PGED S3-01
Exam. Code: ERES

Educational Research & Educational Statistics

SEMESTER- III

EDUCATION

BLOCK-2

KRISHNA KANTA HANDIQUI STATE OPEN UNIVERSITY

Subject Experts

1. Prof. Swarnalata Das, Dept. of Education, Gauhati University
2. Prof. Gayatree Goswamee, Dept. of Education, Gauhati University
3. Dr. Sadhana Goswami, Dept. of Education, Cotton College

Course Coordinator : Dr. Ritimoni Bordoloi

SLM Preparation Team

UNITS	CONTRIBUTORS
Units 9,11,12,13	Prof. Kaberi Saha, Gauahti University
Unit 10	Dr. Dilip Kumar Sarma, Cotton University
Unit 14	Dr. Jayashree Sharma, Rangia College
Unit 15 (I, II)	Dr. Mitali Deka, NERIM

Editorial Team

Content Editors : Units 9,11,12,13,14: Dr. Nilima Deka, LCB College
Unit 10: Harekrishna Deka, KKHSOU
Unit 15 (I, II): Retd. Prof. Tarakeswar Choudhury, Cotton College

Lanquage Editor : Professor Rabin Goswami, Former HOD,English, Department, Cotton College

Format Editor : Dr. Ritimoni Bordoloi, Krishna Kanta Handiqui State Open University

First Printed : January, 2022 **ISBN: 978-93-91026-81-3**

This Self Learning Material (SLM) of the Krishna Kanta Handiqui State Open University is made available under a Creative Commons Attribution-NonCommercial-ShareAlike4.0 License (international): http://creativecommons.org/licenses/by-nc-sa/4.0/

Printed and published by Registrar on behalf of the Krishna Kanta Handiqui State Open University.

Headquarters: Patgaon, Rani Gate, Guwahati 781017; Web : www.kkhsou.in/web/
City Campus : NH 37, Resham Nagar, Khanapara, Guwahati-781022, Assam, India

COURSE INTRODUCTION (COURSE 1, BLOCK 2)

BLOCK INTRODUCTION

In this Second Block of the Course, there are Seven units. *The Ninth Unit* deals with the importance of education and statistics and the important considerations for preparing a table. The Tenth unit explains the normal distribution and normal probability. *The Eleventh and the Twelfth Units* focuses on the measurement of central tendency and variability. The *Thirteenth Unit* calculates the co-efficient of correlation of educational statistics. The *Fourteenth Unit* explain the regression. The part I and Part II of the *Fifteenth Unit* explains the chi-square and ANOVA.

While going through a unit, you will notice some along-side boxes, which have been included to help you know some of the difficult, unseen terms. Some “ACTIVITY’ (s) have been included to help you apply your own thoughts. Again, we have included some relevant concepts in “LET US KNOW” along with the text. And, at the end of each section, you will get “CHECK YOUR PROGRESS” questions. These have been designed to self-check your progress of study. It will be better if you solve the problems put in these boxes immediately after you go through the sections of the units and then match your answers with “ANSWERS TO CHECK YOUR PPROGRESS” given at the end of each unit.

CONTENTS

Total Marks 100 **Page No.**

UNIT 9: Education and Statistics **157-172**

Meaning and importance of statistics, data analysis, important considerations for preparing a table and frequency distribution.

UNIT 10 : Normal Distribution **173-186**

Continuous Probability Distribution, Normal Distribution, Properties of Normal Distribution, Standard Normal Variate

UNIT 11 : Central Tendency **187-198**

Mean, median and mode and its merits and limitations.

UNIT 12 : Measures Variability **199-209**

Computing different measures of variability, merits and limitations of different variability.

UNIT 13 : Co-efficient of Correlation **210-218**

Calculations of Co-efficient of correlation, its merits and demerits.

UNIT 14 : Regression **219-227**

Meaning of Regression; Line of Regression equation; Correlation and Regression analysis

UNIT 15 (I) : Chi- Square Test **228-236**

Distribution, Applications of χ^2 Test, χ^2 Test of Goodness of Fit, χ^2 Test for Independence of Attributes, χ^2 Test for a Specified Population Variance

UNIT 15 (II): F- Distribution and Analysis of Variance **237-281**

Assumptions For F-Test, Properties Of F-Distribution, Applications Of F-Distribution, Hypothesis Testing for Difference between Two Population variances, Analysis Of Variance, Assumptions Of Analysis Of Variance, Classification Of Analysis Of Variance

UNIT 9: EDUCATION AND STATISTICS

UNIT STRUCTURE

9.1 Learning Objectives

9.2 Introduction

9.3 Meaning and Importance of Statistics

9.3.1 Importance of Statistics

9.3.2 Important Concepts in Statistics

9.3.3 Data Analysis

9.3.4 Grouping and tabulation of Data

9.3.5 Some Important Considerations for Preparing a Table

9.4 Frequency Distribution

9.5 Graphical Representation

9.5.1 Graphical Representation of Ungrouped Data

9.6 Let Us Sum Up

9.7 Further Reading

9.8 Answer to Check Your Progress

9.9 Possible Questions

9.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- discuss the meaning and importance of Statistics in Education
- describe frequency distribution
- draw a graphical representation of polygon and histogram.

9.2 INTRODUCTION

Statistics refers to numerical data or facts. However as a subject, 'Statistics' refers to that branch of knowledge which helps in the scientific collection, presentation, analysis and interpretation of numerical facts. The knowledge of statistics provides you a firm foundation which will help you to solve various statistical problems in Education, Psychology and Sociology. In this unit, let us discuss the meaning and importance of

statistics.

9.3 MEANING AND IMPORTANCE OF STATISTICS

The word statistics has the following important meanings –

- Statistics refers to numerical facts. For example, the numerical facts regarding birth and death, school attendance, employment market etc.
- It also signifies the method of dealing with numerical facts. Thus, statistics is the science of collecting, summarizing, analyzing and interpreting numerical facts.
- Again, statistics refers to the summarized figures of numerical facts like percentages, averages, means, medians, standard deviation etc.

In general, we can say that statistics is that branch of knowledge which helps us in scientific collection, presentation, analysis and interpretation of numerical facts.

9.3.1 Importance of Statistics

Do you know the importance of statistics for the students of Education, Psychology and Sociology? When you begin to deal with statistics, you may find that it is a very difficult task, but later on you will find how easy and useful it is. Let us discuss the importance of statistics in Psychology and Education.

Statistics, if clearly understood is a fascinating subject, which has tremendous importance in Psychology and Education and some of its importance can be summarized below –

Statistics helps:

- in collecting data either numerically or otherwise.
- in classification, organisation and summarization.
- in drawing general conclusions.
- Statistics also help in the task of evaluation and measurement.
- It helps us in the construction and standardization of test as well as using them properly.
- It helps the teacher to know the individual differences of the

students, comparing the suitability of one method or technique with another, making predictions for the future etc.

- It also helps us in drawing general conclusions.
- in the task of evaluation and measurement.
- in the construction and standardization of test as well as using them properly.
- to know the individual differences of the students, comparing the suitability of one method or technique with another, making predictions for the future etc.
- and in research work in many ways.

9.3.2 Important Concepts in Statistics

Data : Data represent the facts over a period of time, e.g. the minimum and the maximum temperature recorded in the city in a newspaper, the record of rainfall, the time of sunrise and sunset, etc.

Data may be of various forms–

The backbone of statistical inference is the theory of probability.

- **Qualitative Data :** When the data are classified on the basis of some attribute or quality such data is known as qualitative data. For example sex, colour of hair, literacy, religion etc. In this data, the attibute which is taken for the study, can not be measured, one can only find out whether the attribute is present or absent in the units of the population under the study.

  or when the data is classified on the basis of certain categories, such data is known as Quantitative data.

  For example– Management-wise Number of Schools

Management	No. of Schools
Government	4
Local body	8
Private aided	10
Private Unaided	2
	N = 24

- **Quantitative Data :** When the groups are based on numbers, such data is called Numerical or Quantitative Data, or whenever the measurement of a variable or data is possible on a scale in some appropriate units, it is called quantitative data. E.g. of quantitative data is age, height, income and intellectual ability. Age is measurable in years or months, height in cm, etc.

Height	No. of Students
Upto 4/5//	6
4/5// - 4/8//	15
4/8// - 5/11//	12
6/2// - 6/5//	8
	Total = 41

- **Continuous and Discrete Data :** Numerical or quantitative data may be continuous or discrete depending on the nature of the elements or objects being observed.

  Continuous data arise from the measurement of continuous variables or attributes in which individuals may differ by amounts just approaching zero. Weight and heights of children, temperature of a body, intelligence and achievement level of students etc. are the examples of continuous data.

CONTINUOUS AND DISCRETE DATA

Continuous		**Discrete**	
Height in cm	No. of person	No. of children per couple	No of couples
160 – 162	12	2	50
162 – 164	15	3	75
164 – 166	24	4	40
165 – 168	13	5	28
	N = 64		N = 193

Discrete data is characterised by gaps in the scale for which no real values may ever be found. Such data is usually expressed in whole numbers. e.g., size of the family, enrolment of the children,

number of books etc. are the examples of discrete data.

- **Primary and Secondary data :** The data collected by or on behalf of the person or people who are going to make use of the data refers to Primary Data. For e.g. – the attendance of children, the results of examinations conducted by you are the primary data.

  Sometimes an investigator may use the data already collected by some other person, such as the school attendance of children, or performance of students in various subjects for his or her study. Then the data is Secondary Data.

9.3.3 Data Analysis

Statistical data are analysed in two different ways –

- **Descriptive Statistics :** Descriptive statistics is concerned with describing or summarizing the numerical properties of data. The methodology of descriptive statistics includes classification, tabulation, graphical representation and calculation of certain indicators like mean, median, range etc. through which can be summarised certain important features of data.

  In this method, we cannot draw any conclusion, but we can provide worthy information regarding the nature of a specific group of individuals.
- **Inferential Statistics :** Inferential Statistics is also known as statistical inference. It is concerned with derivation of scientific inference about generalization of results from the study of a few particular cases. The methods of statistical inference help in generalizing the results of a sample as applied to the entire population from which the sample is drawn.

9.3.4 Grouping and Tabulation of Data

When a set of data contain only a few entries, a simple listing of the observations might be sufficient for interpreting the data. But if the

number of observations is too large, it is very difficult to list the observations simply. Here the data is usually organized into groups called classes and presented in a table which gives the number of observations, in each group. Such a table gives a better overall picture of the distribution of data and helps one to rapidly asses the important characteristics of the data.

The simplest way to organize the data is either in ascending way or descending way.

e.g. Heights in cms : 142, 156, 139, 148, 150, 149, 148, 144, 150, 152, 148, 149, 147, 141, 145. (N=15)

Arrangement in ascending way i.e. lowest to highest

139, 141, 142, 144, 145, 147, 148, 148, 148,
149, 149, 150, 150, 152, 156 (N=15)

Arrangement in descending way i.e. highest to lowest

156, 152, 150, 150, 149, 149, 148, 148, 148
147, 145, 144, 142, 141, 139 (N=15)

9.3.5 Some Important Considerations for Preparing a Table

Preparation of a table is an art. This should contain all the information needed within the smallest possible space. An ideal table should contain the following points –

- Table number : For easy reference and identification, a table should be numbered.
- Title : A good table should have a clearly worded, brief and unambiguous title describing the nature of data contained in the table.
- Captions : Captions in a table stand for brief and self explanatory vertical column. Captions may involve headings and sub headings as well.
- Stubs : Stubs stands for brief and self explanatory headings of horizontal rows. Normally, more important classification is given

in rows.

- Body : The body of the table contains the numerical information or frequency of observations in the different cells.
- Footnotes : Footnotes are given at the foot of the table for explanation of any fact or information, included in the table which needs some explanations.
- Sources of Data : Lastly, one should also mention the source of information from which the data is taken. This may include the name of the author, volume, page and the year of publication.

9.4 FREQUENCY DISTRIBUTION

Suppose we obtain a set of numerical data like

72, 75, 77, 67, 81, 68, 65, 86, 73, 67, 69, 82,

76, 76, 70, 83, 71, 63, 72, 72, 61, 84, 64, 67.

But this score does not convey any meaning. From this score, we may usually want to know several things. What is the average score? What is the amount of variability in the group, or how large are the individual differences are? These answer can be got by setting up a frequency distribution table. In frequency distribution table, we organize the collected data into some classes or groups and show the frequency in each group of scores.

How to construct a frequency distribution table : Following are the steps which we have to follow in constructing frequency distribution table.

In frequency distribution table, we organize the collected numerical data into some arbitrarily chosen classes or groups and try to show the frequency in each group of scores or class.

Step-I : Finding the Range : Range of the series of any score can be found out by subtracting the lowest score from the highest. In the above set of scores, the range of the distribution is (86-61)=25.

Step-2 : Determining the class interval : There are two different rules for determining the class interval. The First is – the range is divided by the number of class desired, e.g.

$$i = \frac{\text{range}}{\text{class desired}} \text{ Where, } i = \text{class interval}$$

$$\text{Here, } i = \frac{25}{5} = 5 \text{ and}$$

Second is — class interval 'r' is decided first and then the number of classes is determined. For this purpose usually, the class intervals of 2, 3, 5 or 10 can be chosen.

In the above example the range is 5 and the total number of scores is 25. Therefore i = 25 ÷ 5=5

Step-3 : Writing the contexts of the frequency distribution : Writing the contexts, we need to draw three columns (i) classes of scores, (ii) Tallying the scores into proper classes and (iii) checking the tallies. These have been shown in a tabular form in the table 1. In the above set of scores the range of the distribution is (86-61)=25.

Step-3 : Preparing the Table of frequency distribution.

I	II	III
Classes or Scores	Tallies	Frequencies

Step-4 :Checking the tallies : The total of the third column should be equal to the number of individuals whose scores have been tabulated. Here in Table 1, the total tallies are equal to the total number of scores.

- In the first column, we have to write down all the classes of the given distribution, preferably from the lowest to the highest as shown in the table 1.
- In the second column we write down scores given in the series on the basis of their respective classes which is known as tallies.
- These tally marks against each class are then counted. These counted numbers are called the frequencies of that class. They are written in column 3.

It is to be noted that, the total of the 3rd column should be equal to the number of total scores whose scores have been tallied.

In the above table, let us put the series of scores given in

Steps to construct a frequency distribution table.

1. To find out the range $R = \frac{\text{highest - lowest score}}{\text{class interval}}$
2. Determining the class interval
3. Preparing the table of frequency distribution.
   i) C.I.
   ii) Tallies
   iii) Frequencies

the beginning of this chapter according to the respective class interval in which the scores fall.

Let us take one example : Fifty students of B.A. (Major) course obtained the following scores in an achievement test. Tabulate the data in a frequency distribution table.

