data. Traditionally, the two possible directions of change are coded by signs, with a positive sign indicating an increase and a negative sign



indicating a decrease. When the binomial test is applied to signed data, it is called a *sign test*.

In many cases, data from a repeated-measures experiment can be evaluated using either a sign test or a repeated-measures *t* test. In general, you should use the *t* test whenever possible. However, there are some cases in which a *t* test cannot or should not be used, and in these situations, the sign test can be valuable. Four specific cases in which a *t* test is inappropriate or inconvenient are described below.

1. When you have infinite or undetermined scores, a *t* test is impossible, and the sign test is appropriate.
2. Often it is possible to describe the difference between two treatment conditions without precisely measuring a score in either condition.
3. Often a sign test is done as a preliminary check on an experiment before serious statistical analysis begins.
4. Occasionally, the difference between treatments is not consistent across participants. This can create a very large variance for the difference scores. As we have noted in the past, large variance decreases the likelihood that a *t* test will produce a significant result. However, the sign test only considers the direction of each difference score and is not influenced by the variance of the scores

---

### 10.6 Friedman test

---

The Friedman test evaluates differences among three or more groups from a repeated-measures design. The scores are the ranks obtained by rank ordering the scores for each participant. With three conditions, for example, each participant is measured three times and would receive ranks of 1, 2, and 3.

### Check your Progress- 3

6. What is the logic behind median test?
7. A researcher used a chi-square test for goodness of fit to determine whether people had any preferences among three leading brands of potato chips. Could the researcher have used a binomial test instead of the chi-square test? Explain why or why not.
8. Why do you think *t* test is more powerful than the sign test?

---

### 10.7 Let's sum up:

---

Chi-square tests are nonparametric techniques that test hypotheses about the form of the entire frequency distribution. Two types of chi-square tests are the test for goodness of fit and the test for independence. The data for these tests consist of the frequency or number of individuals who are

### NOTES

**NOTES**

located in each category. The test for independence is used to assess the relationship between two variables. The null hypothesis states that the two variables in question are independent of each other. That is, the frequency distribution for one variable does not depend on the categories of the second variable. On the other hand, if a relationship does exist, then the form of the distribution for one variable depends on the categories of the other variable. The median test is a non-parametric alternative for independent measures t test or ANOVA . The sign test evaluates the difference between two treatments using the data from a repeated measures design. The difference scores are coded as being either increases (+) or decreases (-). The friedman test is a non-parametric alternative for Two way ANOVA

---

**10.8 Unit End Exercises**

1. What are non – parametric tests? List its advantages and disadvantages.
2. A developmental psychologist would like to determinewhether infants display any colour preferences. Astimulus consisting of four colour patches (red, green,blue, and yellow) is projected onto the ceiling above acrib. Infants are placed in the crib, one at a time, andthe psychologist records how much time each infantspends looking at each of the four colours. The colour thatreceives the most attention during a 100-second testperiod is identified as the preferred colour for that infant. The preferred colours for a sample of 60 infants areshown in the following table:

Red	Green	Blue	Yellow
20	12	18	10

Do the data indicate any significant preferences among the four colors?  
Test at the .05 level of significance.

3. What is contingency co-efficient? When is it used?
4. When should a sign test be preferred over t test?

---

**10.9 Answers to check your progress**

1. False. Non parametric tests make very few assumptions about the population distribution and for this reason they are called distribution free statistics
2. False. Non parametric statistics require measurement based upon a nominal scale and ordinal scale.
- 3.True. Observed frequencies are obtained by counting people in the sample.

4. False. Expected frequencies are computed and may be fractions or decimal values.

5. a.  $df= 2$

b. According to the null hypothesis one-third of the population would prefer each design. The expected frequencies should show one-third of the sample preferring each design. The expected frequencies are all 20.

6. The logic behind the median test is that whenever several different samples are selected from the same population distribution, roughly half of the scores in each sample should be above the population median and roughly half should be below.

7. No, the binomial test cannot be used when there are three categories.

8. t test is preferred over sign test as it uses the actual difference scores not just the signs.

## NOTES

---

### 10.10 Suggested Readings

---

1. Tattao, L. (2007). *Basic Concepts in Statistics*. Retrieved from [https://books.google.com/books?id=og4a\\_700L-4C](https://books.google.com/books?id=og4a_700L-4C) on January 10, 2018

2. Mangal S.K. (1987). *Statistics in Psychology and Education*. New Delhi: PHI Learning Private Ltd.

3. Garrett.H.E.(1981). *Statistics in Psychology and Education*. Vakils, Feffer and Simons Ltd.

4. Aron, Aron., E. N., Coups , E.(2012). *Statistics for Psychology*.(3<sup>rd</sup> Ed). Prentice Hall.

**NOTES**


---

## BLOCK IV: ANALYSIS AND INTERPRETATION OF DATA

---



---

### UNIT 11: PREPARATION OF DATA FOR COMPUTER

---

- 11.1 Introduction:**  
**11.2 Software packages that are used for statistical analysis**  
     **11.2.1. Programs for Quantitative Research.**  
     **11.2.2 Programs for Qualitative Research**  
**11.3 Let's sum up**  
**11.4 Unit End Exercises**  
**11.5 Answers to check your progress**  
**11.6 Suggested Readings**
- 

#### **11.1 Introduction:**

---

For statistical analysis we think of data as a collection of different pieces of information or facts. These pieces of information are called variables. A variable is an identifiable piece of data containing one or more values. Those values can take the form of a number or text (which could be converted into number). For data analysis your data should have variables as columns and observations as rows. The first row should have the column headings. Make sure your dataset has at least one identifier (for example, individual id, family id, etc.). That is, it should have a serial number for each observation

Example:

id	Var1	Var2	Var3	Var 4
1	7	3	5	2
2	6	9	10	3
3	5	6	4	5
4	4	7	1	8

In the table above “var” stands for variable. Instead of this, we can actually have the actual variable that is used for the study. If we have more than one group, all data of group 1 has to be entered followed by group 2, group 3 and so on.

Example	id	Gender	Self esteem	Self efficacy	Aspiration
Group 1	1	1	13	15	12
	2	2	19	10	13
Group 2	3	2	16	14	15
	4	1	17	21	18

**NOTES**

Note that a nominal variable like gender has been converted into number. This is done by assigning values to each gender. For example, Male has been assigned a value of 1 and female has been assigned a value of 2. It is also possible to enter the data in terms of category as either “male” or “female” itself.

Once you have your data in the proper format, before you perform any analysis you need to explore and prepare it first. Here are a few things to keep in mind before you start

1. Make sure variables are in columns and observations in rows.
2. Make sure you have all variables you need.
3. Make sure there is at least one id.
4. If times series make sure you have the years you want to include in your study.
5. Make sure missing data has either a blank space or a dot (‘.’)
6. Make sure to make a back-up copy of your original dataset.
7. Have the codebook handy.
8. If you are using datasets with categorical variables you need to clean them by getting rid of the non-response categories like ‘do not know’, ‘no answer’, ‘no applicable’, ‘not sure’, ‘refused’, etc. No response categories not only affect the statistics of the variable, it may also affect the interpretation and coefficients of the variable if we do not remove them.

**Check your Progress 1:**

- While entering data into a software, the variables are entered in rows: True/False
- How do you enter missing data into your data sheet?

**11.2 Software packages that are used for statistical analysis:**

Statistical software are programs which are used for the statistical analysis of the collection, organization, analysis, interpretation and presentation of data. Statistical packages are collections of software designed to aid in statistical analysis and data exploration. In the last two decades more and more software packages have been designed to help with data analysis. The software is designed for questionnaire-based research, called **quantitative** research, and for other types of research, such as interviews and focus groups, which is called **qualitative** research. Let's have a look at the programs and packages for these two types of research.

## NOTES

### 11.2.1. Programs for Quantitative Research

There are numerous software packages for quantitative research and listed below are few common ones:

1. **(IBM) SPSS, (Statistical Package for the Social Sciences):** SPSS is the most widely used statistics software package within human behaviour research. SPSS has built-in data manipulation tools such as recoding, transforming variables. Social scientists generally learn SPSS as their main package. It gives many options for statistical analysis and offers output in the form of tables, graphs, pie-charts and other diagrams. The software is easy to use, but that does not mean it is problem-free. The data put in must be correct and the suitable form of analysis chosen, otherwise the results will be misleading or simply wrong. Thus, you still need a knowledge of statistics to choose the appropriate method and necessary parameters of the analysis.

2. **R (R Foundation for Statistical Computing):** R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis and graphical facilities for data analysis and display either on-screen or on hardcopy.

3. **SAS (Statistical Analysis Software):** SAS is perfect for traditional analysis of variance and linear regression and meets both specialized and enterprise-wide statistical needs. SAS uses the newest techniques for statistical analysis and large data tasks for any size of data sets. The software is organized in such a way that it helps you to access and manage data, build and deploy statistical models.

4. **STATA:** Stata is an unified software which provides you with the complete package required for data analysis, management and graphics. It is one of the most important softwares available online for statistical analysis

5. **Minitab:** Minitab is one of the best statistical softwares available online. The software allows you to easily transfer Microsoft Excel XISX files directly into Minitab Express. You can evaluate confidence intervals for a parameter of interest such as median or proportion using available resampling techniques. Minitab helps you to find the best regression equations with the help of a model reduction technique.

6. **Microsoft Excel:** Microsoft Excel is spread sheet software. In Excel, you can perform some Statistical analysis. Working in excel is a little tedious compared to other software packages as excel does not have built-in data manipulation tools such as recoding, and transforming variables.