56	48	47	42	52	58	42	37	**68**	40
32	46	45	41	38	34	54	57	61	53
26	45	46	44	50	30	40	56	44	39
21	49	45	51	38	63	39	33	42	37
62	48	43	41	35	43	52	28	54	36

In the above example, the total no. of scores = 50,
Highest score = 68
Lowest score = 21
\ Range = 68-21 = 47
Hence, here 10 classes are sufficient.
\ i = 47 ¸ 10 = 4.7
(approx.5)

Table 1 : Frequency distribution table

Class Interval	Tallies	Frequencies
20-24	I	1
25-29	II	2
30-34	IIII	4
35-39	~~IIII~~ III	8
40-44	~~IIII~~ ~~IIII~~ I	11
45-49	~~IIII~~ IIII	9
50-54	~~IIII~~ II	7
55-59	IIII	4
60-64	III	3
65-69	I	I
		N = 50

9.5 GRAPHICAL REPRESENTATION

The data in the form of raw scores is known as ungrouped data, and when it is organized into a frequency distribution, it is known as grouped data. William Playfair, is the first person who introduce the presentation of the data in graphics method. Graphical methods has a number of advantages like–

- Graphs are visual aids which give a bird's eye view of a given set of numerical data. They present the data in a simple,readily comprehensive

form.

- Graphs are generally more attractive, fascinating and impressive than the sets of numerical data.
- They are extensively used to present statistical figures and facts.
- They register a meaningful impression on the mind almost before we think.
- They make the principal characteristics of groups and series visible at a glance.
- If the relationship between two variables is to be studied, graphic form of the data is a useful device. Graphs help us in studying the relation of one part to the other, and to the whole set of data.

9.5.1 Graphical Representation of Ungrouped Data

For the ungrouped data, we usually make use of the following graphical representations:

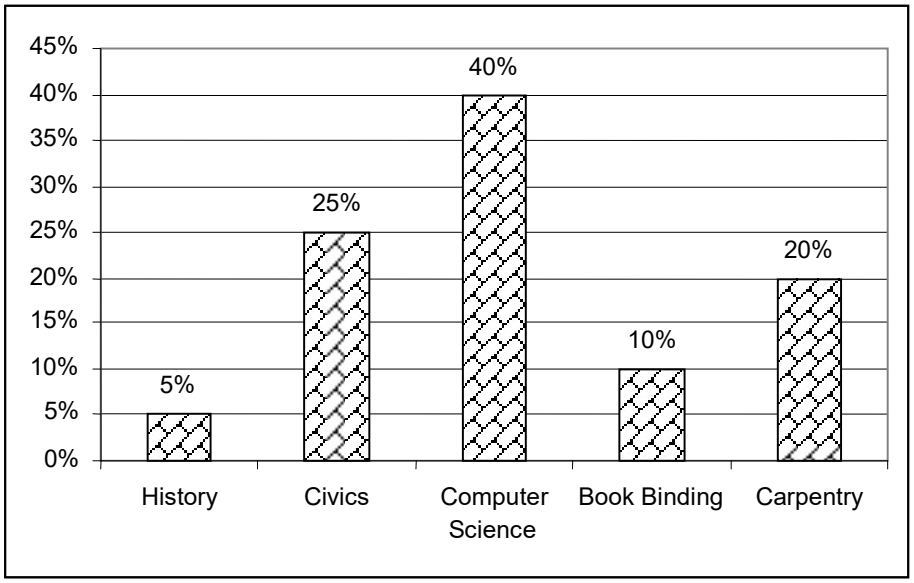
- Bar graph or Bar diagram
- Circle graph or Pie diagram
- Pictograms
- Line graphs.

Bar graph or Bar diagram : In bar graphs or diagram the data is represented by bars. Generally these diagrams or pictures are drawn on graph paper. Let us take the example: 60 class X students of a school were asked to opt for different optional papers. The details of these options are as follows:

Table 2

Subject	No. of students	Percentage
History	6	5%
Civics	30	25%
Computer Science	48	40%
Book Binding	12	10%
Carpentry	24	20%

Fig. 1 : Representation of the above data through a bar graph



Bar Graph : Percentage of students of class X opting for different Optional Subjects

Construction of Bar Diagram : In the construction of bar diagram, the length of the bars is kept proportional to the amount of variables or tract (height, weight, intelligence, cost etc.) possessed. The width of the bars is not governed by any set rules. It is an arbitrary factor. Regarding the space between two bars, it is conventional to have a space about one half of the width of a bar.

Circle Graph or Pie Diagram : In these types of graph, the data is represented through the sections or portions of a circle. Pie diagrams are very popularly used to denote percentage breakdown. It is so called because the entire graph looks like a pie. Pie diagrams are very useful in showing the expenditure of a govt. or of a company. It is also useful in teaching Geography, Science, etc.

How to Construct a Pie diagram : The following steps can be followed in the construction of a Pie diagram.

Steps :

- Draw a circle of appropriate size with a compass.
- Prepare the data in the form of percentage under different heads, the percentage for various sectors should be transposed into

Note:

1. Though graphs are more attractive and more understandable but, graphs cannot be prepared without tabulation.
2. Graphs cannot show so many facts at a time.
3. Information cannot be supplied in much detail. It only provides a quick picture of general situation.

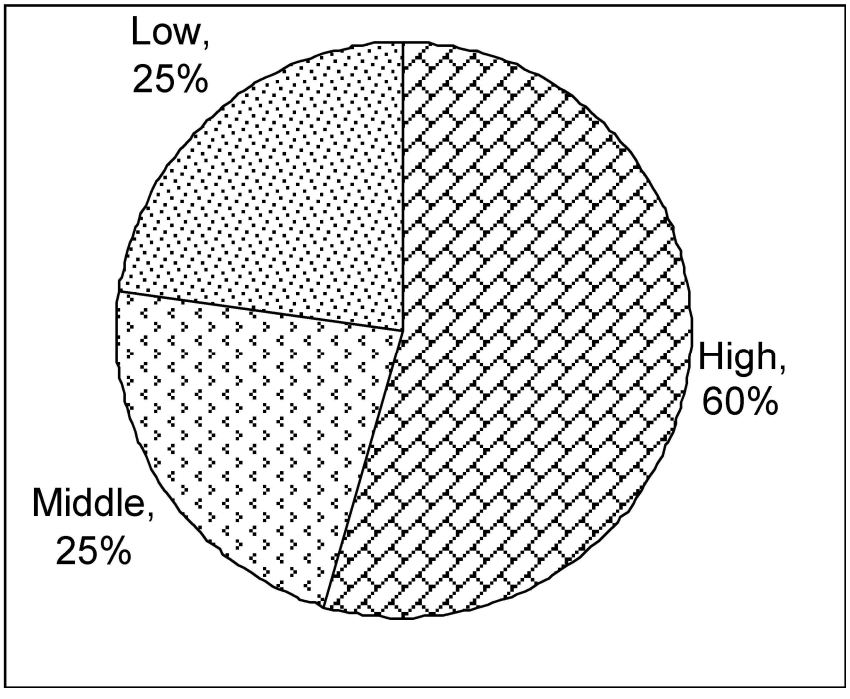
corresponding degrees on the circle.

For this purpose, the angle value of each sub part has to be found out. We know that the value of all the angles on any point is equal to 360o i.e. the entire circle is 360o which represents 100%.

$$\text{Thus, one \% means} = \frac{360^0}{100} = 3.6$$

- Suppose there are 3 components with the value 60% as high achievers, 25% as middle achievers and 15% as low achievers. Therefore they should represent 216o (60x3.6o), 90o(25x3.6o) and 54o(15x3.6o) respectively.
- When the values of all the angles have been determined, their total may not be exactly 360o due to approximation. If this is the case, some of the angle value may have to be slightly adjusted to make the total equal to 360o.
- Measure the points on the circle to represent the size of each sector with the help of a protractor. The labels may be placed inside the sector or outside the circle.

Fig. 2 : Distribution of Academic Achievements of Students through the Pie Diagram



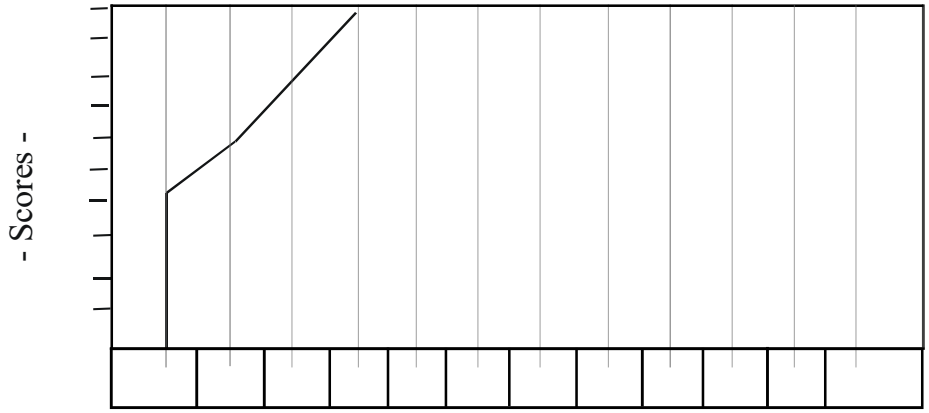
Line Graph : Line graphs are simple mathematical graphs that are drawn on the graph paper by plotting the data concerning one variable on the horizontal X-axis and other variable of the data on the vertical Y-axis. With the help of such graphs, the effect of one variable upon another variable during the experimental or normative study may be clearly demonstrated.

Example : A Nonsense syllables Association Test was administered on the students of class X to demonstrate the effect of practice on learning. The data so obtained may be studied from the following table.

Trial No. –

1	2	3	4	5	6	7	8	9	10	11	12
4	5	8	8	10	13	12	12	14	16	16	16

Fig. 3 : Trials Line graph – The effect of practice on learning





CHECK YOUR PROGRESS

Q 1: What is ment by the term statistics?

.....

.....

Q 2: Why is statistics important in education?

.....

.....

Q 3: Define the meaning of data.

.....

.....

Q 4: What steps will you follow in constructing a frequency distribution table?

.....

.....

Q 5: How can the ungrouped data be represented graphically?

.....

.....



9.6 LET US SUM UP

- The word 'Statistics' may refer to numerical facts, methodology of dealing with numerical facts, methodology of dealing with summarized figure of numerical facts. It has great importance in Education and Psychology.
- The term 'data' refers to numerical facts such as measures of height, weight, and scores on intelligence and achievement test.
- Statistical data may be organized in three different ways: Statistical tables, rank order and frequency distribution tables.
- In arranging data in rank order, we usually arrange it in the descending series by taking the highest score on the top and the lowest in the bottom.
- In organizing data through frequency distribution table, we group it into some arbitrarily chosen classes or groups, and try to show the frequency each group of scores or class.
- For the construction of frequency distribution table we first find out the range (highest – lowest score) and then determine the number of classes and class intervals by the following formula

$$i = \frac{\text{range}}{\text{no. of classes}}$$

- Statistics is helpful in all the social sciences, including Education and Psychology.

9.7 FURTHER READING

1) Garrett : H.E. (1956), Elementary Statistics, Longmans, Green and Co. New York.
2) Guildford, J.P. (1965): Fundamental Statistics in Psychology and Education, McGraw Hill Bokk Company, New York.

9.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: Statistics refers to numerical facts. As a branch of knowledge it refers to collection, presentation, analysis, and interpretation of data.

Ans to Q No 2: Statistics is very important in education as it helps in collecting, presenting analysis and interpreting data. It also helps in drawing general conclusion.

- Statistics helps in the task of measurement and evaluation
- Statistics is very useful in doing research work.

Ans to Q No 3: Data represent the facts over a period of time, e.g. the minimum and maximum temperature recorded, the record of rainfall etc.

Ans to Q No 4: The steps to prepare a frequency distribution table are:

- To find out the range
- To determine the class interval
- For preparing the Table of Frequency Distribution:
  (a) Class interval, (b) Tallies and (c) Frequency

Ans to Q No 5: The ungrouped data can be represented graphically through–

(a) Bar diagram (b) Circle graph or Pie diagram
(c) Pictograms (d) Line graphs etc.

9.9 POSSIBLE QUESTIONS

A. Short Questions (Answer each question in about 150 words)

Q 1: What do you mean by Statistics? What are its different functions?

Q 2: Describe the meaning of descriptive and inferential statistics.

Q 3: What are the importances of graphical representation of the data?

Q 4: Describe how you can represent the grouped and ungrouped data graphically.

Q 5: Distinguish between

i) Discrete and Continuous variable.

ii) Primary and Secondary data.

B. Long Questions (Answer each question in about 300-500 words)

Q 1: What are the essential factors or essentials of a good table?

Q 2: Tabulate the following twenty nine scores into a frequency distribution table taking a class interval of three.

68, 70, 75, 73, 70, 80, 77, 64, 85, 72, 67

82, 76, 76, 70, 83, 71, 63, 72, 61, 71

67, 84, 69, 64, 78, 82, 72, 75

Q 3: Draw a bar diagram to show the absolute as well as relative changes among the students of two colleges, A and B, in the different faculties from 2010-2013.

Faculties	College – A 2010 – 2013	College – B 2010 – 2013
Arts	100 150	130 150
Science	160 500	150 200
Commerce	100 450	130 150
Law	100 300	100 120

Q 4: Draw a Pie diagram to represent the following data -

Monthly expenditure for Grocery	3,800
Monthly expenditure for stationery etc.	1,200
Monthly expenditure for House rent, electricity, water, telephone bills etc.	10,000
Miscellaneous expenditure	5,000
Total	20,000

*** ***** ***

UNIT 10 : NORMAL DISTRIBUTION

UNIT STRUCTURE

10.1 Learning Objectives

10.2 Introduction

10.3 ContinuousProbability Distribution

10.4 Normal Distribution

- 10.4.1 Properties of Normal Distribution
- 10.4.2 Standard Normal Variate

10.5 Let Us Sum Up

10.6 Further Readings

10.7 Answers To Check Your Progress

10.8 Possible Questions

10.1 LEARNING OBJECTIVES

After going through this unit, you will able to-

- continuous probability distribution and probability deusity function
- normal distribution
- standard normal variate
- properties of normal distribution.

10.2 INTRODUCTION

Normal distribution is a continuous distribution. It is a limiting case of the binomial distribution. It has immense application. It is the most popular and commonly used distribution. It was discovered by Abrahoun de Moivre (1667–1754) in 1733.

10.3 CONTINUOUS PROBABILITY DISTRIBUTION

If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution.

A continuous probability distribution cannot be expresed in tabular form. An equation or formula is used to describe a continuous probability

distribution. The equation used to describe a continuous probability distribution is called a probability. density function (p.d.f.)

10.4 NORMAL DISTRIBUTION

Normal distribution is derived from the binomial distribution under two conditions, viz.

(i) n, the number of trials, is sufficiently large, i.e. $n \to \infty$

and (ii) neither p nor q is very small.

A random variable X is said to follow normal distribution, if and only if, its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where x is the real value of X, i.e. $-\infty < X < \infty$

The variable X is said to be distributed normally with mean μ and variance α^2 symbolically we write $X \sim N(\mu, \alpha^2)$. It is read as 'X is distributed as $N(\mu, \alpha^2)$'.

Note :

1. The probability density function given by (1) has two parameters, viz., μ and α.

   μ is any real number, i.e. $-\infty < \mu < \infty$ whereas α is a positive real number, i.e., $\alpha > 0$.