Sharing an Excel spread sheet is a highly tedious affair. Spread sheets create ample opportunities for accidental data loss, which makes it impossible to share crucial data and information.

7. **Max Stat:** This is very easy-to-use and affordable statistical software available online. Since it provides step by step analysis it is handy for students and young scholars.

### 11.2.2 Programs for Qualitative Research

Very often, qualitative research involves dealing with long texts, and so the software involves such functions as text retrieval, coding, and building conceptual networks. For example, this software works if you want to analyze the content of a document for key words or themes and then rank the words relating to a topic by frequency of use. For example, the Computer Assisted Qualitative Data Analysis (CAQDAS) Networking Project has been set up to help researchers find out which is the best software for their needs.

Advantages of using qualitative data analysis software include being freed from manual and clerical tasks, saving time, being able to deal with large amounts of qualitative data, having increased flexibility, and having improved validity and auditability of qualitative research. Concerns include increasingly deterministic and rigid processes, privileging of coding, and retrieval methods; reification of data, increased pressure on researchers to focus on volume and breadth rather than on depth and meaning, time and energy spent learning to use computer packages, increased commercialism, and distraction from the real work of analysis. Hence it is recommended researchers consider the capabilities of the package, their own computer literacy and knowledge of the package, or the time required to gain these skills, and the suitability of the package for their research. (John & Johnson, 2000)

#### Check your Progress: 2

- What do you understand by the term statistical software?
- What do you understand by the term qualitative research?
- What advantage does SPSS have over excel?

---

### 11.3 Let's sum up:

---

For many students, the thought of having to undertake statistical analyses is uncomfortable as they find mathematics and statistics difficult. However, there is lots of support available to make a student comfortable with undertaking statistical analyses such as online courses , you tube videos and even MOOC courses . In addition, there are a multitude of statistical software packages available that can aid in analysis of data

**NOTES**

obtained from both quantitative and qualitative research. It is for most part very easy to learn to use them. To the question “How do I get my data into a statistical package?”. The answer is that most statistical software can read data directly from an Excel spread sheet, so using Excel is often the easiest solution. Secondly, you can always enter data directly into a statistical package, since they nearly all have some form of inbuilt spread sheet. Another solution is to use software like Survey Monkey to collect the data. Survey Monkey has the facility to convert the data into an Excel spread sheet or SPSS format. Though software packages exist to aid the researcher in data analysis, researcher’s intelligence and skill is required in choosing the appropriate statistical tool for data analysis and interpretation

---

**11.4 Unit End Exercises**


---

1. Why is it important to clean the non –response categories?
2. What are the advantages and disadvantages of using data analysis packages for qualitative research?
3. List statistical software packages commonly employed with quantitative research.

---

**11.5 Answers to check your progress**


---

1. False
2. The missing data should either a blank space or a dot (‘.’).
3. Statistical software are programs which are used for the statistical analysis of the collection, organization, analysis, interpretation and presentation of data.
4. *Qualitative research* methods are *defined* as a process that focuses on obtaining data through open-ended and conversational communication such as focus groups and interviews.
5. Excel does not have built-in data manipulation tools such as recoding, and transforming variables like SPSS.

---

**11.6 Suggested Readings:**


---

6. <https://www.princeton.edu/~otorres/DataPrep101.pdf>
7. <https://woofresh.com/statistical-analysis-software/>
8. [https://www.researchgate.net/publication/12189428\\_The\\_pr os\\_and\\_cons\\_of\\_data\\_analysis\\_software\\_for\\_qualitative\\_re search](https://www.researchgate.net/publication/12189428_The_pr os_and_cons_of_data_analysis_software_for_qualitative_re search).



---

# UNIT 12: INFERENCE STATISTICS

---

## NOTES

### 12.1 Introduction to Inferential Statistics

### 12.2 Types of Hypothesis

### 12.3 The Meaning of Statistical Inference

#### 12.3.1 Relationship between probability and Inferential Statistics

### 12.4 Methods of statistical Inference

#### 12.4.1 Statistical Estimation

#### 12.4.2 Statistical hypothesis testing

### 12.5 Let us sum up

### 12.6 Unit end Exercises

### 12.7 Answers to check your progress

### 12.8 Suggested Readings

---

### 12.1 Introduction to Inferential Statistics:

---

Inferential statistics are methods that use sample data as the basis for drawing general conclusions about populations. However, a sample is not expected to give a perfectly accurate reflection of its population. In particular, there will be some error or discrepancy between a sample statistic and the corresponding population parameter. There are two sources of error that may result in samples being different from the population from which it is drawn that include sampling error and sampling bias.

Sampling error is the natural discrepancy, or amount of error, between a sample statistic and its corresponding population parameter. Sampling bias is a bias in which a sample is collected in such a way that some members of the intended population have a lower sampling probability than others. A common cause of sampling bias lies in the design of the study or in the data collection procedure, both of which may favour or disfavour collecting data from certain classes or individuals or in certain conditions

There are two methods of inferential statistics. They are estimation and hypothesis testing. There are few key concepts that form the foundations of inferential statistics. They include knowledge of a) Probability (covered in Chapter 6) b) z score (covered in chapter 7) c) Estimation and Hypothesis testing procedure d) The distribution of sample means. The first two topics has been covered and the next two will be presented in the current and the next chapter respectively

---

### 12.2 Types of Hypotheses:

---

a) Null hypothesis ( $H_0$ ): The null hypothesis states that in the general population, there is no change, no difference, or no relationship. In the

**NOTES**

context of an experiment  $H_0$  predicts that independent variable (treatment) will have no effect on the dependent variable for the population.

b) The alternate hypothesis ( $H_1$ ) states that there is a change, there is a difference, or there is a relationship. In the context of an experiment,  $H_1$  predicts that the independent variable (treatment) will have an effect on the dependent variable.

**Check your Progress - 1**

1. What do you mean by the term statistics and parameters?
2. What are the sources of sampling bias?
3. What is the goal of inferential statistics?
4. Null hypothesis predicts that independent variable have an effect on the dependent variable: T/F

**12.3 The meaning of statistical inference:**

Statistical inference is defined as the process inferring the properties of the given distribution based on the data. In other words, it deduces the properties of the population by conducting hypothesis testing and obtaining estimates from the data obtained from samples.

**12.3.1. Relationship between probability and statistical inference:**

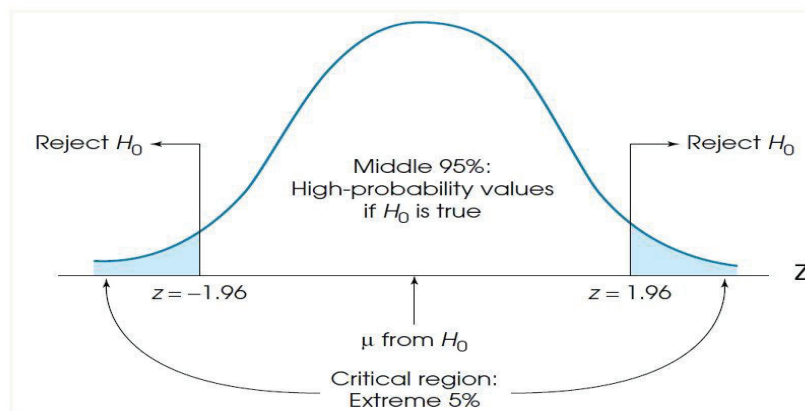
Statistics are, in one sense, all about probabilities. Inferential statistics deal with establishing whether differences or associations exist between sets of data. The data comes from the sample we use, and the sample is taken from a population. So we need to think about whether the sample represents the population from which it has been taken. The larger the sample we take the greater the probability that it is representative of the population. If we took the whole population for our study the probability would = 1, since the sample = the population. A sample smaller means that we cannot guarantee that it is similar to the population. In order for generalizability from sample to population to be meaningful, we aim to reduce sampling error.

Probability forms a direct link between samples and the populations from which they come. The following example provides a brief overview of probability can be used in the context of inferential statistics. The research begins with a population that forms a normal distribution with a mean of  $\mu = 400$  and a standard deviation of  $\sigma = 20$ . A sample is selected from the population and a treatment is administered to the sample. The goal for the study is to evaluate the effect of the treatment. To determine whether the treatment has an effect, the researcher simply compares the treated sample with the original population. If the individuals in the sample have scores around 400 (the original population mean), then we must conclude that the treatment appears to have no effect. On the other hand, if the treated individuals have scores that are noticeably different from 400, then the researcher has evidence that the treatment does have an effect. Notice that

**NOTES**

the study is using a sample to help answer a question about a population; this is the essence of inferential statistics.

The problem for the researcher is determining exactly what is meant by “noticeably different” from 400. If a treated individual has a score of  $X = 415$ , is that enough to say that the treatment has an effect? What about  $X = 420$  or  $X = 450$ ? One way of answering it is to use z scores. We know, from the normal probability distribution that a z-score value beyond  $z + 2.00$  (or  $-2.00$ ) was an extreme value and, therefore, noticeably different. But the choice of  $+2$  or  $-2$  was arbitrary. The other way to answer this question is to use *probability*, to help us decide exactly where to set the boundaries.



Alpha level of .05 -- the probability of rejecting the null hypothesis when it is true is no more than 5%.