2. We know that probability can never be negative.

   $\therefore f(x) \geq 0, \forall x$

3. If $\mu = 0,\ \alpha = 1$ the p.d.f is given by

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$$

In this case $X \sim N(0, 1)$. In this case X is called the standard normal variate, and the p.d.f given by (2) is called standard normal distribution.

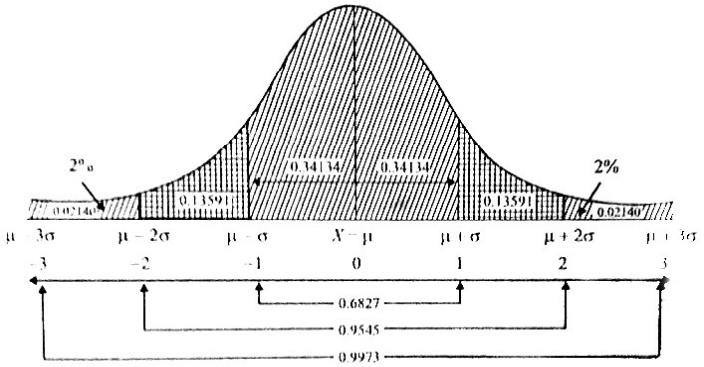
10.4.1 Properties of Normal Distribution

Normal distribution has many properties, some of which are given below.

1. The equation of the normal probability curve is

$$Y = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

It is of the shape given below. It is bell-shaped.



[Source : **Biostatistics** by Dr. P N Arora and Dr. P K Malhan (Himalya Publishing House, 2010)]

2. The curve is symmetrical about the line $x = \infty$, and x lies between $-\infty$ and $+\infty$

3. At $x = \infty$, the ordinate is maximum, and it is $Y = \frac{1}{\sigma\sqrt{2\pi}}$

4. The total area under the normal curve within the limits $-\infty$ to ∞ is 1, i.e.,

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\, dx = 1$$

5. The percentage distribution of area under the normal curve given below, and is shown in the figure given above.
   (i) About 68% of the area falls between $\mu - \alpha$ and $\mu + \alpha$
   (ii) About 95.5% of the area falls between $\mu - 2\alpha$ and $\mu + 2\alpha$.
   (iii) About 99.7% of the area falls between $\mu - 3\alpha$ and $\mu + 3\alpha$.

6. The curve has a single peak, i.e., it is unimodal.

7. The distribution is symmctrical. Mean = median = mode = μ.

8. The curve is asymptotic, i.e., it meets the x-axis at both the

ends at infinity. Theeretically the curve never towches the x-axis.

9. The points of inflexion are $x = \mu - \alpha$ and $x = \mu + \alpha$. [Inflexion points are where the function changes concavity.]
10. Coefficient of skewness = 0, and kurtosis = 3.
11. Mean deviation about mean $\simeq \frac{4}{5}\alpha$
12. $Q.D. \simeq \frac{2}{3}\alpha$

10.4.2 Standard Normal Variate

Let a random variable X be distributed as $N(\mu, \alpha^2)$. The location and shape of the normal curve depends on μ, and α, Where $-\infty < \mu < \infty$, and $\alpha > 0$. Therefore, no master table for the area can be prepared. This hindrance can be removed if we introduce a new variable Z defined by.

$$Z = \frac{X - \mu}{\alpha}$$

In this case Z is called the standard normal variate.

Now we prove that the mean of Z is 0, and the standard deviation is 1. Before that we state some related properties.

We know that mean of a random variable X is given by

mean $= E(X) = \Sigma x_i p_i$

We prove that

(i) E(a) = a; a is a constant

(ii) E (aX + b) = aE(x) + b; a, b are constants

(iii) $E\left(X - \overline{X}\right) = 0$

Proof : (i) $E(a) = \Sigma a p_i$

$= a\ \Sigma p_i$

= a, $\because\ \Sigma p_i = 1$

(ii) $\Sigma(aX + b) = \Sigma(ax_i + b)p_i$

$= a\Sigma x_i p_i + b\Sigma p_i$

$= a\,E(X) + b,\ \because\ \Sigma p_i = 1$

(iii) $E\left(X-\overline{X}\right)$

$= E(X) - E(\overline{X})$

$= \overline{X} - \overline{X}, \quad \because \overline{X}$ is constant

= 0

Also, we note that

(iv) Var (aX) = a^2 Var (X), where a is a constant

(v) Var (aX $\pm b$) = a^2 Var (X), where a, b are constants.

Now we prove that the mean of Z is 0, and the standard deviation is 1.

Mean = E (Z)

$= E\left(\frac{X-\mu}{\alpha}\right)$

$= \frac{1}{\alpha} E(X-\mu),$ property (ii)

$= \frac{1}{\alpha}\left[E(X)-\mu\right],$ property (ii)

$= \frac{1}{\alpha}[\mu-\mu]$

= 0

Variance = Var (Z)

$= Var\left(\frac{X-\mu}{\alpha}\right)$

$= \frac{1}{\alpha^2}\text{var}(X-\mu),$ property (v)

$= \frac{1}{\alpha^2}\text{var } X,$ property (iv)

$= \frac{1}{\alpha^2} . \alpha^2$

= 1

Thus, $Z = \frac{X-\mu}{\alpha}$ is distributed as N (0, 1), i.e., Z ~ N(0, 1).

All the properties for a normal variable hold good for standard normal variable $(\mu = 0,\ \alpha = 1)$

Let Z be a standard normal variate N(0, 1). The area under the standard normal curve is given by a table which is addended at the end of this chapter.

We know that the total area under the normal curve within the limits $-\infty$ to ∞ is 1. The standard normal curve is symmetrical about Z = 0. Therefore, the area to the left of z = 0 is 0.5, and the area to the right of z = 0 is 0.5.

Suppose, we are to find the area between z = 0, and z = 1.89, i.e, we are to find out $P(0 \le Z \le 1.89)$.

First we find 1.8 at the left of the table and single out that row.

Next we find 0.09 at the top of the table and single out that column.

Then we obtain the entry where the row for 1.8 and the column for 0.09 meet. the entry is 0.47062.

Thus $P(0 \le Z \le 1.89)$

= (the are from z = 0 to z = 1.89)

= 0.47062.

Example 1 : Find the probability of the standard normal variate z ehen z is greater than 1.09.

Solution : We know that the total area under the normal curve within the limits $-\infty$ to ∞ is 1. Therefore, the area to the right of z = 0 is 0.5, as the curve is symmetrical about z = 0.

$\therefore$ P(Z > 1.09) = 0.5 – 0.36214 = 0.13786

Example 2 : Find the probability of the standard normal variate z when z is less than 1.09.

Solution : We know that the total area under the normal curve within the limits $-\infty$ to $+\infty$ is 1. Therefore, the area to the left of z = 0 is 0.5, and the area to the right or z = 0 is 0.5.

Again, from the table, we get that the area from z = 0 to z = 1.09 is 0.36214.

$\therefore$ P(z < 1.09)

= (area to the left of z = 0) + (area from z = 0 to z = 1.09)

= 0.5 + 0.36214

= 0.86214

Example 3 : Find P(0.5 < z < 1.09)

Solution : We know that the area to the right of z = 0 is 0.5. From the table, the area from z = 0 to z = 0.5 is 0.19146. Again, the area from z = 0 to z = 1.09 is 0.36214.

$\therefore$ P(0.5 < z < 1.09)

= (Area from z = 0 to z = 1.09) – (area from z = 0 to x = 0.5)

= 0.36214 – 0.19146

= 0.17068

Example 4 : A large number of measurements is normally distributed with mean 65.5 cm and standard deviation 6.2 cm. Find the percentage of measurements that fall between 54.8 cm and 68.8 cm.

Solution : Let X be the normal variate showing the large number of measurements.

We have $\mu = 65.5,\ \alpha = 6.2$

Now, $Z = \frac{X - \mu}{\alpha} = \frac{X - 65.5}{6.2}$

When, X = 54.8 $\quad Z = \frac{54.8 - 65.5}{6.2} = \frac{-10.7}{6.2} = -1.7$

When, X = 68.8, $\quad Z = \frac{68.8 - 65.5}{6.2} = 0.5$

Thus, when X lies between 54.8 and 68.8, Z lies between – 1.7 and 0.5

Now P(–1.7 < X < 0.5)

= P(–1.7 < Z < 0) + P(0 < Z < 0.5)

= 0.45543 + 0.19146

= 0.64689 = 64.69%

$\therefore$ the percentage of measurements is 64.69%

Example 5 : For the distribution of makrks of candidates in an examination, mean is 14 and standard deviation is 2.5. Assuming the normality of the distribution, find the probability that a randomly

selected student will score above 15.

Solution : Let X be the normal variate showing the scores of candidates.

We have $\mu = 14,\ \alpha = 2.5$

Now, $Z = \frac{X - \mu}{\alpha} = \frac{X - 14}{2.5}$

When X, $Z = \frac{15 - 14}{2.5} = \frac{1}{2.5} = 0.4$

$\therefore$ P(x > 15) = P(z > 0.4)

= 0.5 – (Area between z = 0 and Z = 0.4)

= 0.5 – 0.15542

= 0.34458.



CHECK YOUR PROGRESS 1

Q 1: Find (i) P (Z > 2.58), (ii) P(Z < 1.96), (iii) P (Z > –3)

Q 2: For a certain normal distribution, mean is 50, and standard deviation is 5. Find $P(45 \le X \le 55)$, X being a normal variate.

Q 3: A normal variate X has mean 100 and standard deviation 4. Find the probability that X lies between 92 and 10.8

Q 4: The distribution of scores of students in an examination is normal. If the mean is 45, and the standard deviation is 5, find the percentage of students who scored between 40 and 50.

Q 5: The marks obtained in a certain examination follow normal distribution with mean 45, and standard deviation 10. If 1000 students appear at the examination, calculate the number of students scoring (i) less than 30 marks, and (ii) more thaw 80 marks.



10.5 LET US SUM UP

- Normal distribution is a continous distribution. It is a limiting case of the binomial distribution.
- Normal distribution is derived from the binomial distribution under two conditions, viz.,
  - (i) n, the number of trials, is sufficiently large, i.e., $n \to \infty$ and
  - (ii) neither p nor q is very small.

  A random variable X is said to follow normal distribution, if and only if, its probability density function is given by

  $$f(x) = \frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\alpha}\right)^2}$$

  Where x is the real value of X, i.e., $-\infty < X < \infty$

  The variable X is said to be distributed normally with mean μ and variance α^2. Symbolically, we write $X \sim N(\mu, \alpha^2)$. It is read as 'X is distributed as $N(\mu, \alpha^2)$.
- The normal curve is bell-shaped, and the curve is symmetrical about the line $x = \mu$, and x lies between $-\infty$, and $+\infty$.
- The total area under the normal curve within the limits $-\infty$ to ∞ is 1.
- If a new variable Z is defined by $Z = \frac{X - \mu}{\alpha}$, Z is called the standard normal variate. The mean of Z is 0, and the standard deviation is 1.



10.6 FURTHER READING

1) Blommers. P. and Lindquist. E. F. (1959), Elementary Statistical Methods. Houghton Mifflin, Company, Boston.

2) Gieldford, J. P. (1956). Fundamental Statistics in Psychology and Education, McGraw Hill Book Company, New York.

3) Rath, R. K. (1999). Fundamentals of Educational Statistics and Measurement. Taratarini Pustakalaya, Orissa.

4) Srivastava, A.B.L. and Sarma, K. K. (1985). Elementary Statistics in Psychology and Education. Sterling Publishers Pvt. Ltd., New Delhi.



10.7 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) We know that tha total area under the normal curve within the limits $-\infty$ to $+\infty$ is 1. Therefore, the area to the left of z = 0 is 0.5, and the area to the night of z = 0 is 0.5. Again, from the table, we get that the area from z = 0 to z = 2.58 is 0.49506

$\therefore$ P (z > 2.58)

= (area to the right of z = 0) – (area from z = 0 to z = 2.58)

= 0.5 – 0.49506

= 0.00494

(ii) We know that the total area under the normal curve within the limits $-\infty$ to $+\infty$ is 1. Therefore, the area to the left of z = 0 is 0.5, and the area to the right of z = 0 is 0.5. Again from the table, we get that the area from z = 0 to z = 1.96 is 0.47500.

$\therefore$ P (z < 1.96)

= (area to the left of z = 0) + (area from z = 0 to z = 1.96)

= 0.5 + 0.47500

= 0.97500

(iii) We know that the total area under the normal curve within the limits $-\infty$ to ∞ is 1. Therefore, the area to the left of z = 0 is 0.5, and the area to the right of z = 0 is 0.5.

Now, P(Z > –3)

= area from z = –3 to z = 0) + (area to the right of z = 0)

= (area from z = 0 to z = 3) + (area to the right of z = 0)

= 0.49865 + 0.5

= 0.99865

Ans to Q No 2: X is a normal variate.

We have $\mu = 50,\ \alpha = 5$

Now, $Z = \frac{X - \mu}{\alpha} = \frac{X - 50}{5}$

When X = 45 $Z = \frac{45 - 50}{5} = -1$

When X = 55, $Z = \frac{55 - 50}{5} = 1$

Thus, when X lies between 45 and 55, Z lies between –1 and 1.

Now $P(-1 \le Z \le 1)$

$= P(-1 \le Z \le 0) + P(0 \le Z \le 1)$

= 0.34134 + 0.34134

= 0.68268

Ans to Q No 3: Do yourself.

Ans to Q No 4: Do yourself.

Ans to Q No 5: Let X be the normal variate showing marks.

We have $\mu = 45,\ \alpha = 10$

Now $Z = \frac{X - \mu}{\alpha} = \frac{X - 45}{10}$

When X = 30, $Z = \frac{30 - 45}{10} = \frac{-15}{10} = -1.5$

When X = 80, $Z = \frac{80 - 45}{10} = \frac{35}{10} = 3.5$

(i) Now $P(X \le 30)$

= (area to the left of z = 0) – (area from z = – 1.5 to z = 0)

(area to the left of z = 0) – (area from z = 0) to z = 1.5)

= 0.5 – 0.43319

= 0.06681

$\therefore$ the required number of students

= 1000 x 0.06681

= 67 (approximately)

(ii) $P(Z \geq 80)$

$= P(Z \geq 3.5)$

= (area to the right of z = 0) – (area from z = 0 to z = 3.5)

= 0.5 – 0.49977

= 0.00023

Now 1000 x 0.00023 = 0.23

$\therefore$ there are no students scoring more than 80 marks.



10.8 POSSIBLE QUESTIONS

Q 1: Define standard normal variate and write down the probability density function of this variate.

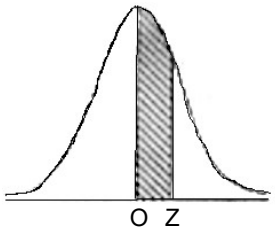
Q 2: State the conditions under which the normal distribution is obtained as a dimiting case of the binomial distribution.

Q 3: Find (i) $P(Z \leq -1.96)$ (ii) $P(Z \leq 1.96$ (iii) $P(Z > 2.58)$

Q 4: A normal variate X has mean 50 and standard deviation 5. Find $P(40 \leq X \leq 60)$.

Q 5: The mean of a normal distribution is 50 and the standard deviation is 4. What percentage of total frequency lies between 42.16 and 57.84?

Q 6: Let a continuous variable X follow a normal distribution with mean 12 and standard deviation 2. What is the probability that the value of X selected at random lies between 11 and 14 ?



Area under the Standard Normal Curve

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.00000	.00399	.00798	.01197	.01595	.01994	.02392	.02790	.03188	.03586
0.1	.03983	.04383	.04776	.05172	.05567	.05962	.06356	.06749	.07142	.07535
0.2	.07926	.08317	.08706	.09095	.09483	.09871	.10257	.10642	.11026	.11409
0.3	.11791	.12172	.12552	.12930	.13307	.13683	.14058	.14431	.14803	.15173
0.4	.15542	.15910-	.16276	.16640	.17003	.17364	.17724	.18082	.18439	.18793
0.5	.19146	.19497	.19847	.20194	.20540	.20884	.21226	.21566	.21904	.22240
0.6	.22575	.22907	.23237	.23565	.23891	.24215	.24537	.24857	.25175	.25490
0.7	.25804	.26115	.26424	.26730	.27035	.27337	.27637	.27935	.28230	.28524
0.8	.28814	.29103	.29389	.29673	.29955	.30234	.30511	.30785	.31057	.31372
0.9	.31594	.31859	.32121	.32381	.32639	.32894	.33147	.33398	.33646	.33891
1.0	.34134	.34375	.34614	.34849	.35083	.35314	.35543	.35769	.35993	.36214
1.1	.36433	.36650	.36864	.37076	.37286	.37493	.37698	:37900	.38100	.38298
1.2	.38493	.38686	.38877	.39065	.39251'	.39435	.39617	.39796	.39973	.40147
1.3	.40320	.40490	.40658	.40824	.40988	.41198	.41309	.41466	.41621	.41774
1.4	.41924	.42073-	.42220	.42364	.42507	.42647	.42785	.42922	.43056	.43189
1.5	.43319	.43448	.43574	.43699	.43822	.43943	.44062	.44179	.44295	.44408
1.6	.44520	.44630	.44738	.44845	.44950	.45053	.45154	.45254	.45352	.45449
1.7	.45543	.45637	.45728	.45818	.45907	.45994	.46080	.46164	.46246	.46327
1.8	.46407	.46485	.46562	.46638	.46712	.46784	.46856	.46926	.46995	.47062
1.9	.47128	.47193	.47257	.47320	.47381	.47441	.47500	.47558	.47615	.47670
2.0	.47725	.47784	.47831	.47882	.47932	.47982	.48030	.48077	.48124	.48169
2.1	.48214	.48257	.48300	.48341	.48382	.48422	.48461	.48500	.48537	.48574
2.2	.48610	.48645	.48679	.48713	.48745	.48778	.48809	.48840	.48870	.48899

2.3	.48928	.48956	.48983	.49010	.49036	.49061	.49086	.49111	.49134	.49158
2.4	.49180	.49202	.49224	.49245	.49266	.49286	.49305	.49324	.49343	.49361
2.5	.49379	.49396	.49413	.49430	.49446	.49461'	.49477	.49492	.49506	.49520
2.6	.49534	.49547	.49560	.49573	.49585	.49598	.49609	.49621	.49632.	.49643
2.7	.49653	.49664	.49674	.49683	.49693	.49702	.49711	.49720	.49728	.49736
2.8	.49744	.49752	.49760	.49767	.49774	.49781	.49788	.49795	.49801	.49807
2.9	.49813	.49819	.49825	.49831	.49836	.49841	.49846	.49851	.49856	.49861
3.0	.49865	.49869	.49874	.49878	.49882	.49886	.49889	.49893	.49896	.49900
3.1	.49903	.49906	.49910	.49913	.49916	.49918	.49921	.49924	.49926	.49929
3.2	.49931	.49934	.49936	.49938	.49940	.49942	.49944	.49946	.49948	.49950
3.3	.49952	.49953	.49955	.49957	.49958	.49960	.49961	.49962	.49964	.49965
3.4	.49966	.49968	.49969	.49970	.49971	.49972	.49973	.49974	.49975	.49976
3.5	.49977	.49978	.49978	.49979	.49980	.49981	.49981	.49982	.49983	.49983
3.6	.49984	.49985	.49985	.49986	.49986	.49987	.49987	.49988	.49988	.49989
3.7	.49989	.49990	.49990	.49990	.49991	.49991	.49992	.49992	.49992	.49992
3.8	.49993	.49993	.49993	.49994	.49994	.49994	.49994	.49995	.49995	.49995
3.9	.49995	.49995	.49996	.49996	.49996	.49996	.49996	.49996	.49997	.49997
4.0	.49997	-	-	-	-	-	-	-	-	-
4.5	.499997	-	-	-	-	-			-	-
5.0	.4999997	-	-	-	-	-	-	-	-	-

[Source : Basic statisties by B. L. Agarwal (New Age International Publishers, 2009)]

*** ***** ***

UNIT 11 : CENTRAL TENDENCY

UNIT STRUCTURE

11.1 Learning Objectives

11.2 Introduction

11.3 Central Tendency

- 11.3.1 The Mean
- 11.3.2 The Median
- 11.3.3 The Mode

11.4 Merits and Limitations of The Mean, The Median and The Mode.

11.5 Let Us Sum Up

11.6 Answers to Check Your Progress

11.7 Further Reading

11.8 Possible Questions

11.1 LEARNING OBJECTIVES

After going through this unit you will be able to:

- explain the meaning of central tendency: Mean, Median and Mode.
- identify the merits and demerits of Mean, Median and Mode.

11.2 INTRODUCTION

We have already introduced you to the meaning of statistics, its importance in education, organisation of data and normal distribution etc. in UNIT 9 and in UNIT 10. In this Unit, we shall deal with the meaning of the Central tendency. Let us see what the measures of the central tendency are:

11.3 CENTRAL TENDENCY

Central Tendency is a sort of average or a typical value of the items in the series or some characteristics of the members of a group. Suppose,

we take the achievement scores of the students of a class and then arrange them in frequency distribution. We will find that most of the students score will center round in the middle and few will score either very high or very low. This tendency of the scores is known as **central tendency**.

The most common measures of central tendencies are

- The Arithmetic Mean
- The Median
- The Mode.

11.3.1 Mean (M)

Mean is nothing but the most useful measures of central tendency. It is the average. It is the sum of all values of the items in a series divided by the total number of the items.

The calculation of the mean can be done by using the following formula.

1. **Mean for ungrouped data (M)**

   $M = \frac{\Sigma X}{N}$, here M = mean,

   X = score in a distributes

   Σ = Sum of

   N = Total Number of scores

   e.g. We can find out the average performance of 5 boys securing marks respectively 64, 80, 42, 70, 51 as

   $$M = \frac{64 + 80 + 42 + 70 + 51}{5}$$

   = 61.4

2. **Mean for grouped data :** Let us see how we can find out the mean for grouped data.

Mean is Average
Formula for Mean (M)

1. $M = \frac{\Sigma X}{N}$ (ungrouped)
2. $M = \frac{\Sigma f(X)}{N}$ (grouped data)

Table No. 1: Calculation of Mean from Data grouped into a frequency Distribution not italies

Class Interval	X	*f*	*f*x
5 - 9	7	9	63
10-14	12	5	60
15-19	17	7	119

20-24	22	8	176
25-29	27	8	216
30-34	32	6	192
35-39	37	2	74
40-44	42	3	126
45-49	47	2	94
		N = 50	Σfx = 1120

$\therefore$ Mean = $\frac{\sum fX}{N} = \frac{1120}{50}$ =22.40

Mean for the grouped data can be easily computed by Assumed mean or Short cut method too by using the formula.

$$M = AM + \frac{\sum fX}{N} \times i$$

Where, AM = Assumed Mean, *i* = Class interval, *f* = Respective frequency of the mid values of the class intervals, N = Total frequency

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

Let us take one example

Table 2

Class Interval	*f*	X (Mid point)	X'	fX'\
20-24	1	22	-4	-4
25-29	2	27	-3	-6
30-34	4	32	-2	-8
35-39	8	37	-1	-8
40-44	11	42	0	0
45-49	9	47	1	9
50-54	7	52	2	14
55-59	4	57	3	12
60-64	3	62	4	12
65-69	1	67	5	5
	N = 50			$\Sigma fX'=26$

Assumed Mean = 42

$$\therefore \text{Mean} = A + \frac{\sum fX'}{N} \times i$$

$$= 42 + \frac{26}{50} \times 5$$

$$= 42 + 2.6 = 44.6$$

Note :

1. Select any class interval near the middle of the distribution and call this interval zero in the deviation column. The midpoint of this class interval is the Assumed Mean or A.M., here it is 42.
2. Determine the deviation of each C.I. from the zero interval ($X' = \frac{X - A}{i}$ Where x is the raw score; A is the Assumed Mean and *i* = class interval.
3. Multiply each X' with each *f* to get fX' and add to get $\Sigma fX'$.

11.3.2 The Median

Another measure of Central Tendency is the Median. The median is the midpoint of a set of scores. If the number of scores is even, the median is half way between the two middle most scores. When the number of scores is odd, the middle is the middle score. Let us take one example. Suppose following are the given scores of 13 students. We require the Median. The scores are :

Ungrouped data

20
25
27
28
28
28
29
32
33

Median is the exact mid point of a given distribution

Median (Mid point, i.e. 7th number out of 13 students)

36
40
42
44

Calculation of Median from grouped data : Median can be calculated by using the following formula –

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

Here, L = exact lower limit of the median class.

F = Total of all frequencies before in the median class.

f = Frequency of the median class.

i = Class interval

N = Total of all the frequencies.

Let us calculate median by taking one example

Table 3 : Calculation of Median

Class Interval	*f*	*F* (*Cumulative frequency*)
20-24	1	1
25-29	2	3
30-34	4	7
35-39	8	15
40-44	11	26
45-49	9	35
50-54	7	42
55-59	4	46
60-64	3	49
65-69	1	50
	N = 50	

$$\therefore M_d = 39.5 + \frac{(50/2 - 15)}{11} \times 5$$

$$M_d = 39.5 + 4.55 = 44.05$$



Mode, the most frequently occurring score.

11.3.3 The Mode

The Mode is defined as the size of a variable (score) which occurs most frequently. In any series, it is the value of the item which is most common and is usually repeated the maximum number of time. Suppose we have to find out the value of the mode from the following scores of students -

25, 29, 24, 25, 27, 25, 28, 25, 29

Here, the score 25 occurs repeatedly. Hence, the value of mode in this case is 25.

Mode in case of Grouped data : In case of data which is available in the form of frequency distribution table, the mode can be computed by the following formula –

$$\text{Mode } (M_o) = 3 \text{ Mdn} - 2 \text{ M}$$

11.4 MERITS AND LIMITATIONS OF THE MEAN, THE MEDIAN AND THE MODE

THE MEAN : The mean is more commonly used and is also known as '**Common Average'**. Some of the **Merits** of Mean are –

- It can be easily calculated and is simple to understand
- It satisfies most of the conditions laid down for ideal average
- It can be used to compute further statistics like standard deviation, coefficient of correlation etc.

But the mean has some **limitations** too.

- Sometimes mean can give us absurd result. For example, the mean of a family size in a locality may be 4.62 which seems to be absurd.
- Sometimes the mean of a distribution is highly misleading especially when some of the observations are too large or too small as compared to others.

  E.g. a) The mean of 2, 84, 88, 80 and 85

  b) The mean of 15, 12, 16, 2925 etc.

THE MEDIAN : The median **has some merits** too. It is an

- Ideal average, for it is rigidly defined.
- It is not affected by the value of extreme items. Median can also be used with scores expressing grades or ranks such as A, B, C, D.

Median has some **limitation** too.

- It does not give equal weightage to each score in the series. Thus it may not be the true representative score.
- A median is more likely to be affected by the fluctuations of sampling than the mean.

MODE : The most important **merits** of mode are

- It is very simple. It is an average which people use in their day to day expression like sale of commodity, size of families etc.
- It is not affected by the extreme items.

The **limitations** of mode are

- It is not defined rigidly like mean
- It cannot be determined from unequal class intervals.



LET US KNOW

The mean takes into account the value of each score. Even one extremely high or low score could have an appropriate effect on the mean. Mean is the most accurate and reliable measure of Central Tendency.

The median is only a point which divides a set of scores into two equal halves. It is not affected by the extreme scores in the series.

The mode is the most frequent score in the set. It is the least reliable measure of Central Tendency.



ACTIVITY-1

(i) Find out the mean, median and mode for the following sets of score -

a) 9, 14, 8, 13, 10, 10, 11, 12, 10.

.....

.....

b) 24, 18, 19, 12, 24, 20, 21, 22.

.....

.....

(ii) Compute the mean, median and mode of the following distrabution.

Class Interval	Frequency
45-49	3
50-54	4
55-59	6
60-64	6
65-69	14
70-74	9
75-79	8
80-84	2
85-89	4
90-94	1
	N = 57



CHECK YOUR PROGRESS

Q.1: Point out the most common measures of Central Tendency.

.....

Q.2: In what particular situation will we use the mean, the median and the mode?

.....

.....

.....

Q.3: What are the differences between Mean and Median?

.....

.....

.....

Q.4: What do you mean by Central Tendency?

.....

.....

Q.5: Which measure of Central Tendency will you prefer in the following situation?

a) The average achievement of a group.

.....

b) The most popular fashion of the day.

.....

c) Determining the mid-point of the scores of a group in an entrance examination.

.....

Q.6: a) A random sample of 10 boys had the following intelligent quotients (I.Q.) 120, 70, 110, 101, 88, 83, 95, 98, 105, 100. calculate the mean.

.....

.....

b) What technique will you use to find out the median mental age of 8 students. 7, 10, 6, 8, 13, 9, 10, 11? Find out the median.

.....

.....



11.5 LET US SUM UP

- Central Tendency is a sort of average or a typical value of the items in the series or some characteristics of the members of a group.
- The most common measures of central tendencies are Arithmetic Mean, Median and Mode.
- Mean is the sum of all values of the items in a series divided by the total number of items.
- Median is the midpoint of a set of scores.
- Mode is defined as the size of a variable (score) which occurs most frequently. It is the least reliable measure of central tendency.
- Measures of Central tendency serve the purpose of summing up the whole distribution into one representative figure and this figure contains the gist of the complex group.
- There are three measures of Central Tendency. Mean, Median and Mode. Mean is used when scores are nearly symmetrically distributed around a central point. Median is used when the exact mid point – the 50% point of the distribution is wanted. Mode is used, when a quick and appropriate measure of Central Tendency is required.
- In the statistics of Mean, Median and Mode, all have some merits as well as demerits in their calculations and measures.



11.6 FURTHER READING

1) Garrett : H. E. (1956), Elementary Statistics, Longmans, Green and Co. New York.
2) Guildford, J. P. (1965): Fundamental Statistics in Psychology and Education, McGraw Hill Bokk Company, New York.
3) Asthana, B. (2000):Measurement and Evaluation in Psychology and Education, Vinod Pustak mandir, Agra-2.