Figure shows the setting of boundaries for alpha level of 0.05. The middle 95% of the distribution consist of high probability values if the null hypothesis were true and the extreme 5%, which when divided in half produces proportions of 0.0250 in the right-hand tail and 0.0250 in the left-hand tail. Using the unit normal table, we find that the z-score boundaries for the right and left tails are  $z = +1.96$  and  $z = -1.96$ , respectively. The boundaries set at  $z = +$  or  $- 1.96$  provide objective criteria for deciding whether our sample provides evidence that the treatment has an effect. Specifically, we use the sample data to help decide between the following two alternatives:

1. The treatment has no effect. After treatment, the scores still average  $\mu = 400$ .
2. The treatment does have an effect. The treatment changes the scores so that, after treatment, they no longer average  $\mu = 400$ .

As a starting point, we assume that the first alternative is true and the treatment has no effect. In this case, treated individuals should be no different from the individuals in the original population, which is shown in the above figure. If our assumption is correct, it is extremely unlikely

**NOTES**

(probability less than 5%) for a treated individual to be outside the  $\pm 1.96$  boundaries. Therefore, if we obtain a treated individual who is outside the boundaries, we must conclude that the assumption is probably not correct. In this case, we are left with the second alternative (the treatment does have an effect) as the more likely explanation. Notice that we are comparing the treated sample with the original population to see if the sample is noticeably different. If it is different, we can conclude that the treatment seems to have an effect. Now we are defining “noticeably different” as meaning “very unlikely.” We are using the sample data and the  $\pm 1.96$  boundaries, which were determined by probabilities, to make a general decision about the treatment. If the sample falls outside the boundaries we make the following logical conclusion:

- a. This kind of sample is very unlikely to occur if the treatment has no effect.
- b. Therefore, the treatment must have an effect that changed the sample.

On the other hand, if the sample falls between the  $\pm 1.96$  boundaries, we conclude:

- a. This is the kind of sample that is likely to occur if the treatment has no effect.
- b. Therefore, the treatment does not appear to have had an effect.

---

## 12.4 Methods of statistical Inference

---

**12.4.1 Statistical Estimation:** An important aspect of statistical inference is using estimates to approximate the value of an unknown population parameter. For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

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

**a) Point estimate:** A point estimate of a population parameter is a single value of a statistic. For example, the sample mean  $\bar{x}$  is a point estimate of the population mean  $\mu$ . Similarly, the sample proportion  $p$  is a point estimate of the population proportion  $P$ . Point estimates are rarely stated.

**b) Interval estimate:** An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example,  $a < x < b$  is an interval estimate of the population mean  $\mu$ . It indicates that the population mean is greater than  $a$  but less than  $b$ . Of course we may be wrong in supposing that the stated limits contain the population value. Other things being equal, if we set wide limits, the likelihood that the limits will include the population value is high and if we set narrow limits, there is a greater risk of being wrong. Because the

**NOTES**

option of setting wider or narrower limit exists, any statement of limits must be accompanied by indication of the probability that the limits contain the population parameters. The limits themselves are usually referred to as **confidence interval**. To express a confidence interval, you need three pieces of information.

- [Confidence level](#)
- [Statistic](#)
- [Margin of error](#)

Given these inputs, the range of the confidence interval is defined by the *sample statistic + margin of error*. And the uncertainty associated with the confidence interval is specified by the confidence level.

**Check your progress – 2**

5. What are confidence intervals?

**12.4.2 Statistical hypothesis testing:** Hypothesis testing is a statistical procedure that allows researchers to use sample data to draw inferences about the population of interest. The general goal of hypothesis testing is to determine whether or not a particular treatment has an effect on a population. The test is performed by selecting a sample, administering the treatment to the sample, and then comparing the result with the original population. If the treated sample is noticeably different from the original population, then we conclude that the treatment has an effect, and we reject null hypothesis ( $H_0$ ). On the other hand if the treated sample is still similar to the original population, then we conclude that there is no evidence for treatment effect and we fail to reject  $H_0$ .

The critical factor in this decision is the size of the difference between the treated sample and the original population. A large difference is evidence that the treatment worked, a small difference is not sufficient to say that the treatment has any effect. The hypothesis testing can either be in the one tailed or two tailed test format. In a directional hypothesis test or a one tailed test, the statistical hypothesis (null hypothesis ( $H_0$ ) and alternate hypothesis ( $H_1$ ) specify either an increase or a decrease in the population mean score that is, they make a statement about the directions of the effect.

**In very simple terms, the hypothesis-testing procedure is as follows:**

For the purpose of illustration, let's take the following example. Let's say our researcher selects a test for which adults older than 65 have an average score of  $\mu = 80$  with a standard deviation of  $\sigma = 20$ . The distribution of test scores is approximately normal. The researcher's plan is to obtain a sample of  $n = 25$  adults who are older than 65, and give each participant a daily dose of a blueberry supplement that is very high in antioxidants. After taking the supplement for 6 months, the participants are given the

**NOTES**

neuropsychological test to measure their level of cognitive function. If the mean score for the sample is noticeably different from the mean for the general population of elderly adults, then the researcher can conclude that the supplement does appear to have an effect on cognitive function. On the other hand, if the sample mean is around 80 (the same as the general population mean), the researcher must conclude that the supplement does not appear to have any effect.

**Step 1: State the hypothesis:**

First, we state a hypothesis about a population. Usually the hypothesis concerns the value of a population parameter. For the study in example , the null hypothesis states that the blueberry supplement has no effect on cognitive functioning for the population of adults who are more than 65 years old.

In symbols, this hypothesis is  $H_0: \mu$  (with supplement) = 80

For this example, the alternative hypothesis states that the supplement does have an effect on cognitive functioning for the population and will cause a change in the mean score. In symbols, the alternative hypothesis is represented as  $H_1: \mu$  with supplement  $\neq$  80

**Step 2: Set the criteria for decision:**

Eventually the researcher uses the data from the sample to evaluate the credibility of the null hypothesis. The data either provide support for the null hypothesis or tend to refute the null hypothesis. In particular, if there is a big discrepancy between the data and the null hypothesis, then we conclude that the null hypothesis is wrong. To formalize the decision process, we use the null hypothesis to predict the kind of sample mean that ought to be obtained. Specifically, we determine exactly which sample means are consistent with the null hypothesis and which sample means are at odds with the null hypothesis.

For our example, the null hypothesis states that the supplement has no effect and the population mean is still  $\mu = 80$ . If this is true, then the sample mean should have a value around 80. Therefore, a sample mean near 80 is consistent with the null hypothesis. On the other hand, a sample mean that is very different from 80 is not consistent with the null hypothesis. To determine exactly which values are “near” 80 and which values are “very different from” 80, we examine all of the possible sample means that could be obtained if the null hypothesis is true. For our example, this is the distribution of sample means for  $n = 25$ . According to the null hypothesis, this distribution is centered at  $\mu = 80$ . The distribution of sample means is then divided into two sections:

1. Sample means that are likely to be obtained if  $H_0$  is true; that is, sample means that are close to the null hypothesis

2. Sample means that are very unlikely to be obtained if  $H_0$  is true; that is, sample means that are very different from the null hypothesis

**The alpha level** To find the boundaries that separate the high-probability samples from the low-probability samples, we must define exactly what is meant by “low” probability and “high” probability. This is accomplished by selecting a specific probability value, which is known as the *level of significance*, or the *alpha level*, for the hypothesis test. The alpha ( $\alpha$ ) value is a small probability that is used to identify the low probability samples. By convention, commonly used alpha levels are  $\alpha = .05$  (5%),  $\alpha = .01$  (1%), and  $\alpha = .001$  (0.1%). For example, with  $\alpha = .05$ , we separate the most unlikely 5% of the sample means (the extreme values) from the most likely 95% of the sample means (the central values).

The extremely unlikely values, as defined by the alpha level, make up what is called the *critical region*. These extreme values in the tails of the distribution define outcomes that are not consistent with the null hypothesis; that is, they are very unlikely to occur if the null hypothesis is true. Whenever the data from a research study produce a sample mean that is located in the critical region, we conclude that the data are not consistent with the null hypothesis, and we reject the null hypothesis.

Technically, the critical region is defined by sample outcomes that are *very unlikely* to occur if the treatment has no effect (that is, if the null hypothesis is true). Reversing the point of view, we can also define the critical region as sample values that provide *convincing evidence* that the treatment really does have an effect. For our example, the regular population of elderly adults has a mean test score of  $\mu=80$ . We selected a sample from this population and administered a treatment (the blueberry supplement) to the individuals in the sample. What kind of sample mean would convince you that the treatment has an effect? It should be obvious that the most convincing evidence would be a sample mean that is really different from  $\mu=80$ . In a hypothesis test, the critical region is determined by sample values that are “really different” from the original population.

### Step 3: Collect data and compute sample statistics:

At this time, we select a sample of adults who are more than 65 years old and give each one a daily dose of the blueberry supplement. After 6 months, the neuropsychological test is used to measure cognitive function for the sample of participants. Notice that the data are collected *after* the researcher has stated the hypotheses and established the criteria for a decision. This sequence of events helps to ensure that a researcher makes an honest, objective evaluation of the data and does not tamper with the decision criteria after the experimental outcome is known. Next, the

**NOTES**

raw data from the sample are summarized with the appropriate statistic like  $z$  or  $t$ .

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\text{obtained difference between data \& hypothesis}}{\text{Standard distance expected by chance}}$$

**Step 4: Make a decision**

In the final step, the researcher uses the  $z$ -score value or  $t$  score obtained in step 3 to make a decision about the null hypothesis according to the criteria established in step 2. There are two possible outcomes:

1. The sample data are located in the critical region. By definition, a sample value in the critical region is very unlikely to occur if the null hypothesis is true. Therefore, we conclude that the sample is not consistent with  $H_0$  and our decision is to *reject the null hypothesis*. Remember, the null hypothesis states that there is no treatment effect, so rejecting  $H_0$  means that we are concluding that the treatment did have an effect.

2. The second possibility is that the sample data are not in the critical region. In this case, the sample mean is reasonably close to the population mean specified in the null hypothesis (in the center of the distribution). Because the data do not provide strong evidence that the null hypothesis is wrong, our conclusion is to *fail to reject the null hypothesis*. This conclusion means that the treatment does not appear to have an effect.

**Check your progress -3**

6. A researcher selects a sample of  $n=16$  individuals from a normal population with the mean of  $\mu=40$  and  $\sigma=8$ . A treatment is administered to the sample and, after treatment, the sample mean is  $M=43$ . If the researcher uses a hypothesis test to evaluate the treatment effect, What  $z$  score would be obtained for this sample?

7. A  $z$  score value in the critical region means you should accept the null hypothesis – T/F

8. A small value near zero for the  $z$  score statistic is evidence that a sample data are consistent with null hypothesis – T/F

**12.5 Let us sum up:**

In this Chapter, we introduced inferential statistics as a basis for drawing statistical inferences about the population from the sample data. Two main methods of statistical inference namely estimation and hypothesis testing has been covered. The two types of hypothesis and the procedure underlying hypothesis testing has been structured into a four step process for recall.



**12.6 Unit end Exercises:**

1. Explain why probability is the basis for statistical inference.
2. Discuss the logic underlying the hypothesis testing procedure.
3. An automobile manufacturer claims that a new model will average  $\mu = 45$  miles per gallon with  $\sigma = 4$ . A sample of  $n = 16$  cars is tested and averages only  $M = 43$  miles per gallon. Is this sample mean likely to occur if the manufacturer's claim is true? Specifically, is the sample mean within the range of values that would be expected 95% of the time? (Assume that the distribution of mileage scores is normal.)

**12.7 Answers to check your progress:**

1. Measures such as mean, median, variance etc descriptive of the sample are called statistics and measures descriptive of the population are called parameters.
2. An improper study design as well as a faulty data collection procedure can contribute to sampling bias.
3. The goal of inferential statistics is to infer or make conclusions about the population parameters based on sample statistics
4. False
5. A confidence interval, in statistics, refers to the probability that a population parameter will fall between two set values for a certain proportion of times. Confidence intervals measure the degree of uncertainty or certainty in a sampling method.
6. The standard error is 2 points and  $z = 3/2 = 1.50$
7. False
8. True

**12.8 Suggested Readings**

1. Garrett.H.E.(1981). *Statistics in Psychology and Education*. Vakils, Feffer and Simons Ltd.
2. Gravetter,J., &Wallnau,B.(1995). *Essentials of statistics* (2<sup>nd</sup> Ed.).West Publishing Company.
3. Gravetter,F.J., &Wallnau,L.B.(2013). *Statistics for behavioural sciences* (9<sup>th</sup> Ed.). Cengage Learning India Private Ltd.
4. Mangal S.K. (1987). *Statistics in Psychology and Education*. New Delhi: PHI Learning Private Ltd.

NOTES

---

## UNIT 13: THE RANDOM SAMPLING DISTRIBUTION

---

- 13.1 Population and Samples
- 13.2 The Distribution of sample means
- 13.3 Properties of the distribution of sample means
  - 13.3.1 The shape of the distribution of sample means
  - 13.3.2 The mean of the distribution of sample means: The expected value of  $M$
  - 13.3.3 The standard error of  $M$
- 13.4 Sampling distribution of differences between means
- 13.5 Introduction to  $t$  statistics
  - 13.5.1: Illustration of hypothesis testing with  $t$  statistics for Independent measures research design
  - 13.5.2 : Illustration of hypothesis testing with  $t$  statistics Repeated measures research design
- 13.6 Choice of alternative hypothesis: One tailed and two tailed tests of significance
- 13.7 Let us sum up
- 13.8 Unit end exercises
- 13.9 Answers to check your progress
- 13.10 Suggested Readings

---

### 13.1 Population and Samples

---

Before you begin, reading this unit, make sure you have refreshed your knowledge on random sampling,  $z$  scores, probability and normal distribution. Whenever a score is selected from a population, you should be able to compute a  $z$ -score that describe exactly where the score is located in the distribution. If the population is normal, you also should be able to determine the probability value for obtaining any individual score. In a normal distribution, for example, any score located in the tail of the distribution beyond  $z$  (+) or (-) 2.00 is an extreme value, and a score this large has a probability of only  $p = 0.0228$ . However, the  $z$ -scores and probabilities that we have considered so far are limited to situations in which the sample consists of a single score

Most research studies involve much larger samples, such as  $n = 25$  preschool children or  $n = 100$  American Idol contestants. In these situations, the sample mean, rather than a single score, is used to answer questions about the population. In this chapter we extend the concepts of  $z$ -scores and probability to cover situations with larger samples. In particular,

we introduce a procedure for transforming a sample mean into a z-score. Thus, a researcher is able to compute a z-score that describes an entire sample. As always, a z-score near zero indicates a central, representative sample; a z-score beyond 2.00 or  $-2.00$  indicates an extreme sample. Thus, it is possible to describe how any specific sample is related to all the other possible samples. In addition, we can use the z-scores to look up probabilities for obtaining certain samples, no matter how many scores the sample contains.

Suppose, for example, a researcher randomly selects a sample of  $n = 25$  students from the state college. Although the sample should be representative of the entire student population, there are almost certainly some segments of the population that are not included in the sample. In addition, any statistics that are computed for the sample are not identical to the corresponding parameters for the entire population. For example, the average IQ for the sample of 25 students is not the same as the overall mean IQ for the entire population. This difference, or error, between sample statistics and the corresponding population parameters is called **sampling error**. Furthermore, samples are variable; they are not all the same. If you take two separate samples from the same population, the samples are different. They contain different individuals, they have different scores, and they have different sample means. How can you tell which sample gives the best description of the population? Can you even predict how well a sample describes its population? What is the probability of selecting a sample with specific characteristics? These questions can be answered once we establish the rules that relate samples and populations

---

### 13.2 Distribution of sample means:

---

Two separate samples probably are different even though they are taken from the same population. The samples have different individuals, different scores, different means, and so on. In most cases, it is possible to obtain thousands of different samples from one population. With all these different samples coming from the same population, it may seem hopeless to try to establish some simple rules for the relationships between samples and populations. Fortunately, however, the huge set of possible samples forms a relatively simple and orderly pattern that makes it possible to predict the characteristics of a sample with some accuracy. The ability to predict sample characteristics is based on the distribution of sample means.

*The distribution of sample means is the collection of sample means for all of the possible random samples of a particular size ( $n$ ) that can be obtained from a population. Since a sampling distribution is a distribution of statistics obtained by selecting all of the possible samples of a specific size from a population, the distribution of sample means is an example of a sampling distribution. In fact, it often is called the sampling distribution of  $M$ .*

## NOTES

**NOTES**

If you actually wanted to construct the distribution of sample means, you would first select a random sample of a specific size ( $n$ ) from a population, calculate the sample mean, and place the sample mean in a frequency distribution. Then you would select another random sample with the same number of scores. Again, you would calculate the sample mean and add it to your distribution. You would continue selecting samples and calculating means, over and over, until you had the complete set of all the possible random samples. At this point, your frequency distribution would show the distribution of sample means.

What do you expect the distribution to look like? Since samples tend to be representative of the population, it is reasonable to expect the sample means should pile up around the population mean. The pile of sample means should tend to form a normal-shaped distribution as most of the samples should have means close to  $\mu$ , and it should be relatively rare to find sample means that are substantially different from  $\mu$ . The distribution of sample means is finally used to answer probability questions about sample means.

---

### **13.3 Properties of the distribution of sample means**

---

#### **13.3.1 The shape of the distribution of sample means**

It has been observed that the distribution of sample means tends to be a normal distribution. In fact, this distribution is almost perfectly normal, if either of the following two conditions is satisfied:

1. The population from which the samples are selected is a normal distribution.
2. The number of scores ( $n$ ) in each sample is relatively large, around 30 or more. (As  $n$  gets larger, the distribution of sample means more closely approximates a normal distribution. When  $n > 30$ , the distribution is almost normal, regardless of the shape of the original population. However when  $n < 30$ , it tends to take the form of “t” distribution)

The fact that the distribution of sample means tends to be normal is not surprising. Whenever you take a sample from a population, you expect the sample mean to be near to the population mean. When you take lots of different samples, you expect the sample means to “pile up” around  $\mu$ , resulting in a normal-shaped distribution.

#### **13.3.2 The mean of the distribution of sample means: The expected value of $M$**

The distribution of sample means is centered around the mean of the population from which the samples were obtained. In fact, the average value of all the sample means is exactly equal to the value of the population mean. This fact should be intuitively reasonable; the sample

means are expected to be close to the population mean, and they do tend to pile up around  $\mu$ . The formal statement of this phenomenon is that the mean of the distribution of sample means always is identical to the population mean. This mean value is called the expected value of  $M$ .

### 13.3.3 The standard error of $M$

To completely describe this distribution, we need one more characteristic, variability. The standard deviation of the distribution of sample means,  $M$ , is called the standard error of  $M$ . The standard error, like standard deviation serves the two purposes for the distribution of sample means.

1. *The standard error describes the distribution of sample means.* It provides a measure of how much difference is expected from one sample to another. When the standard error is small, then all of the sample means are close together and have similar values. If the standard error is large, then the sample means are scattered over a wide range and there are big differences from one sample to another. The magnitude of the standard error is determined by two factors:

(a) The size of the sample: In general, as the sample size increases, the error between the sample mean and the population mean should decrease. This rule is also known as the *law of large numbers*

(b) The standard deviation of the population from which the sample is selected: When the sample consists of a single score ( $n=1$ ), the standard error is the same as the standard deviation ( $M$ ).