11.7 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1 : The most common measure of the Central Tendency is the mean.

Ans to Q No 2 : ***Mean*** is used to get a reliable and accurate measure of Central Tendency

Median is used to get an exact midpoint of the distribution

Mode is used to get a quick and appropriate measure of Central Tendency.

Ans to Q No 3 : Mean represents the 'average' of the data given, whereas median is the score or value of the central item, which divides the series into two equal parts.

Ans to Q No 4 : A measure of central tendency is a sort of average or a typical value of the item in the series or some characteristics of the whole graph.

Ans to Q No 5 : a) Mean, b) Mode, c) Median

Ans to Q No 6 : a) Mean

$$= \frac{120+70+110+101+88+83+95+98+105+100}{10}$$

$\therefore$ MEAN = 97

b) 7, 10, 6, 8, 13, 9, 10, 11

The midpont may be either in 8 or 13.

$\therefore$ The exact midpoint will be 8+13/2= 10.5

11.8 MODEL QUESTIONS

Q.1: Calculate mean, median and mode from the following test scores:

8, 9, 12, 14, 16, 18, 20, 22, 28, 30, 31, 34, 36

Q.2: Make a frequency distribution from the following scores and then calculate Mean, Medfian and Mode.

17	15	36	18	16
22	16	38	24	18
18	19	39	27	19
20	18	42	29	21
29	27	40	35	21
26	31	32	27	22
28	40	34	19	23
30	45	40	15	28

Q.3: Calculate Mean, Median and Mode from the following frequency distributions:

C.I	Frequency
5-9	1
10-14	0
15-19	6
20-24	12
25-29	10
30-34	8
35-39	8
40-44	5
	N = 50

*** ***** ***

UNIT 12: MEASURES OF VARIABILITY

UNIT STRUCTURE

12.1 Learning objectives
12.2 Introduction
12.3 Measures of Variability
12.3.1 Computing different Measures of Variability (Q.D., A.D. and S.D.)
12.4 Computation of the Standard Deviation by direct method
12.5 Merits and limitations of Quartile Deviation, Average Deviation and Standard Deviation
12.6 Let Us Sum Up
12.7 Further Reading
12.8 Answers to Check Your Progress
12.9 Model Questions

12.1 LEARNING OBJECTIVES

After going through this unit you will be able to

- define the meaning of Variability and different measures of Variability.
- calculate the different measures of Variability of both grouped and ungrouped data.
- identify the merits and demerits of both grouped and ungrouped data.

12.2 INTRODUCTION

In the first three units, we have already discussed the meaning of statistics, its graphical representation as well as the very meaning and measures of Central Tendency. In this unit, we are going to explain the measures of variability.

12.3 MEASURES OF VARIABILITY

Variability helps in knowing about the nature of the distribution of individual scores in a given series or a set of data.

Measures of Variability indicate the performance of a group of students by a single score. Sometimes we may like to know about the variability, spread or dispersion of the scores as there is a tendency of the scores to be dispersed, scattered or to show variability around the average or the central value. This tendency is known as Dispersion or Variability.

There are different measures of variability. These measures of variability are :

- **Variability :**
- **1. Range :** The difference between the highest and lowest sores
- **2. Quartile Deviation :** It is the distance between the third and first quartiles divided by two. Quartiles are merely points which divide a set of scores into quartiles, and are indicated by Q_3, Q_2 and Q_1. The third and first quartiles (Q_3 and Q_1) respectively are points below which 75% and 25% of the score lie. Q_2 is the middle quartile or the median.
- **3. Average Deviation :** Average deviation is the mean of the deviation of all the separate scores in the series taken from the mean or some other measures of Central Tendency. It takes into account the variation of all the items in a series.
- **4. Standard Deviation :** Standard Deviation is also the mean of the degree to which a set of scores deviates from the mean or some other measures of Central Tendency. It is the mean and more stable measure of Variability than any other measure.

The formula for Range

1. Range = Highest – Lowest

The formula for Q.D.

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

The formulas for A.D. are

1. AD= $\frac{\Sigma|X|}{N}$ [ungrouped data]
2. AD= $\frac{\Sigma f|X|}{N}$ [grouped data]

The formulas for S.D are

1. $\delta = \sqrt{\frac{\Sigma fX'^2}{N}}$ [ungrouped data]
2. $\delta = \sqrt{\frac{\Sigma fX'^2}{N} - \left(\frac{\Sigma fX'}{N}\right)^2}$

12.3.1 Computing Different Measures of Variability (Q.D., A.D., S.D.)

A. Calculating of AD and SD for Ungrouped Data : We can compute the Average Deviation and Standard Deviation from the following data (ungrouped)

30, 35, 36, 39, 42, 44, 46, 38, 34, 35.

The mean of the given series

$$M = \frac{30+35+36+39+42+44+46+38+34+35}{10} = 37.9$$

Table 1 : Calculation of A.D.

Scores	$X^{/}$ (Deviation from the Mean)	$\lvert X \rvert$
30	30-37.9 = -7.9	7.9
35	35-37.9 = -2.9	2.9
36	36-37.9 = -1.9	1.9
39	39-37.9 = 1.1	1.1
42	42-37.9 = 4.1	4.1
44	44-37.9 = 6.1	6.1
46	46-37.9 = 8.1	8.1
38	38-37.9 = 0.1	0.1
34	34-37.9 = -3.9	3.9
35	35-37.9 = -2.9	2.9
N=10		39 = $\lvert X \rvert$

Here mean = 37.9

By applying the formula of mean AD = $\frac{39}{10}$=3.9

$\therefore$ Average deviation of the given score = 3.9

In the same way, we can also find out the standard deviation from ungrouped data by applying the formula of Standard Deviation.

i.e. $\delta = \sqrt{\frac{\sum X^2}{N}}$

From the above example we find the mean as

$$\frac{30+35+36+39+42+44+46+38+34+35}{10}$$

$\therefore$ M = 37.9

$\therefore$ S.D. will be

Table 2 : Calculation of S.D.

Scores	X'	X^2
30	-7.9	62.41
35	-2.9	8.41
36	-1.9	3.61
39	1.1	1.21
42	4.1	16.81
44	6.1	37.21
46	8.1	65.61
38	0.1	0.1
34	-3.9	15.21
35	-2.9	8.41
N=10		$\Sigma X^2 = 218.90$

$$\therefore \delta = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{218.90}{10}} = \sqrt{21.89} = \delta = 4.68$$

Thus we find that Average Deviation of the given score is 3.9 and Standard Deviation of the given score is 4.68.

Calculation of Q.D. and S.D. from Grouped data : We can calculate the quartile deviation and standard deviation for the frequency distribution given below.

Class-Interval	Frequency
0-4	1
5-9	2
10-14	2
15-19	15
20-24	25
25-29	30
30-34	17
35-39	2
40-44	3
45-49	2

Table 3 : Calculation of Quartile Deviation

Solution : How to find out the Quartile deviation.

Scores	*f*	Cu.f.
0-4	1	1
5-9	2	3
10-14	3	6
15-19	15	21
20-24	25	46
25-29	30	76
30-34	17	93
35-39	2	95
40-44	3	98
45-49	2	100
	N=100	

The formula for Quartile deviation is

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

$$\text{Where, } Q_1 = L + i\frac{\left(\frac{N}{4} - Cuf\right)}{f_q}$$

$$Q_3 = L + i\frac{\left(3\frac{N}{4} - Cuf\right)}{f_q}$$

Where,

L = the exact lower limit of the interval in which the quartile fall.

i = the length of the interval

Cuf = Cumulative f up to the interval which contains the quartile

F_q = the f on the interval containing the quartile.

Here, N = 100, so $Q_1 = \frac{N}{4} = \frac{100}{4} = 25$

And $Q_3 = 3\frac{N}{4} = \frac{3 \times 100}{4} = 75$

Now from the column of the cumulative frequency, we find that 25 will fall under the C.I. of 20-24 whose actual limits are 19.5 – 25.5.

Hence the Q_1 will be in the interval of 19.5-25.5.

The value of L=19.5, Cuf(F) = 21, f=25, i=5

$$\therefore Q_1 = 19.5 + 5\left(\frac{25 - 21}{25}\right) = 19.5 + 5(0.16)$$

$$= 19.5 + 0.8 = 20.30$$

Again Q_3 will be in the interval of 24.5 – 29.5 as 75 will fall under the C.I. of 25-29.

Hence the value of L = 24.5, Cuf (F) = 46, f=30, i=5

$$\therefore Q_3 = 24.5 + 5\left(\frac{75 - 46}{30}\right)$$

$$= 24.5 + 5(0.96)$$

$$= 24.5 + 4.8$$

$$\therefore Q_3 = 29.3$$

Therefore, the value of Q will be

$$Q = \frac{Q_3 - Q_1}{2} = \frac{29.3 - 20.30}{2} = 4.50$$

$\therefore$ Quartile deviation of the given data is 4.50

Computation of Standard Deviation : There are two methods in finding out Standard Deviation (S.D.)

1. Direct
2. Short Cut

12.4 COMPUTATION OF THE STANDARD DEVIATION BY DIRECT METHOD

A. The Standard Deviation (by Long Method) : Let us find out the standard deviation of the given frequency table.

Table 4 : Calculation of S.D. (Long Method)

C.I.	*f*	X (mid point)	fX	X'	X'^2
0-4	1	2	2	-23.05	531.30
5-9	2	7	14	-18.05	325.80
10-14	3	12	36	-13.05	170.30
15-19	15	17	255	-8.05	64.80
20-24	25	22	550	-3.05	9.30
25-29	30	27	810	1.95	3.80
30-34	17	32	544	6.95	48.30
35-39	2	37	74	11.95	142.80
40-45	3	42	126	16.95	287.30
45-49	2	47	94	21.95	481.80
N =	100		2505		2065.50

Steps :

1. To find out the mean
2. To find out the deviation of the mid point of C.I.
3. To square the deviation and to find the product of the squared deviation
4. To use the formula

$$\delta = \sqrt{\frac{\sum fX'^2}{N}}$$

Here Mean = $\sqrt{\frac{\sum fX'}{N}} = \frac{2505}{100}$ = 25.05

Standard Deviation = $\sqrt[3]{\frac{\sum fX'^2}{N}} = \sqrt{\frac{2065.50}{100}}$

Standard Deviation = $\sqrt{20.65}$

δ = 4.54

$\therefore$ The required Standard Deviation (δ) = 4.54

B. The Standard Deviation (by Short Cut Method) : Standard Deviation can also be computed by using short cut method. Let us take the same example.

Table 5 : Calculation of Standard Deviation (Short Cut Method)

C.I.	*f*	X (mid point)	X'	fX'	fX'^2
0-4	1	2	-5	-5	25
5-9	2	7	-4	-8	64
10-14	3	12	-3	-9	81
15-19	15	17	-2	-30	900
20-24	25	22	-1	-25	625

Steps :

1. To obtain deviation of the varieties from assumed mean D = X-M/i
2. To multiply the deviation by corresponding frequencies to get the column of fX/. The sum of this column will give $\Sigma fX'$.
3. Multiply to get the column fX'^2. The sum of this column is Σfd^2
4. To use the formula

$$\sigma = \iota\sqrt{\frac{\Sigma fx'^2}{N} - \left(\frac{\Sigma fx'^2}{N}\right)}$$

25-29	30	27	0	0	0
30-34	17	32	1	17	289
35-39	2	37	2	4	19
40-45	3	42	3	9	81
45-49	2	47	4	8	64
N =	100				2145

Here Assumed Mean = 27

$\Sigma fX' = \frac{-39}{100}$

$$\therefore \quad \delta = i\sqrt{\frac{\Sigma fX^2}{N} - \left(\frac{\Sigma fX}{N}\right)^2}$$

$$\delta = 5\sqrt{\frac{2145}{100} - \left(\frac{-39}{100}\right)^2}$$

$$\delta = 5\sqrt{21.45 - (0.15)}$$

$\delta = 5 \times 4.61$

$\therefore$ The required standard deviation = 23.05

12.5 MERITS AND LIMITATIONS OF QUARTILE DEVIATION, AVERAGE DEVIATION AND STANDARD DEVIATION

Quartile Deviation : The Quartile Deviation (Q) is one half the scale distances between the Third Quartile (Q_3) and First Quartile (Q_1). Some of the ***merits*** of Q.D. are

- It is useful in indicating the skewness of a distribution.
- Q.D. is preferred as measures of dispersion.

The **limitations** of Q.D. are :

- It roughly gives an idea of dispersion.
- It ignores the score above the third quartile and scores below the first quartile.

Average Deviation : Average Deviation is also known as the **first moment of dispersion**. Some **merits** of A.D. are –

- It is easily comprehensible even by a person not well versed in statistics.
- It is not very much affected by the value of extreme items.

Some ***limitations*** of AD are –

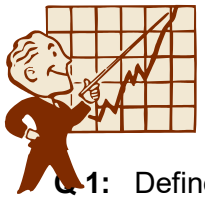
- It is rarely used in modern statistics.
- AS it ignores the algebraic signs of the deviation, it is as such not capable of further mathematical treatment.

Standard Deviation : S.D. is also known as **"Root mean square deviation from mean"**. Some of its **merits** are

- It is the mostly widely used measure of dispersion and most reliable measure of dispersion.
- Here the signs of deviation are not disregarded, they are eliminated by squaring each of the deviation.

Some of its ***limitations*** are –

- It is not easy to calculate and also to understand
- It gives more importance to extreme items and less to those which are near the mean.



CHECK YOUR PROGRESS

Q 1: Define Quartile Deviation and Standard Deviation.

.....

.....

Q 2: Which measures of variability will you prefer in the following situation

a) The measure of central tendency is available in the form of median.

.....

.....

b) The distribution is badly skewed.

.....

.....

c) The coefficient of correlation is subsequently to be computed.

.....

.....

d) The variability is to be calculated within no time.

.....

.....



12.6 LET US SUM UP

- Measures of variability or dispersion, supplement the information given by the measures of Central Tendency. It makes possible the comparison between different groups. It also serves as a useful check on drawing wrong conclusions from the comparison of average or measures of Central Tendency.
- There are four commonly used measures of variability or dispersion. Range is the index of variability. If the range is more, the group is more variable. Quartile deviation is used when the median is a measure of Central Tendency. Standard deviation is used, when the most accurate, reliable and stable measure of variability is wanted.



12.7 FURTHER READING

1) Blommers. P. and Lindquist. E. F. (1959), Elementary Statistical Methods. Houghton Mifflin, Company, Boston.

2) Gieldford, J. P. (1956). Fundamental Statistics in Psychology and Education, McGraw Hill Book Company, New York.

3) Rath, R. K. (1999). Fundamentals of Educational Statistics and Measurement. Taratarini Pustakalaya, Orissa.

4) Srivastava, A.B.L. and Sarma, K. K. (1985). Elementary Statistics in Psychology and Education. Sterling Publishers Pvt. Ltd., New Delhi.



12.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1 : Quartile deviation is the difference between 75 and 25

percentiles of a distribution; and Standard deviation is the square root of the sum of squred deviations of scores from their mean.