2. Standard error measures how well an individual sample mean represents the entire distribution. Specifically, *it provides a measure of how much distance is reasonable to expect between a sample mean and the overall mean for the distribution of sample means.*

However, because the overall mean is equal to  $\mu$ , the standard error also provides a measure of how much distance to expect between a sample mean ( $M$ ) and the population mean ( $\mu$ ). Remember that a sample is not expected to provide a perfectly accurate reflection of its population. Although a sample mean should be representative of the population mean, there typically is some error between the sample and the population. The standard error measures exactly how much difference is expected on average between a sample mean,  $M$ , and the population mean,  $\mu$ .

#### Check your Progress - 1

1. How is sampling error different from standard error?
2. When does the distribution of sample means take the shape of a normal curve?

## NOTES

### Sampling distribution of differences between means:

#### NOTES

Until this point, all the inferential statistics we have considered involve using one sample as the basis for drawing conclusions about one population. Although these *single sample* techniques are used occasionally in real research, most research studies require the comparison of two (or more) sets of data. For example, a social psychologist may want to compare men and women in terms of their political attitudes, an educational psychologist may want to compare two methods for teaching mathematics, or a clinical psychologist may want to evaluate a therapy technique by comparing depression scores for patients before therapy with their scores after therapy. In each case, the research question concerns a mean difference between two sets of data. There are two general research designs that can be used to obtain the two sets of data to be compared:

1. The two sets of data could come from two completely separate groups of participants. For example, the study could involve a sample of men compared with a sample of women. Or the study could compare grades for one group of freshmen who are given laptop computers with grades for a second group who are not given computers.
2. The two sets of data could come from the same group of participants. For example, the researcher could obtain one set of scores by measuring depression for a sample of patients before they begin therapy and then obtain a second set of data by measuring the same individuals after 6 weeks of therapy.

The first research strategy, using completely separate groups, is called an *independent measures* research design or a *between-subjects* design. These terms emphasize the fact that the design involves separate and independent samples and makes a comparison between two groups of individuals. The second research strategy using same participants is called *repeated measures research design*.

Whether it is a single sample mean whose population mean we are hypothesizing or the difference two sample means, the procedure however remains the same. That is instead of having a single mean, if we have data related to differences between means of a number of samples, the sampling distribution of differences between these means will also look like a normal curve in case of large samples and like a **t distribution** in case of **small samples**. The standard deviation of this distribution is called standard error of the differences between means and may be taken as a yardstick against which we may safely test the obtained difference between sample means.

### 13.4 Introduction to t statistics :

It is usually impossible or impractical for a researcher to observe every individual in a population. Therefore, researchers usually collect data from a sample. There exist statistical procedures that permit researchers to use a sample mean to test hypothesis about a population. These statistical procedures are based on a few basic notions, which can be summarized as follows.

1. A sample mean ( $\bar{x}$ ) is expected more or less to approximate its population mean ( $\mu$ ). This permits us to use the sample mean to test a hypothesis about the population mean.
2. The standard error provides a measure of how well a sample mean approximates the population mean

$$\sigma_{\bar{x}} \text{ (standard error of the mean)} = \frac{\sigma}{n} \text{ (Standard deviation of}$$

the population)

3. To quantify our inference about the population, we compare the obtained sample mean ( $\bar{x}$ ) with the hypothesized population mean ( $\mu$ ) by computing a z-score statistic.

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\text{obtained difference between data \& hypothesis}}{\text{Standard distance expected by chance}}$$

The short coming of z score as an inferential statistic is that the z score formula requires more information than is usually available. Z scores require that we know the value of the population standard deviation which is needed to compute standard error. Without standard error, we have no way of quantifying the expected amount of distance (or error) between sample mean and population mean. We have no way of making precise, quantitative inferences about the population based on z - score.

The 't' statistic is relatively simple solution to the problem of not knowing the population standard deviation ( $\sigma$ ). When the value of  $\sigma$  is not known, we use the sample standard deviation in its place.

The sample standard deviation was developed specifically to be an unbiased estimate of the population standard deviation.

$$s = \sqrt{\frac{SS}{n-1}}$$

(Sample standard deviation)

SS = sum of squares

using this sample statistic we can estimate the standard error.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \text{ (sample standard deviation)}$$

### NOTES

**NOTES**

$S_{\bar{X}}$  = Estimated standard error

The estimated standard error ( $s_{\bar{X}}$ ) is used as an estimate  $\sigma_x$  when the value of  $\sigma$  (population standard deviation) is unknown. It is computed from the sample standard deviation and provides an estimate of the standard distance (expected by chance) between the sample mean ( $\bar{X}$ ) and the population mean ( $\mu$ ).

The 't' statistic is computed now by substituting the estimated standard error in the denominator of the z score formula.

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

The only difference between the 't' formula and z formula is that the z score formula uses the actual population standard deviation ( $\sigma$ ) and t statistic uses the sample standard deviation as an estimate when  $\sigma$  is unknown.

**Degrees of freedom and 't' statistic**

We must know the sample mean before we can compute the sample standard deviation. This places a restriction on sample variability such that only  $n-1$ , scores in a sample are free to vary. The value  $n-1$  is called the degrees of freedom (or df) for the sample standard deviation. The greater the value of df is for a sample, the better  $s$  represents  $\sigma$ , and the better the 't' statistic approximates the z - score. This should make sense because the larger the sample ( $n$ ) is, the better the sample represents its population. Thus, the degrees of freedom associated with a sample mean also describe how well "t" represents "z".

**The t distribution**

The t distribution is not a normal distribution. How well the 't' distribution approximates a normal distribution is determined by degrees of freedom. In general, the greater the sample size ( $n$ ) is, the larger the degrees of freedom ( $n-1$ ) are, better the t distribution approximates the normal distribution.

There is a different sampling distribution of 't' (a distribution of all possible sample t values) for each possible number of degrees of freedom. The 't' distribution has more variability than the normal distribution, especially when of values are small. The 't' distribution tends to be flatter and more spread out, whereas the normal z distribution has more of a central peak. This is because of the fact that the standard error in the 't' formula is not a constant because it is estimated. That is estimated standard error is based on the sample standard deviation, which will vary in value from sample to sample.

However when the value of df increases, the variability in the 't' distribution decreases and it more clearly resemble the normal



distribution because with greater df,  $s_x$  will more clearly estimate  $\sigma_x$ , and when df is very large, they are nearly the same.

### The 't' distribution table

The numbers in the 't' distributions table are the values of t that separate the tail from the main body of the distribution. Proportions for one or two tails are listed on the top of the table, and df values for "t" will be listed in the first column.

## NOTES

### 13.5.1 't' STATISTICS FOR AN INDEPENDENT MEASURES RESEARCH DESIGN

#### Hypothesis tests with the independent measures 't' statistics.

**Step – 1:** State the hypothesis.  $H_0$  &  $H_1$  and select an alpha level for the independent measures t test. The hypotheses concern the difference between the two population means.

$$H_0 = M_1 - M_2 = 0$$

$$H_1 = M_1 - M_2 \neq 0$$

The hypothesis testing procedure will determine whether or not the data provide evidence for a mean difference between the two populations. At the conclusion of the hypothesis test, we will decide either to.

- a) Reject  $H_0$  (we conclude that the data indicate a significant difference between the two population mean) or to
- b) Fail to reject  $H_0$  (The data do not provide sufficient evidence to conclude that the difference exist).

**Step 2:** Locate the critical region. The critical region is defined as sample data that would be extremely unlikely if the null hypothesis were true. In this case, we will be locating extremely unlikely "t" values.

**Step 3:** Get the data and compute the test statistic.

$$t = \frac{\text{Sample statistic} - \text{hypothesized population parameter}}{\text{Estimated standard error}}$$

**Step 4:** Make a decision: If the 't' statistic we compute is in the critical region, we reject  $H_0$ . Otherwise we conclude that the data do not provide sufficient evidence that the two populations are different.

#### Computation of difference between two means for small but independent samples

Example: Two groups of 10 students each got the following scores on an attitude scales.

**NOTES**

**Group I:** 10, 9, 8, 7, 7, 8, 6, 5, 6, 4

**Group II:** 9, 8, 6, 7, 8, 8, 11, 12, 6, 3

Compute the means for both groups and test the significance of the difference between these two means.

Sample 1			Sample 2		
X <sub>1</sub>	x <sub>1</sub>	x <sub>1</sub> <sup>2</sup>	X <sub>2</sub>	x <sub>2</sub>	x <sub>2</sub> <sup>2</sup>
10	3	9	7	1	1
9	2	4	8	0	0
8	1	1	6	-2	4
7	0	0	7	-1	1
7	0	0	8	0	0
8	1	1	8	0	0
6	-1	1	11	3	9
5	-2	4	12	4	16
6	-1	1	6	-2	4
4	-3	9	5	-3	9
Total = 70		Σx <sub>1</sub> <sup>2</sup> =30	Total = 80		Σx <sub>1</sub> <sup>2</sup> =44

$$M_1 = 70/10 = 7$$

$$M_2 = 80/10 = 8$$

$$\text{Pooled S.D or } \sigma = \sqrt{\frac{\sum x_1^2 + \sum x_2^2}{(N_1 - 1) + (N_2 - 1)}}$$

$$\sqrt{\frac{30 + 44}{9 + 9}} = 2.03$$

$$SE_D \text{ or } \sigma_D = \sigma \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

$$= 2.03 \sqrt{\frac{1}{10} + \frac{1}{10}}$$

$$= \frac{2.03}{2.2361} = 0.908$$

$$t = \frac{M_1 - M_2}{\sigma_D} = \frac{7 - 8}{0.908} = \frac{-1}{0.908} = -1.1$$

$$df = (N_1 + N_2) - 2$$

$$= 10 + 10 - 2$$

$$= 18$$

From the 't' distribution, we find that the critical value of 't' with 18 degrees of freedom at 5% level of significance is 2.10. One computed value of i.e. 1.1 is quite smaller than the critical table value 2.10 and hence is not significant. Therefore, the null hypothesis cannot be rejected and as a result, the given difference in sample means being insignificant, can only be attributed to some chance factors or sampling fluctuations.