Ans to Q No 2 : a) Quartile deviation

b) Quartile deviation

c) Standard deviation

d) Range.

12.9 POSSIBLE QUESTIONS

Q 1: Compute the Average Deviation Quartile Deviation, Standard Deviation from the following data

Scores	*f*
80-84	4
85-89	4
90-94	3
95-99	5
100-104	3
105-109	3
110-114	1
115-119	2
N =	25

Q 2: Calculate the Standard Deviation and Quartile Deviation from the following data

C.I.	F
10-14	5
15-19	8
20-24	10
25-29	15
30-34	8
35-39	4

*** ***** ***

UNIT 13 : CO-EFFICIENT OF CORRELATION

UNIT STRUCTURE

13.1 Learning Objectives

13.2 Introduction

13.3 Coefficient of Correlation

13.3.1 Measures: Rank Difference Method

13.3.2 Calculation of ρ from Ranked data

13.4 Calculation of ρ (Rho) for tied Rank

13.4.1 Characteristics of ρ (Rho)

13.4.2 Merits and Limitations of Rho (ρ)

13.5 Let Us Sum Up

13.6 Further Reading

13.7 Answers to Check Your Progress

13.8 Model Questions

13.1 LEARNING OBJECTIVES

After going through this unit you will be able to know about

- meaning of Co-efficient of Correlation
- measures Rank Difference Method
- measures Product Moment Method.

13.2 INTRODUCTION

You have already learned about the measures of Central Tendency, Mean, Median and Mode and their methods of computation. You have also come to know about the different measures of variability. Let us know about co-efficient of correlation and their methods of computation. The details of coefficient of correlation will be discussed here.

13.3 COEFFICIENT OF CORRELATION

In education and psychology, there are times when one needs to know whether there exists any relationship between two variables. As for example we may want to know –

- Is there any relationship between scholastic achievement and general achievement?
- Is it true that dull children tend to be more neurotic than bright children?

In such type of problems, we need to find out the relationship between two variables (intelligence and achievement, neurotic and dull children etc.). These problems can be tackled by correlation technique. Thus it can be said that, **if the change in one variable appears to be accompanied by a change in the other variable, the two variables are said to be correlated and this interdependence is called correlation or co-variation.**



LET US KNOW

"Correlation is an analysis of the co-variation between two or more variables." – A.M. Tuttle.

Coefficient of Correlation : Coefficient of correlation is a numerical index that tells us to what extent the two variables are related and to what extent the variations in one variable changes with the variation in the other.

The co-efficient of correlation is always symbolized either by γ or (ρ (Rho).

There are many types of correlation like Linear, Biserial, Partial or Multiple etc. But in this unit we will discuss the Linear correlation and its computation.

Correlation can vary from perfect positive correlation and perfect negative correlation. The perfect positive correlation started from +1 and

1. Correlation means the relationship between two sets of scores or variables.
2. Coefficient of correlation is a kind of ratio which expresses the extent to which changes in one variable are accompanied by changes in another variable.

then it will pass through zero, indicating absence of correlation, and it will end with -1. Thus r varies from +1. Accordingly correlation may be positive, null or zero, and negative.

Perfect Negative (n) (-)	Zero (r) (0)	Perfect Positive (r) (+)
If the increase in one variable (x) results in a corresponding decrease in another variable (y), the correlation is said to be negative correlative. It ranges from (0 to -1) e.g. Increase of age and decrease of dependency.	Zero correlation means no relationship between the two variables X and Y. e.g. body weight and intelligence, show size and monthly income.	When the increase in one variable (X) is followed by a corresponding increase in other variable (Y), the correlation is positive. E.g. Increase of heat and increase of temperature, age and body weight.

13.3.1 Measures : Rank Difference Method

Rank difference method only consider the rank of the individual and it is designated by the Greek letter ρ (rho).

Rank difference method is also known as Spearman's Rank Correlation coefficient. As for example, there are some situations in Education or Psychology, where the objects or individuals may be ranked and arranged in order to merit on two variables. When these two sets of rank covar or have agreement between them, we measure the degrees of relationship by rank correlation. E.g. grouping of students on the basis of leadership ability, social adjustment etc.

Spearman has developed a formula called Rank Correlation Coefficient to measure the extent or degree of correlation between 2 sets of ranks. This coefficient of correlation is denoted by Greek letter ρ (called rho) and is given by

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

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

D = Difference between paired ranks (in each case)

N = Total number of items / individual rank

Calculation of ρ from test scores :

Example : The following data give the scores of 5 students in Mathematics and History respectively. Let us compute the Correlation between the two sets of scores by Rank difference method.

	A	B	C	D	E
Scores (Maths)	8	7	9	5	1
Scores (His)	10	8	7	4	5

Table 1 : Computation of ρ

Students in Maths.	Scores in His.	Scores Maths.	Rank 1 His.	Rank 11 in Rank	Difference	D^2
A	8	10	2	1	1	1
B	7	8	3	2	1	1
C	9	7	1	3	-2	4
D	5	4	4	5	-1	1
E N = 5	1	5	5	4	1 $\Sigma D = 0$	1 $\Sigma D2 = 8$

$$\therefore \rho = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6\times 8}{5(5^2-1)} = 1 - \frac{48}{120}$$

$\therefore \rho = 1 - 0.4 = 0.6.$

The coefficient is positive.

Steps :

1. List the students in column 1.
2. W rite the scores in column 2 and 3
3. Assign Rank according to scores of set 1.
4. Assign scores according to set II.
5. Calculate the difference of Rank of each student.
6. Check the sum of difference which is zero.
7. Squared each difference and add.
8. Apply formula.

13.3.2 Calculation of ρ From Ranked Data

In a debating competition Prof. Sarma and Prof. Das judged 10 students as follows. There judgement is given in Ranks. Determine the extent to which their judgements were in agreement.

	A	B	C	D	E	F	G	H	I	J
Prof. Sarma	1	3	4	7	6	9	8	10	2	5
Prof. Das	1	2	5	9	6	8	10	7	4	3

Table 2 : Computation of ρ (Rho)

Pupil	Prof. Sarma's Rank I	Prof. Das's Rank II	Difference	D^2
A	1	1	0	0
B	3	2	1	1
C	4	5	-1	1
D	7	9	-2	4
E	6	6	0	0
F	9	8	1	1
G	8	10	-2	4
H	10	7	3	9
I	2	4	-2	4
J	5	3	2	4
N = 10			$\Sigma D = 0$	$\Sigma D^2 = 28$

$$\therefore \rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

$$\therefore \rho = 1 - \frac{6 \times 28}{10(10^2-1)} = 1 - \frac{168}{10(100-1)}$$

$$\rho = 1 - \frac{168}{10(99)} = 1 - \frac{168}{990}$$

$\rho = 1 - 0.17$

$\rho = 0.83$

The correlation is positive.

13.4 CALCULATION OF ρ (RHO) FOR TIED RANK

Following are the data of 10 students on two trials of test with a gap of two weeks. let us compute the correlation between the scores of two trials by rank difference method.

Table 3 : Calculation of ρ (Rho)

Student	Trial I (X)	Trial II (Y)	Rank on Trial I	Rank on Trial II	Diff. D	D^2
A	10	16	6.5	5.5	1	1.00
B	15	16	3	5.5	-2.5	6.25
C	11	24	5	1.5	3.5	12.25
D	14	18	4	4	0	0.00
E	16	22	2	3	-1	1.00
F	20	24	1	1.5	-0.5	0.25
G	10	14	6.5	7.5	-1	1.00
H	8	10	9	10	-1	1.00
I	7	12	10	9	1	1.00
J N = 10	9	14	8	7.5	0.5 $\Sigma D = 0$	0.25 $\Sigma D^2 = 24.00$

Note :
When the scores are repeated, the procedure for assigning Rank is different.

1. In Trial I, the student A and G has same score, hence there rank will be (6^{th} and 7^{th}) instead of assigning 6^{th} and 7^{th}, we have to assign $\frac{6+7}{2}$ =6.5 to them.
2. Same procedure has to be followed in case of others too.

$$\therefore \rho = 1 - \frac{6 \times \sum D^2}{N(N^2 - 1)}$$

$$\therefore \rho = 1 - \frac{6 \times 24}{10(10^2 - 1)} = 1 - \frac{144}{990}$$

$$\therefore \rho = 1 - 0.14$$

$$\therefore \rho = 0.86$$

The co-efficient of correlation is positive and the value of Rho = 0.86.

13.4.1 Characteristics of Rho (ρ)

1. In Rank Correlation Coefficient, the observations or measurements of the bivariate variable is based on the ordinal scale in the form of ranks.
2. The size of the coefficient is directly affected by the size of the rank difference.
   a) If the ranks are the same for both tests, each rank difference will be zero and ultimately D^2 will be zero.
   b) If the rank differences are very large, and the fraction is greater than one, then the correlation will be negative.

13.4.2 Merits and Limitations of Rho (ρ)

Merits :

1. Spearman's Rank Order Coefficient of Correlation can be easily and quickly computed.
2. It is an acceptable method if data are available only in ordinal form or number of paired variables is more than 5 and not greater than 30 or a few ties in Rank.
3. It is quite easy to interpretρ.

   Some of the ***limitations*** of Rank difference methods are –

   1) If the number of cases is more, giving rank becomes a tedious job.
   2) When the interval data is converted into rank ordered data, the information about the size of the score difference is lost.



CHECK YOUR PROGESS

Q 1: What is Rank difference coefficient of correlation?

.....

.....

Q 2: What do you mean by –

a) Perfect positive correlation?

.....

.....

b) Perfect negative correlation?

.....

.....

c) Zero correlation?

.....

.....

Q.3. What is the formula of Rank difference coefficient of

correlation?

.....

.....

Q.4. Measures of correlation describe the relationship of

.....

.....



13.5 LET US SUM UP

- Spearman's rank order coefficient of correlation is used with ordinal data.
- The symbol ρ (Rho) is used to represent this coefficient of correlation.
- Perfect positive relationship (ρ=+1.00) means when the pairs of variables are related. As one increases the other tends to increase. As for example Intelligence and Academic achievement. Height and shoe size.
- Perfect negative relationship (ρ = -1.00) means when the pairs are negatively correlated. As one increases the other tends to decrease. As for example Academic achievement, and time spent per week of T.V. viewing; Time spent in practice, number of typing errors.
- Null correlation or zero correlation (ρ=0.00) when there is no correlation between two variables. E.g. Intelligence and body weight, Monthly salary and show size etc.

  Spearman's co-efficient of correlation can be determined by using the formula $\rho = 1 - \frac{6 \times \sum D^2}{N(N^2 - 1)}$



13.6 FURTHER READING

1) Ferguson, G. A. (1981). Statistical Analysis in Psychology and Education. (5th ed.). New York. Mc Graw Hill.

2) Guildford, J. P. and Fruchter B. (1978). Fundamental Statistics in Psychology and Education (6th ed.), New York. Mc Graw Hill.

3) Mangal S. K. (2003). Statistics in Psychology and Education (2nd ed.). New Delhi, Prentice Hall of India.

13.7 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: Rank difference method is also known as Spearman's Correlation Coefficient; where the objects or individuals are ranked and arranged in order of merit on two variables

Ans to Q No 2: a) Perfect positive correlation means when the increase in one variable (X) is followed by a corresponding increase in other variable(Y) and the value of r = +0.1, it is known as perfect positive correlation.

b) When the increase in one variable(X) results in a corresponding decrease in another variable(Y), the correlation is said to be negative correlation. (The value of r = -0.1)

c) Zero correlation means no relationship between the two variables. (r= 0)

Ans to Q No 3: P= 1- $\frac{6 \times \sum D^2}{N(N^2 - 1)}$

Ans to Q No 4 : Paired variables.

13.8 POSSIBLE QUESTIONS

Q 1: Discuss the need and importance of computing coefficient of correlation.

Q 2: Discuss the important methods of computing coefficient of correlation.

Q 3: Find rank correlation coefficient from the data given below and interpret the result.

Student:

Marks in Hindi – 30, 40, 50, 20, 10, 45, 22, 18

Marks in English – 55, 75, 60, 12, 35, 38, 25, 15.

*** ***** ***

Unit 14: REGRESSION

UNIT STRUCTURE

14.1 Learning Objective

14.2 Introduction

14.3 Uses of Regression Analysis

14.4 Objectives of Regression Analysis

14.5 Difference Between Correlation and Regression

14.6 Simmilarity Between Correlation and Regression

14.7 Regression Lines

14.8 Regression Equations

14.9 Lets Us Sum Up

14.10 Further Reading

14.11 Answer to Check Your Progress

14.12 Possible Questions

14.1 LEARNING OBJECTIVE

After going through this unit, you will be able to know

- what is regression
- difference between correlation and regression analysis
- uses of regression
- what are regression lines
- what is regression equation

14.2 INTRODUCTION

The term regression was first used by Sir Francis Galton in 1877. The dictionary meaning of the word regression is the act of returning or going back. Sir Galton introduced the term regression in the paper “Regression towards Mediocrity in Hereditary Stature”, where he studied the relationship between height of father and son. Sir Galton coined the term “regression line” to describe the tendency to regress or going back. Although this term has become popular, but modern writers nowadays prefer

to use the term "estimating line" as regression used for estimating or predicting the value of one variable given the value of another. For example, we know that advertising can have an impact on sales. So producers can use the cost on advertising to predict the amount of sales, or otherwise use the estimated amount of sale to know how much advertising cost needs to be incurred. In other words, regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

Some definitions of Regression

- "Regression analysis attempts to establish the *'nature of the relationship'* between variables – that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting." – Ya-Lun Chau
- "One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related casually, is regression analysis" – Taro Yamane
- The term regression analysis refers to the methods by which estimates are made of the values of a variable from the knowledge of the values of one or more other variables and to the measurement of errors involved in this estimation process". – Morris Hamburg

The above definition reveals that regression analysis is a statistical device with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable. The variable which is used to predict the variable of interest is called the independent variable or explanatory variable and the variable we are trying to predict is called the dependent variable or "explained" variable. The independent variable is denoted by X and the dependent variable by Y. The analysis used is called the simple linear regression analysis. It is called simple because there is only one independent variable, and linear because of the assumed linear relationship between the dependent and independent variables. Here the term "linear relationship" indicates that the relationship between dependent and independent variables forms a straight line of the equation Y=a+bx, where a and b are constants.

It may be pointed out that there is only mathematical or functional dependence between the two variables. Mere existence of dependence does not necessarily indicate any cause and effect relationship between the two variables.

14.3 USES OF REGRESSION ANALYSIS:

Regression analysis is widely used in various fields. Its applications are extended to almost all natural, physical and social sciences. Regression analysis is widely used for prediction and forecasting. It is particularly used to understand the relationship between dependent variables and independent variable, and to explore the forms of these relationships. Regression analysis is also widely used in business and in framing economic policies, especially because it has the ability to predict and forecast. It thereby helps in crucial decision making, among others.

14.4 OBJECTIVES OF REGRESSION ANALYSIS

The objectives of regression analysis are as follows:

- Regression analysis provides estimates of values of the dependent variable from values of the independent variable. The device used to accomplish this estimation procedure is the regression line. The regression line describes the average relationship existing between X and Y, i.e., it displays mean values of X for given values of Y. The equation of this line, known as the regression equation, provides estimates of the dependent variable when values of the independent variable are inserted into the equation.
- Another objective of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimator on. For this purpose the standard error of estimate is calculated.
- With the help of regression coefficients we can calculate the correlation coefficient.

14.5 DIFFERENCE BETWEEN CORRELATION AND REGRESSION

Correlation and regression are the most widely used economic tools. Both furnish important information, although in most of the situations, it is not clear as to which one should be adopted. The following are some of the points of difference between correlation and regression.

	CORRELATION	REGRESSION
1.	It measures the degree of covariability between X and Y.	It studies the nature of relationship between variables so that we can predict the value of one on the basis of another.
2.	Correlation analysis only provides the information that there is relationship between two variables, asserting the degree of relationship, but tells nothing about dependence.	In regression, on the other hand, one variable is dependent and the other is independent. Thus, it tells us about the mathematical or functional dependence of the two variables.
3.	In correlation analysis r_{xy} is a measure of direction and degree of linear relationship between two variables X and Y. r_{xy} and r_{yx} are symmetric ($r_{xy} = r_{yx}$), i.e., it is immaterial which of X and Y is dependent variable and which is independent variable.	In regression analysis, the regression coefficients *bxy* and *byx* are not symmetric, i.e., *bxy?byx* and hence it definitely makes a difference as to which variable is dependent and which is independent.
4.	There may be correlation between two variables without any logical sense, such as increase in income and weight.	In regression analysis there cannot be any illogical dependence.
5.	Correlation coefficient is independent of change of origin and scale.	Regression coefficient is independent of change of origin but not of scale.

14.6 SIMMILARITY BETWEEN CORRELATION AND REGRESSION

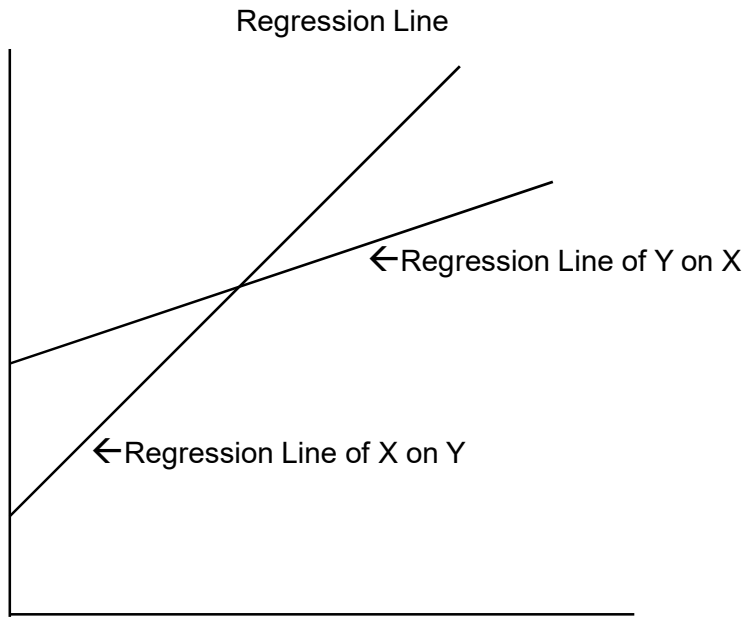
There is also some similarity between correlation and regression analysis.

- The coefficient of correlation (r) takes the same sign as the regression coefficients (b_{xy} and b_{yx}).
- If the value of *b* is significant at a given level of significance *r* is also significant at that level.

14.7 REGRESSION LINES

Let us consider two variables X and Y. We will have two regression lines as the regression of X on Y and the regression of Y on X. The regression line of Y on X gives the most probable values of Y for given values of X and the regression line of X on Y gives the most probable values of X for given values of Y. However, we will have only one regression line, i.e, regression lines will coincide when there is either perfect positive or perfect negative correlation between two variables (r=±1). The further the two regression lines from each other, the lesser will be the degree of correlation between the variables and nearer the two regression lines with each other, the higher will be the degree of correlation between the variables. If the variables are independent, r is zero and the lines of regression are at right angles, i.e., parallel to OX and OY.

It may be pointed out that regression lines cut each other at the point of average of X and Y, i.e., if from the point where both the regression lines cut each other a perpendicular is drawn on the X-axis, we will get the mean value of X and if from that point a horizontal line is drawn on the Y-axis, we will get the mean value of Y.



14.8 REGRESSION EQUATIONS

Regression equations, also known as estimating equations, are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations – the regression equation of X on Y is used to describe the variation in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

Regression equation of Y on X

The regression equation of Y on X is expressed as follows:

$Y = a+bX$

Here, *Y*= dependent variable, i.e., its value depends on X

X=independent variable, i.e., we can take a given value of X and compute the value of Y

a=Y-intercept, because its value is the point at which the regression line crosses the Yaxis or the vertical axis.

b=slope of the line. It represents change in Y variable for a given change in X variable.

'*a*' and '*b*' are called numerical constants because for any given straight line, their values does not change.

Regression equation of X on Y

The regression equation of X on Y is expressed as follows:

$$X = a+bY$$

Here, *X*= dependent variable, i.e., its value depends on X

Y=independent variable, i.e., we can take a given value of X and compute the value of Y

a=X-intercept, because its value is the point at which the regression line crosses the Y-axis or the vertical axis.

b=slope of the line. It represents change in X variable for a given change in Y variable.



CHECK YOUR PROGRESS

Q1: What is the regression equation of Y or X?

.....

.....

Q2: What is the regresion equation of X and Y?

.....

.....



14.9 LETS SUM UP

- Regression analysis reveals the average relationship between two variables and this makes possible estimation or prediction.
- Regression analysis is widely used in various fields. Its applications are extended to almost all natural, physical and social sciences.
- Regression analysis provides estimates of values of the dependent variable from values of the independent variable.
- With the help of regression coefficients we can calculate the correlation coefficient.
- There are similarities as well as differences between correlation and regression.
- The regression equation of Y on X is expressed as follows:
  $Y = a+bX$
- The regression equation of X on Y is expressed as follows:
- $X = a+bY$



14.10 FURTHER READING

1) Gupta S.P. 2010. Statistical Method. New Delhi: Sultan and Sons.

14.11 ANSWER TO CHECK YOUR PROGRESS

Ans to QNo 1: The regression equation of Y on X is expressed as follows:

Here, Y = a + bX

Here, Y = dependent variable, I.e., its value depends on X

X = independent, i.e., we can take a given value of X and compute the value of Y

a = Y - intercept, because its value is the point at which the regression line crosses the Y axis or the vertical axis.

b = slope of the line. It represents change in Y variable for a given change in X variable.

'a' and 'b' are called numerical constants because for any given straight line, their values does not change.

Ans to QNo 2: The regression equation of X on Y is expressed as follows:

X = a + bY

Here, X = dependent variable, i.e., its value depends on X

Y = independent variable, i.e., we can take a given value of X and compute the value of Y

a = X- intercept, because its value is the point at which the regression line crosses the Y-axisor the vertical axis.

b = slope of the line. It represents change in X variable for a given change in Y variable.

14.12 POSSIBLE QUESTIONS

A. Short answer type: (Answer within 150 words)

Q 1: Define regression.

Q 2: Write any two uses of regression analysis.

Q 3: Write any two objectives of regression analysis.

Q 4: Write two similarities of correlation and regression.

Q 6: Write two differences of correlation and regression.

Q 7: What are regression lines?

Q 8: Write the regression equation of X on Y.

Q 9: Write the regression equation of Y on X.

B. Long answer type (Answer within 150 words)

Q 1: What is regression? What are its different uses?

Q 2: What are the objectives of regression analysis?

Q 3: Discuss the differences between correlation and regression.

Q 4: Discuss in detail the regression eqution.

*** ***** ***

UNIT 15: CHI-SQUARE TEST (I)

UNIT STRUCTURE

15.1 Learning Objectives

15.2 Introduction

15.3 Chi square test

15.4 Chi square test for the Goodness of Fit

15.5 Chi square test for the independence of variables

15.6 Chi square test for the equality of more than two population proportions

15.7 Let Us Sum Up

15.8 Further Reading

15.9 Answers to Check Your Progress

15.10 Possible Questions

15.1 LEARNING OBJECTIVES

After going through this unit, you will be able to

- discuss the non-parametric test
- discuss the Independence of attributes
- describe the degrees of freedom
- dicuss chi- Square Distribution
- discuss the Goodness of Fit

12.2 INTRODUCTION

In this Unit we are going to discuss the various concepts of Chi-Square Analaysis.Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. This unit will give us a fair idea about the Chi square test for the Goodness of Fit, Chi square test for the independence of variables and Chi square test for the equality of more than two population proportions.

Let us dicuss this concepts in the following sections.

15.3 CHI SQUARE (χ^2) TEST

Chi square test is applies to those problems in which we study whether the frequency with which a given event has occurred, is significantly different from the one as expected theoretically. The measure of Chi square enables us to find out the degree of discrepancy between observed frequencies and theoretical frequencies and thus to determine whether the discrepancy so obtained between observed frequencies and theoretical frequencies is due to error of sampling or due to chance

The Chi-square is computed on the basis of frequencies in a sample and thus the value of Chi-square so obtained is a statistic. Chi-square is not a parameter as its value is not derived from the observations in a population. Hence Chi-square test is a non-parametric test.The Chi Square (χ^2) Test is defined as

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \text{ where}$$

$O_i = \text{observed frequency of the } i^{th} \text{ event}$

$E_i = \text{expected frequency of the } i^{th} \text{ event}$

Degrees of Freedom (df)

The number of data that are given in the form of a series of variables in a row or column or the number of frequencies that are put in cells in a contingency table, which can be calculated independently is called the degrees of freedom and is denoted by ν.

Case I If the data is given in the form of a series of variables in a row or column, then the degrees of freedom = (number of items in the series) - 1, i.e., $\nu = n - 1$.

Case II When the number of frequencies are put in cells in a contingency table, the degrees of freedom will be the product of (number of rows less one) and the (number of columns less one) i.e., ν = (R - 1)(C-1), where R is the number of rows and C is the number of columns.

Uses of Chi Square (χ^2) Test

The χ^2 test is a very powerful test for testing the hypothesis of a number of statistical problems. The important uses of χ^2 test are:

15.4 CHI SQUARE TEST FOR GOODNESS OF FIT

χ^2 test is a measure of probabilities of association between the attributes. It gives us an idea about the divergence between the observed and expected frequencies. Thus the test is also described as the test of "Goodness of Fit". If the curves of these two distributions, when superimposed do not coincide or appear to diverge much we say the fit is poor. On the other hand if they do not diverge much, then the fit is less poor.

Example: A survey with 320 families with 5 children each revealed the

No. of Boys:	5	4	5	2	1	0
No. of Girls:	0	1	2	3	4	5
No. of Families:	14	56	110	88	40	12

Is the result consistent with the hypothesis that the male and female births are equally probable?

Solution:

Null Hypothesis H_0 : Male and female births are equaly probable

Alternative Hypothesis H_1 : Male and female births are not equaly probable

Calculation of χ^2: Based on the null hypothesis the expected frequencies are given by

$$320\left[\frac{1}{2}+\frac{1}{2}\right]^5 = 320\left[\frac{1}{32},\frac{5}{32},\frac{10}{32},\frac{10}{32},\frac{5}{32},\frac{1}{32}\right] = 10,50,100,100,50,10$$

Table for χ^2

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
14	10	4	16	1.60
56	50	6	36	1.72
110	100	10	100	1.00
88	100	-12	144	1.44
40	50	10	100	2.00
12	10	2	4	0.40
				$\chi^2 = 7.16$

Level of Significance: α = 0.05

Critical Value: The table value of χ^2 at 5% level of significance with 6 - 1 = 5 df is 11.07.

Decision: Since the calculated value of χ^2 is less than table value of χ^2 so null hypothesis is accepted and can be concluded that the male and female births are equally probable.

15.5 CHI SQUARE TEST FOR THE INDEPENDENCE OF VARIABLES

χ^2 test can be applied to test, the independence between various attributes when the sample data is presented in the form of a contingency table with any number of rows 'r' and columns 'c' the null hypothesis and alternative hypothesis are set as follows:

Null Hypothesis H_0 : The attributes are independent

Alternative Hypothesis H_1 : The attributes are not independent

Example: Samples of 400 students of undergraduate and 400 postgraduate classes were taken to know their opinion about autonomous college. 290 undergraduate and 310 postgraduate students favoured the autonomous status. Test whether the opinion about the autonomous status of students independent of their education level or not.

Solution: The contingency table of observed frequencies is

Level of Education	Favour	Not in Favour	Total
Undergraduate	290	110	400
Post graduate	310	90	400
Total	600	200	800

Table of expected frequencies will be

Level of Education	Favour	Not in Favour
Undergraduate	(600 x 400)/800 = 300	(600 x 400)/800 = 300
Post graduate	(200 x 400)/800 = 100	(200 x 400)/800 = 100

Null Hypothesis H_0: Opinion regarding an autonomous status are independent of level of education.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(290-300)^2}{300} + \frac{(310-100)^2}{100} + \frac{(110-300)^2}{300} + \frac{(90-100)^2}{300} = 556$$

df = (2-1) (2-1) = 1

Critical Value: Table value of χ^2 at 5% level of significance with 1 df is 3.84.

Decision: Since calculated χ^2 is greater than table value of χ^2, so H_0 is rejected and can be concluded that opinion about autonomous is not independent of level of education.



CHECK YOUR PROGRESS

Q.1: Write short notes about Chi-square test. Why it is called a non parametric test?

.....

.....

Q.2. Define degrees of freedom.

.....

.....

Q.3. Give the test statistic of chi-square test.

.....

.....

Q.4. What are the uses of chi-square test?

.....

.....

15.6 CHI SQUARE TEST FOR THE EQUALITY OF MORE THAN TWO POPULATION PROPORTIONS

Here, the null hypothesis is that the proportion in each the response variable categories is the same for all treatment groups. The alternate hypothesis is that the set of proportions differs somewhere among the treatment groups. It is possible to write the hypotheses in terms of population parameters. Let π_{ij} be the proportion of treatment population i that is found in response category j. There are G treatment groups and K response categories. The null hypothesis is:

H : $\pi_{11} = \pi_{21} = \pi_{31} = \ldots = \pi_{G1}$

The basic summary statistic in a chi-square analysis is the contingency table which summarizes the number of observations for each combination of the explanatory and response variable. Sample percentages are often computed (denoted as p_{ij} to distinguish them from the π_{ij} in the population).

Example: Suppose assembly elections are announced in three major states. A major political party wishes to test if the proportions of its supporters in the three states are the same or not. The party conducts a sample survey of 1000 people in each state and finds that there are 300, 350 and 425 supporters in the sample in the surveyed state. Using these data test if the three proportions are the same or not.

Solution: Here, we wish to test

$H_0: \Pi_1 = \Pi_2 = \Pi_3$ (All the proportions are the same)

$H_1: \text{Not all } \Pi_1 = \Pi_2 = \Pi_3$ are equal

Note that, we only test if all the proportions are the same or not. In case we conclude that all the proportions are not the same, we cannot find which one is different.