### 13.5.2. 't' STATISTIC FOR REPEATED MEASURES RESEARCH DESIGN

A repeated measures study is one in which a single sample of individuals is tested more than once on the dependent variable. The same subject are used for every treatment condition.

Occasionally researchers will try to approximate the advantages of a repeated measures study by using a technique known as matched subjects. In matched subjects study, each individual in one sample is matched with a subject in the other sample. The matching is done so that the two individuals are equivalent with respect to a specific variable that the researcher would like to control.

The single sample 't' statistic is defined by the formula

$$t = \frac{\bar{X} - M}{s_x}$$

$\bar{X}$  = Sample mean

M = Population Mean

$s_x$  = Estimated standard error

The sample mean comes from the data. The standard error (also computed from the sample data), gives a measure of the error between the sample mean and the population mean, M.

For the repeated measures design, the sample data are difference scores and one identified by the letter D, rather than X. Therefore, we will substitute D's in the formula in place of X's to emphasize that we are dealing with difference scores instead of X values. Also, the population mean that is of interest to us is the population mean difference (the mean amount of change for the entire population). With these simple changes, the formula for the repeated measures design becomes,

$$t = \frac{\bar{D} - \mu_D}{s_D}$$

## NOTES

**NOTES**

$\mu_D$  = Hypothesized population difference Estimated standard error of difference

$s_{\frac{D}{n}}$  = Estimated standard error of difference

**Illustration:**

A researcher in behavioural medicine believes that stress often makes asthma symptom worse for people who suffer from this respiratory disorder. Because of the suspected role of stress the investigator decides to examine the effect of relaxation training on the severity of asthma symptoms. A sample of five patients is selected for the study. During the week before treatment, the investigator records the severity of their symptoms by measuring how many doses of medicine are needed for asthma attacks. Then the patients receive relaxation training. For the week following training; the researcher once again records the number of doses required by each patient. Do these data indicate that relaxation training alters the severity of symptoms.

Patient	Week before training	Week after training	D	D <sup>2</sup>
A	9	4	-5	25
B	4	1	-3	9
C	5	5	0	0
D	4	0	-4	16
E	5	1	-4	16
			$\Sigma D = -16$	$\Sigma D^2 = 66$

$$\bar{D} = \frac{\Sigma D}{n} = \frac{-16}{5} = -3.2$$

$$SS = \Sigma D^2 - \frac{(\Sigma D)^2}{n} = 66 - \frac{(-16)^2}{5}$$

$$= 66 - 51.2 = 14.8$$

$$s = \frac{\sqrt{SS}}{n-1} = \sqrt{\frac{14.8}{4}} = 1.92$$

Next, use sample standard deviation to compute the estimated standard error.

$$SD = \frac{s}{\sqrt{n}} = \frac{192}{\sqrt{5}} = 0.86$$

Finally, use the sample mean ( $\bar{D}$ ) and the hypothesized population mean along with the estimated standard error to compute the value for the 't' statistic.

$$t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}} = \frac{-3.2 - 0}{.86} = -3.72$$

Locate the critical region from the 't' distribution table, you should find that the critical values are  $\pm 2.776$ . The t value we obtained falls in the critical region the investigators rejects the null hypothesis and concludes that relaxation training does affect the amount of medication needed to control the asthma symptoms.

### Check your progress – 2

3. When do we employ independent measures research design?
4. When is a t statistic preferred over z?

---

### 13.6 Choice of Alternate Hypothesis: One tailed and Two tailed test of significance

---

The general goal of hypothesis testing is to determine whether a particular treatment has any effect on a population. The test is performed by selecting a sample, administering the treatment to the sample, and then comparing the result with the original population. Generally sample should represent the population and the extent to which sample statistics represent population parameters is an indication of the significance of the computed sample mean. If the treated sample is noticeably different from the original population, we reject  $H_0$  and conclude that computed sample mean is statistically significant then and that the treatment has an effect. On the other hand, if the treated sample is still similar to the original population, then we fail to reject  $H_0$  conclude that there is no convincing evidence for a treatment effect, and. The critical factor in this decision is the size of the difference between the treated sample and the original population. A large difference is evidence that the treatment worked; a small difference is not sufficient to say that the treatment had any effect.

**One tailed tests of significance:** In a directional hypothesis test, or a one-tailed test, the statistical hypotheses ( $H_0$  and  $H_1$ ) specify either an increase or a decrease in the population mean. That is, they make a statement about the direction of the effect. For example, a special training program is expected to increase student performance, or alcohol consumption is expected to slow reaction times. In these situations, it is possible to state the statistical hypotheses in a manner that incorporates the directional

### NOTES

**NOTES**

prediction into the statement of  $H_0$  and  $H_1$ . The result is a directional test, or what commonly is called a one-tailed test. The following example demonstrates the elements of a one-tailed hypothesis test.

**Two tailed tests of significance:** The term *two-tailed* comes from the fact that the critical region is divided between the two tails of the distribution. This format is by far the most widely accepted procedure for hypothesis testing. In case of two tailed test, we are interested if the difference sample means obtained from two populations is likely or unlikely. We are interested only in the magnitude of difference and not the direction of difference. Consequently when an experimenter wishes to test the null hypothesis,  $H_0 : M_1 - M_2 = 0$ , against its possible rejection and finds that it is rejected, he concludes that a difference really exist between the means. No assertion is made about the direction of difference. Hence two tailed tests are named non-directional tests of significance

The major distinction between one-tailed and two-tailed tests is the criteria that they use for rejecting  $H_0$ . A one-tailed test allows you to reject the null hypothesis when the difference between the sample and the population is relatively small, provided that the difference is in the specified direction. A two-tailed test, on the other hand, requires a relatively large difference independent of direction.

**Check your progress – 3**

5. What is directional test of significance?
6. What is the difference between one tailed and two tailed tests of significance?
7. If a researcher predicts that a treatment will increase scores, then the critical region for a one-tailed test would be located in the right-hand tail of the distribution. (True or false?)
8. If the sample data are sufficient to reject the null hypothesis for a one-tailed test, then the same data would also reject  $H_0$  for a two-tailed test. (True or false?)

**13.7 Let us sum up:**

The distribution of sample means is defined as the set of  $M_s$  for all the possible random samples for a specific sample size ( $n$ ) that can be obtained from a given population. The distribution of sample means is normal when the population from which the samples are selected is normal and when the size of the samples is relatively large ( $n = 30$  or more). The mean of the distribution of sample means is identical to the mean of the population from which the samples are selected. The standard deviation of the

distribution of sample means is called the standard error of  $M$ . The sampling distribution of means is then used to answer probability questions about sample means. Proceeding on the same lines, if we had two different samples taken from two different population, we would need to estimate standard error of difference between the means and use the sampling distribution of differences between means to answer probability questions about the differences in sample means.

The “ $t$ ” statistic was introduced as an alternate to  $z$  score for employing it in situation where the population standard deviation is unknown to the researcher, which most often is the case. Testing the significance of difference between means from two independent samples as well as related samples have been highlighted. In addition, directional and non –directional tests of significance have been presented. This shows that appropriate statistics need to be employed keeping in mind the research design of the study, for proper conclusions to be drawn about the population taken for study.

## NOTES

---

### 13.8 Unit end exercises

---

1. What is “ $t$ ” distribution? How is it different from  $z$  distribution?

2. What do you understand by the term degrees of freedom?

3. In a study examining overweight and obese college football players, Mathews and Wagner (2008) found that on average both offensive and defensive linemen exceeded the at-risk criterion for body mass index (BMI). BMI is a ratio of body weight to height squared and is commonly used to classify people as overweight or obese. Any value greater than 30 kg/m<sup>2</sup> is considered to be at risk. In the study, a sample of  $n = 17$  offensive linemen averaged  $M = 34.4$  with a standard deviation of  $s = 4.0$ . A sample of  $n = 19$  defensive linemen averaged  $M = 31.9$  with  $s = 3.5$ . Use an independent-measures  $t$  test to determine whether there is a significant difference between the offensive linemen and the defensive linemen. Use a two-tailed test with  $\alpha = .01$ .

4. A major oil company would like to improve its tarnished image following a large oil spill. Its marketing department develops a short television commercial and tests it on a sample of  $n = 7$  participants. People’s attitudes about the company are measured with a short questionnaire, both before and after viewing the commercial. The data are as follows:

Person	Before (X1)	After (X2)
A	15	15
B	11	13
C	10	18
D	11	12

**NOTES**

E	14	16
F	10	10
G	11	19

Was there a significant change? Note that participants are being tested twice—once before and once after viewing the commercial. Therefore, we have a repeated-measures design.

---

**13.9 Answers to check your progress:**


---

1. A sampling error is a statistical error that occurs when an analyst does not select a sample that represents the entire population of data and the results found in the sample do not represent the results that would be obtained from the entire population. The standard error of the sample mean is an estimate of how far the sample mean is likely to be from the population mean. It is the standard deviation of the distribution of sample means.
2. The distribution of sample means will take the shape of a normal curve when  $n$  is large and when the population from which the sample is drawn is normal.
3. The term independent refers to separate samples which are not related to one another. Independent measures research design is used to make a comparison between two groups of individuals. Example Boys Vs Girls on numerical ability.
4.  $t$  statistic is preferred over  $z$ , when the value of population standard deviation is unknown.
5. In a directional hypothesis test, the statistical hypotheses ( $H_0$  and  $H_1$ ) specify either an increase or a decrease in the population mean. That is, they make a statement about the direction of the effect.
6. The major distinction between one-tailed and two-tailed tests is the criteria that they use for rejecting  $H_0$ .
7. True. A large sample mean, in the right-hand tail, would indicate that the treatment worked as predicted.
8. False. Because a two-tailed test requires a larger mean difference, it is possible for a sample to be significant for a one-tailed test but not for a two-tailed test.



---

### 13.10 Suggested Readings

---

1. Haslam, S.M., & McGarty, C. (2003). *Research Methods and statistics in Psychology*. Sage Publications Ltd.
2. King, B.M., & Bear, G. (2001). *Statistical Reasoning in Psychology and Education* (3<sup>rd</sup> Ed.). John Wiley & Sons Inc.
4. Gravetter, F.J., & Wallnau, L.B. (2013). *Statistics for behavioural sciences* (9<sup>th</sup> Ed.). Cengage Learning India Private Ltd.

•

### NOTES

---

## UNIT 14: INTERPRETING THE RESULTS OF HYPOTHESIS TESTING

---

### NOTES

---

#### 14.1 A statistically significant difference versus practically important difference.

---

A statistical significance does not necessarily mean that the results are practically significant in a real-world sense of importance. The hypothesis testing procedure determines whether the sample results that you obtain are likely if we assume the null hypothesis is correct for the population. If the results are sufficiently improbable under that assumption, then we can reject the null hypothesis and conclude that an effect exists. In other words, the strength of the evidence in our sample has passed our defined threshold of the significance level ( $\alpha$ ) and hence the results are statistically significant. We use p-values to determine statistical significance in hypothesis tests. Consequently, it might seem logical that p-values and statistical significance relate to importance. However, that is false because conditions other than large effect size such as a large sample size and low sample variability can produce tiny p-values.

As the sample size increases, the hypothesis test gains greater statistical power to detect small effects. With a large enough sample size, the hypothesis test can detect an effect that is so miniscule that it is meaningless in a practical sense. When our sample data have low variability, hypothesis tests can produce more precise estimates of the population's effect. This precision allows the test to detect tiny effects. Statistical significance indicates only that you have sufficient evidence to conclude that an effect exists. It is a mathematical definition that does not know anything about the subject area and what constitutes an important effect. While statistical significance relates to whether an effect exists, practical significance refers to the magnitude of the effect. However, no statistical test can tell us whether the effect is large enough to be important in your field of study. Instead, we need to apply our subject area knowledge and expertise to determine whether the effect is big enough to be meaningful in the real world.

For example, suppose you are evaluating a training program by comparing the test scores of program participants to those who study on their own. Further, we decide that the difference between these two groups must be at least five points to represent a practically meaningful effect size. An effect of 4 points or less is too small to care about. After performing the study, the analysis finds a statistically significant difference between the two

groups. Participants in the study program score an [average](#) of 3 points higher on a 100-point test. While these results are statistically significant, the 3-point difference is less than our 5-point threshold. Consequently, our study provides evidence that this effect exists, but it is too small to be meaningful in the real world. The time and money that participants spend on the training program are not worth an average improvement of only 3 points. We can use confidence intervals to determine practical significance.

## NOTES

### Check your progress - 1

#### 1. What is statistical significance based on?

---

#### 14.2 Errors in hypothesis testing:

It is possible that the data will lead you to reject the null hypothesis when in fact the treatment has no effect. Remember: Samples are not expected to be identical to their populations, and some extreme samples can be very different from the populations that they are supposed to represent. If a researcher selects one of these extreme samples by chance, then the data from the sample may give the appearance of a strong treatment effect, even though there is no real effect.

Let's say a research study examines how a food supplement that is high in antioxidants affects the cognitive functioning of elderly adults. Suppose that the researcher selects a sample of  $n = 25$  people who already have cognitive functioning that is well above average. Even if the blueberry supplement (the treatment) has no effect at all, these people will still score higher than average on the neuropsychological test when they are tested after 6 months of taking the supplement. In this case, the researcher is likely to conclude that the treatment does have an effect, when in fact it really does not. This is an example of what is called a *Type I error*.

**Type I error occurs** when a researcher rejects a null hypothesis that is actually true. In a typical research situation, a Type I error means that the researcher concludes that a treatment does have an effect when, in fact, it has no effect. The problem is that the information from the sample is misleading. In most research situations, the consequences of a Type I error can be very serious. Because the researcher has rejected the null hypothesis and believes that the treatment has a real effect, it is likely that the researcher will report or even publish the research results. A Type I error, however, means that this is a false report. Thus, Type I errors lead to false reports in the scientific literature. Other researchers may try to build theories or develop other experiments based on the false results. A lot of precious time and resources may be wasted. The Probability of a Type I Error occurs when a researcher unknowingly obtains an extreme, non-representative sample. Specifically, the probability of a Type I error is equal to the  $\alpha$  (alpha) level. Occasionally, however, sample data are in the

## **NOTES**

critical region just by chance, without any treatment effect. When this occurs, the researcher makes a Type I error; that is, the researcher concludes that a treatment effect exists when in fact it does not. Fortunately, the hypothesis test is structured to minimize the risk that this will occur.

**Type II error** occurs when a researcher fails to reject a null hypothesis that is really false. In a typical research situation, a Type II error means that the hypothesis test has failed to detect a real treatment effect. A Type II error occurs when the sample mean is not in the critical region even though the treatment has had an effect on the sample. Often this happens when the effect of the treatment is relatively small. In this case, the treatment does influence the sample, but the magnitude of the effect is not big enough to move the sample mean into the critical region. Because the sample is not substantially different from the original population (it is not in the critical region), the statistical decision is to fail to reject the null hypothesis and to conclude that there is not enough evidence to say that there is a treatment effect. The consequences of a Type II error are usually not as serious as those of a Type I error. In general terms, a Type II error means that the research data do not show the results that the researcher had hoped to obtain. Unlike a Type I error, it is impossible to determine a single, exact probability for a Type II error. Instead, the probability of a Type II error depends on a variety of factors and therefore is a function, rather than a specific number.

### **Check your progress – 2**

2. What is the relationship between alpha level and Type I error?
3. Under what circumstances is a Type II error likely?

---

### **14.3 Statistical power**

---

An approach to determine the size or strength of a treatment effect is to measure the power of the statistical test. The power of a test is defined as the probability that the test will reject the null hypothesis if the treatment really has an effect. The power of a statistical test is the probability that the test will correctly reject a false null hypothesis. That is, power is the probability that the test will identify a treatment effect if one really exists. However, the second outcome, rejecting  $H_0$  when there is a real effect, is the power of the test. Researchers typically calculate power as a means of determining whether a research study is likely to be successful. Thus, researchers usually calculate the power of a hypothesis test before they actually conduct the research study. In this way, they can determine the probability that the results will be significant (reject  $H_0$ ) before investing time and effort in the actual research. To calculate power, however, it is first necessary to make assumptions about a variety of factors that

influence the outcome of a hypothesis test. Factors such as the sample size, the size of the treatment effect, and the value chosen for the alpha level can all influence a hypothesis test.

---

#### 14.4 Statistical decision making: Levels of Significance versus p – values

---

**Levels of significance:** To find the boundaries that separate the high-probability samples from the low-probability samples, we must define exactly what is meant by “low” probability and “high” probability. This is accomplished by selecting a specific probability value, which is known as the *level of significance*, or the *alpha level*, for the hypothesis test. The alpha ( $\alpha$ ) value is a small probability that is used to identify the low probability samples. By convention, commonly used alpha levels are  $\alpha = .05$  (5%),  $\alpha = .01$  (1%), and  $\alpha = .001$  (0.1%). For example, with  $\alpha = .05$ , we separate the most unlikely 5% of the sample means (the extreme values) from the most likely 95% of the sample means (the central values). The extremely unlikely values, as defined by the alpha level, make up what is called the *critical region*. These extreme values in the tails of the distribution define outcomes that are not consistent with the null hypothesis; that is, they are very unlikely to occur if the null hypothesis is true. Whenever the data from a research study produce a sample mean that is located in the critical region, we conclude that the data are not consistent with the null hypothesis, and we reject the null hypothesis.

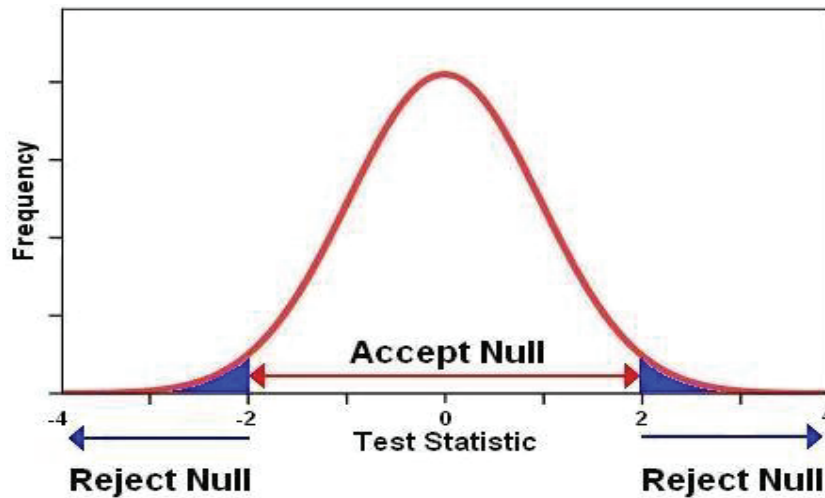
The alpha level, or the level of significance, is a probability value that is used to define the concept of “very unlikely” in a hypothesis test. The critical region is composed of the extreme sample values that are very unlikely (as defined by the alpha level) to be obtained if the null hypothesis is true. The boundaries for the critical region are determined by the alpha level. If sample data fall in the critical region, the null hypothesis is rejected. Technically, the critical region is defined by sample outcomes that are *very unlikely* to occur if the treatment has no effect (that is, if the null hypothesis is true). Reversing the point of view, we can also define the critical region as sample values that provide *convincing evidence* that the treatment really does have an effect.

**The boundaries for the critical region:** To determine the exact location for the boundaries that define the critical region, we use the alpha-level probability and the unit normal table. In most cases, the distribution of sample means is normal, and the unit normal table provides the precise z-score location for the critical region boundaries. With  $\alpha = .05$ , for example, the boundaries separate the extreme 5% from the middle 95%. Because the extreme 5% is split between two tails of the distribution, there is exactly 2.5% (or 0.0250) in each tail. Thus, for any normal distribution, the extreme 5% is in the tails of the distribution beyond  $z = +1.96$  and  $z = -$

## NOTES

## NOTES

1.96. These values define the boundaries of the critical region for a hypothesis test using  $\alpha=.05$ . Similarly, an alpha level of  $\alpha=.01$  means that 1%, or .0100, is split between the two tails. In this case, the proportion in each tail is .0050, and the corresponding z-score boundaries are  $z = + 2.58$  and  $z = - 2.58$ . Below is a figure showing the critical regions



**p-value** : A p-value is a probability. It is usually expressed as a proportion which can also be easily interpreted as a percentage:

- $P = 0.50$  represents a 50% probability or a half chance.
- $P = 0.10$  represents a 10% probability or a one in ten chance.
- $P = 0.05$  represents a 5% probability or a one in twenty chance.
- $P = 0.01$  represents a 1% probability or a one in a hundred chance.

P-values become important when we are looking to ascertain how confident we can be in accepting or rejecting our hypotheses. Because we only have data from a [sample](#) of individual cases and not the entire [population](#) we can never be absolutely (100%) sure that the [alternative hypothesis](#) is true. However, by using the properties of the [normal distribution](#) we can compute the probability that the result we observed in our sample could have occurred by chance. To clarify, we can calculate the probability that the effect or relationship we observe in our sample could have occurred through sampling variation and in fact does not exist in the population as a whole. The strength of the effect (the size of the difference between the mean scores), the amount of variation in scores (indicated by the [standard deviation](#)) and the sample size are all important in making the decision.

Conventionally, where *there is less than a 5% probability that the results from our [sample](#) are due to chance* the outcome is considered [statistically significant](#). Another way of saying this is that we are 95% confident there is a 'real' difference in our [population](#). This is our *confidence level*. We are therefore looking for a p-value that is less than .05, commonly written as  $p < .05$ . Results significant at the 1% level ( $p < .01$ ), or even the 0.1% level

( $p < .001$ ), are often called "highly" significant and if we want to be more sure of our conclusions you can set your confidence level at these lower values. It is important to remember these are somewhat arbitrary conventions - the most appropriate confidence level will depend on the context of the study.

The way that the [p-value](#) is calculated varies subtly between different statistical tests, which each generate a *test statistic* (called, for example, t, F or  $X^2$  depending on the particular test). This test statistic is derived from your data and compared against a known distribution (commonly a [normal distribution](#)) to see how likely it is to have arisen by chance. If the probability of attaining the value of the test statistic by chance is less than 5% ( $p < .05$ ) we typically conclude that the result is [statistically significant](#).

#### Check your progress - 4

4. When do we reject a null hypothesis?
5. What is the need for a power test?
6. If a sample mean is in the critical region with  $\alpha = .05$ , it would still (always) be in the critical region if alpha were changed to  $\alpha = .01$ . (True or false?)

---

#### 14.5 Let us sum up:

---

In this chapter we introduced chief ideas relating to interpretation of a hypothesis test. It is worthy to note that a statistically significant difference may not have any practical significance. Whatever decision is reached in a hypothesis test, there is always a risk of making the incorrect decision. There are two types of errors that can be committed A Type I error is defined as rejecting a true  $H_0$ . This is a serious error because it results in falsely reporting a treatment effect. The risk of a Type I error is determined by the alpha level and, therefore, is under the experimenter's control. A Type II error is defined as the failure to reject a false  $H_0$ . In this case, the experiment fails to detect an effect that actually occurred. The probability of a Type II error cannot be specified as a single value and depends in part on the size of the treatment effect. The power of a hypothesis test is defined as the probability that the test will correctly reject the null hypothesis.

---

#### 14.6 Unit End exercises:

---

1. Write short notes on errors in hypothesis testing with the suitable example.
2. What are probability values and how do you interpret them?

#### NOTES

## NOTES

---

### 14.7 Answers to check your progress

---

1. Statistical significance is the likelihood that a relationship between two or more variables is caused by something other than chance and is based on probability.
2. The alpha level for a hypothesis test is the probability that the test will lead to a Type I error. That is, the alpha level determines the probability of obtaining sample data in the critical region even though the null hypothesis is true
3. Type II error is likely when the effect of the treatment is relatively small. That is, the magnitude of the effect is not big enough to move the sample to the critical region.
4. We reject the null hypothesis when we find that our sample mean is in the critical region.
5. Power analysis is normally conducted before the data collection. The main purpose underlying power analysis is to help the researcher to determine the smallest sample size that is suitable to detect the effect of a given test at the desired level of significance. The reason for applying power analysis is that, ideally, the investigator desires a smaller sample because larger samples are often costlier than smaller samples.
6. False. With  $\alpha = .01$ , the boundaries for the critical region move farther out into the tails of the distribution. It is possible that a sample mean could be beyond the .05 boundary but not beyond the .01 boundary.

---

### 14.8 Suggested Readings

---

1. Gravetter, F.J., & Wallnau, L.B. (2013). *Statistics for behavioural sciences* (9<sup>th</sup> Ed.). Cengage Learning India Private Ltd.
2. Aron, Aron, E. N., Coups, E. (2012). *Statistics for Psychology*. (3<sup>rd</sup> Ed). Prentice Hall.
3. Garrett, H.E. (1981). *Statistics in Psychology and Education*. Vakils, Feffer and Simons Ltd.



**Distance Education – CBCS – (2018-19 Academic Year  
Onwards)  
MODEL QUESTION PAPER  
B.Sc(Psychology),  
11933 - PSYCHOLOGICAL STATISTICS**

*Model Question Paper*

**Time : 3 Hours  
Marks :75**

**PART – A (10X 2 = 20 Marks)**

**I. Answer all questions.**

1. What is a continuous variable?
2. What is a histogram?
3. Give the z score formula.
4. When do you use the mean?
5. What do you mean by the term Statistical inference?
6. Define Standard Error
7. Define Probability
8. Define variance
9. What is Contingency co-efficient?
10. Define ANOVA.

**PART – B (5X 5 = 25 Marks)**

**II . Answer all questions choosing either (a) or (b).**

11. a) Explain the importance of statistics.  
or  
b) Construct a frequency distribution table from the scores given below:

18	29	0	1	4	6	2	0
11	10	8	21	17	8	5	6
15	21	19	17	16	13	14	1
24	25	22	21	22	28	27	26

12. a) Compute Pearson's product moment correlation for the following data.

X- 10, 12, 14, 14, 16

- Y- 18, 20, 22, 17, 18 Compute mean and median from the following data.

Scores	Frequency
14-15	3
12-13	8
10-11	15
8-9	20
6-7	10
4-5	4

13. a) Explain the process of hypothesis testing.

or

b) . Compute Range, Average deviation, Standard deviation and Variance for the following data.

17, 15, 13, 15, 16, 14, 12, 18.

14. a) Explain scales of measurement with suitable examples.

or

b) List statistical software packages that are useful in analysing quantitative and qualitative research data

15. a) Explain one-tailed and two-tailed tests of significance.

Or

b) Explain errors in hypothesis testing

**PART – B (3X10 = 30 Marks)**

**III. Answer any 3 out of 5 questions.**

16. Explain Chi- square as a goodness of fit.

17. Give the null hypothesis for simple regression. How would you interpret the results of the simple regression?

18. A developmental psychologist is examining problem solving ability of grade school children. Each child is given a standardized problem solving task, and the psychologist records the number of errors. Use the data given to test whether there is a significant differences among the three age groups.

5 year olds	6 year olds	7 year olds
5	6	0
4	4	1
6	2	2

19. Explain measures of asymmetry.

20. a) At Smart university , all new freshmen are given an English proficiency test during the week of registration. Scores are normally distributed with the mean of 70 and SD of 10. The university has decided to place the top 25% into honors English and the bottom 20 % into remedial English. What scores separate the upper 25% and lower 20% of the students from the rest?

b) In a normal distribution of 1000 aptitude test scores with the mean of 60 and standard deviation of 8, How many scores fall i) Above 76 ii) Below 80 iii) Between 48 and 52.