We first make the 2X3 contingency table of the given data as below:

Supporter	State 1	State 2	State 3	Total
Yes	300	350	425	1075
No	700	650	575	1925
Total	1000	1000	1000	3000

The expected cell frequencies will be as follows:

Supporter	State 1	State 2	State 3
Yes	358.33	358.33	358.33
No	641.67	641.67	641.67

Test Statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(300-358.33)^2}{358.33} + \frac{(350-358.33)^2}{358.33} + \frac{(425-358.33)^2}{358.33}$$

$$+ \frac{(700-641.67)^2}{641.67} + \frac{(650-641.67)^2}{641.67} + \frac{(575-641.67)^2}{641.67}$$

$$= 9.495 + 0.194 + 12.404 + 5.302 + 0.108 + 6.927$$

$$= 34.43$$

The test statistic follows chi-square distribution with (2-1)(3-1)=2 d.f. Therefore critical value at 5% level of significance is 5.991. Since computed test statistic is larger than the critical value, we reject the null hypothesis at 5% level of significance and conclude that all proportions of supporters of the political party in three states are not the same.



15.7 LET US SUM UP

Chi square test is computed on the basis of frequencies in a sample and thus the value of chi square so obtained is a statistic.

Chi-square is not a parameter as its value is not derived from the observations in a population. Hence chi-square test is a non parametric test.

It is not concerned with any population distribution and its observation.

The number of data that are given in the form of a series of variables in a row or column or the number of frequencies that are put in cells in a contingency table, which can be calculated independently is called the degrees of freedom and is denoted by ν.

Chi square test can be applied for

- Test of Goodness of Fit
- Test for the independence of variables
- Test for the equality of more than two population proportions

FURTHER READING

1) Kothari, C.R. and Garg, G.(2014), 'Research Methodology - Methods and Techniques', New Age International Publishers

2) Arora, P.N. and Arora, S.(2009), ' Statistics for Management', S. Chand

15.9 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q.1: The Chi-square is computed on the basis of frequencies in a sample and thus the value of Chi-square so obtained is a statistic. Chi-square is not a parameter as its value is not derived from the observations in a population. Hence Chi-square test is a non-parametric test.

Ans to Q.2: The number of data that are given in the form of a series of variables in a row or column or the number of frequencies that are put in cells in a contingency table, which can be calculated independently is called the degrees of freedom and is denoted by ν.

Ans to Q.3: The χ^2 test is a very powerful test for testing the hypothesis of a number of statistical problems. The important uses of χ^2 test are:

- Test of Goodness of Fit
- Test for the independence of variables
- Test for the equality of more than two population proportions

15.10 POSSIBLE QUESTIONS

Q 1: A sample of 300 students of undergraduate and 300 students of Postgraduate classes of a University were asked to give their opinion toward the autonomous colleges. 190 of the undergraduate and 210 of the Postgraduate students favoured the autonomous status.

Test at 5% level of significance whether the opinion of students towards the status are independent of the classes or not.

Q 2: In 60 throws of a dice, face one turned up 6 times, face two or three 18 times, face four or five 24 times and face six 12 times. Test at 10% significance level whether the dice is unbiased?

Q 3: One thousand girls in a college were graded according to their I.Q. and the economic conditions of their homes. Use chi-square test to find out whether there is any association between economic conditions at home and I.Q. of girls.

Q 4: From the table given below, test whether the colour of son's eye is associated with that of father's eye.

EYE COLOUR OF SONS

Eye Colour of Father	Light	Dark
Light	471	151
Dark	148	230

*** ***** ***

UNIT15: F-DISTRIBUTION AND ANALYSIS OF VARIANCE (ANOVA)

UNIT STRUCTURE

15.1 Learning Objectives
15.2 Introduction
15.3 The F Distribution
15.4 Assumptions For F-Test
15.5 Properties Of F-Distribution
15.6 Applications Of F-Distribution
15.6.1 Hypothesis Testing for Difference between Two Population variances
15.7 Analysis Of Variance
15.8 Assumptions Of Analysis Of Variance
15.9 Classification Of Analysis Of Variance
15.9.1 One -Way Classification
15.9.2 Two-Way Classification
15.10 Let Us Sum Up
15.11 Further Readings
15.12 Answer To Check Your Progress
15.13 Model Questions

15.1 LEARNING OBJECTIVES

After going through this unit you will be able to

- introduce F distribution and learn how to use it in statistical inferences
- understand the assumptions involved in the use of analysis of variance
- to compare more than two population means using analysis of variance

15.2 INTRODUCTION

In several statistical applications a decision maker might want to compare the difference in two population variances. For example, (i) Quality of products, as measured by some indicators like diameter, length, etc.

manufactured by two different production processes. The process with lower variance, implying more consistent quality, is preferred. (ii) Price or Earning Per Share (EPS) of two shares on day-to-day basis, say for a week or month, etc. The share with lower variance implying lesser volatility could be considered less risky and as a result preferred by some investors. (iii) Service time taken by two systems or agencies—the one having lower variance is preferred and so on are some of the areas where comparison of variances is needed. In testing the hypotheses of the difference in two population variances, sample variances are used. Thus the ratio of the two sample variances $\frac{S_1^2}{S_2^2}$ is termed as F test. Therefore the test statistic used for the difference in two population variances

$$F = \frac{S_1^2}{S_2^2}$$

follows F distribution with $d.f. = n_1 - 1$ and $d.f. = n_2 - 1$, where S_1^2 and S_2^2 are defined by

$$S_1^2 = \frac{\sum\left(x_1 - \overline{x_1}\right)^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\sum\left(x_2 - \overline{x_2}\right)^2}{n_2 - 1}$$

15.3 THE F DISTRIBUTION

F distribution is based on the assumption that the populations from which samples are drawn are normally distributed. This distribution was first originated by G.W.Snedecor and the distribution was named by Snedecor as F distribution in honour of the famous English statistician Sir Ronald A Fisher. However, this distribution is also called Snsdecore's F distribution. As in the case with t and Chi-square distributions, the shape of a particular F distribution curve depends on the number of degrees of freedom. A major difference between and Chi-square distributions and F distribution is that the former two distributions have only one number of degrees of freedom whereas the latter has two numbers of degrees of freedom. This is because in the F distribution, the degrees of freedoms are attached to both numerator and denominator. It is to be noted that the F distribution is a continuous probability distribution and is positively skewed. The probability density function of F distribution is given by

$$f(F) = y_0 (F)^{(\nu_1/2)-1} \left(1 + \frac{\nu_1}{\nu_2} F\right)^{-(\nu_1+\nu_2)/2} ; \quad 0 \le F < \infty$$

where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ and y_0 is a constant which is so determined that the total area under the curve is unity.

15.4 ASSUMPTIONS FOR F -TEST

Few assumptions for the ratio $\frac{S_1^2}{S_2^2}$ to have an F distribution are given below:

(i) Independent random samples are drawn from each of two normal populations.

(ii) The variability of the measurements in the two populations is same and can be measured by a common variance σ^2, *i.e.*, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. The assumption is needed in order to combine the variances within the populations into a single source of variation 'within populations'.

15.5 PROPERTIES OF F DISTRIBUTION

Some important properties of F distribution are given below:

(i) The total area or probability under the curve is unity.

(ii) The F distribution is positively skewed with a range 0 to ∞ and its degree of skewness decreases with the increase in degrees of freedom $\nu_1 = n_1 - 1$ for numerator and $\nu_2 = n_2 - 1$ for denominator. For $\nu_2 \ge 30$, F distribution is approximately normal.

(iii) The mean of the F distribution is

$$\text{Mean} = \frac{\nu_2}{\nu_2 - 2}; \nu_2 > 2$$

Thus mean of F distribution is independent of the first degree of freedom ν_1. Obviously mean is greater than unity.

(iv) Variance of F distribution is

$$\text{Variance} = 2\left(\frac{1}{\nu_1} + \frac{1}{\nu_2}\right)$$

(v) Mode of the distribution is

Mode = $\frac{\nu_2(\nu_1 - 1)}{\nu_1(\nu_2 + 2)}$

Since F>0, mode exists only if $\nu_1 > 2$. Clearly mode is less than unity.

(vi) The curve of the F distribution rises steadily from 0 till the model value and then decreases slowly.

Critical values of F distribution: Critical values of F for different pairs of degrees of freedom at 1% level and 5% level of significance for the right-tailed test are available in F-table. The critical region corresponding to a specified level of significance and degrees of freedom is determined by the right-tail area. The significant value (i.e., critical value) $F_\alpha(\nu_1, \nu_2)$ at level of significance α and $(\nu_1, \nu_2)\, d.f.$ is determined by the equation

$$P[F > F_\alpha(\nu_1, \nu_2)] = \alpha$$

as drawn in the following figure.

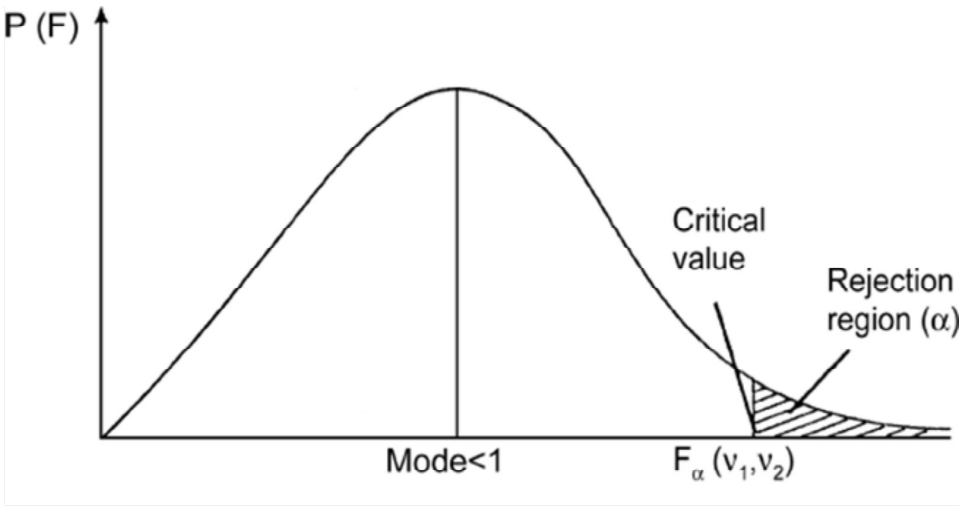


Figure 10.1

CHECK YOUR PROGRESS

Q 1. State whether the following statements are true or false:

(i) The F -distribution is positively skewed.

(ii) The -distribution is based on only one degrees of freedom.

(iii) The test statistic $F = S_1^2 / S_2^2$ is used to test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$.

15.6 APPLICATIONS OF F DISTRIBUTION

F distribution is widely used in business environment by decision maker. Some of the applications are outlined below:

(i) F test is applied to test the equality of population variances.

(ii) F test is used to test the significance of an observed sample multiple correlation.

(iii) It is used to test the significance of an observed sample correlation ratio.

(iv) It is applied for testing the linearity of regression.

(v) It is used to test the equality of several population means.

Hereunder we shall, of course, discuss only the application of testing the equality of population variances.

15.6.1 HYPOTHESIS TESTING FOR DIFFERENCE BETWEEN TWO POPULATION VARIANCES

Let us consider two independent random samples of sizes n_1 and n_2 drawn from two normal populations with variances σ_1^2 and σ_2^2. We want to test the hypothesis whether the variances of the two populations are equal or if there exists any significant difference between the population variances.

Thus the null hypothesis is $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$

The alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Under the null hypothesis, the test statistic to be constructed is

$$F = \frac{S_1^2}{S_2^2}$$

follows F distribution with $d.f. = \nu_1 = n_1 - 1$ and $d.f. = \nu_2 = n_2 - 1$, where S_1^2 and S_2^2 are unbiased estimates of the common population variance σ^2 defined by

$$S_1^2 = \frac{\sum\left(x_1 - \overline{x_1}\right)^2}{n_1 - 1} \text{ and } S_2^2 = \frac{\sum\left(x_2 - \overline{x_2}\right)^2}{n_2 - 1}$$

x_1 and x_2 denote respectively the sample observations of the two random samples drawn from the two normal populations.

Since F test is based on the ratio of two variances it is also termed as Variance Ratio Test.

The criteria of acceptance or rejection of null hypothesis H_0 are as under:

Accept H_0 if the calculated value of the test statistic F is less than its critical value $F_\alpha(\nu_1, \nu_2)$ and otherwise reject the null hypothesis.

An Important Note: 1. Since the available tables of the significant values of F at different levels of significance are for the right-tail test, therefore in numerical problems we will take greater of the variances S_1^2 and S_2^2 in the numerator and adjust for the degrees of freedom accordingly. Thus if $F \sim$ then refers to the degrees of freedom of the larger variance, which must be taken in the numerator while calculating

2. In numerical problems, usually sample varianceis given from which can be obtained on using the relation

$$ns^2 = (n-1)S^2$$

where $s^2 = \frac{\sum\left(x - \overline{x}\right)^2}{n}$

Relation between t, F and χ^2 Distributions:

We give below (without proof), the relationship between and Distributions:

Relation between and distributions:

If a statistic t follows Student's t distribution with then its square follows Snedecores's F distribution with (1, n) degrees of freedom. Symbolically,

$$t \sim t_n \quad \Rightarrow \quad t^2 \sim F(1, n)$$

with the help of this relation, all the applications of t distribution may be taken as applications of F distribution. For example, for testing the significance of sample mean, we may consider the test statistic

$$F = t^2 = \left[\frac{\bar{x} - \mu}{S/\sqrt{n}}\right]^2 \sim F(1, n)$$

and apply F-test.

Relation between F and χ^2 Distributions:

In F distribution with $(n_1, n_2)d.f.$ if we put

$$n_1 F = \chi^2 \Rightarrow F = \chi^2 / n_1$$

and let $n_2 \to \infty$, then we get Chi-square distribution with $n_1\, d.f.$

Example 10.1: The time taken by workers in performing a job by method I and method II is given below:

Method I:	19	22	24	27	24	18	20	19	25		
Method II:	26	37	40	35	30	30	40	26	30	35	45

Do the data show that the variance of time distributions for populations from which these samples are drawn do not differ significantly?

Solution: Let us set up null and alternative hypotheses as

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ i.e., there is no significant difference between the variances of the time distribution by the workers in performing a job by Method I and Method II.

Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Under the null hypothesis, the test statistic to be constructed is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) \quad provided\ S_1^2 > S_2^2$$

where S_1^2 and S_2^2 are unbiased estimates of the common population variance σ^2 defined by

$$S_1^2 = \frac{\sum (x_1 - \overline{x_1})^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\sum (x_2 - \overline{x_2})^2}{n_2 - 1}$$

We now calculate sample variances

Method I			Method II		
x_1	x_1-22	$(x_1-22)^2$	x_2	x_2-34	$(x_2-34)^2$
19	-3	9	26	-8	64
22	0	0	37	3	9
24	2	4	40	6	36
27	5	25	35	1	1
24	2	4	30	-4	16
18	-4	16	30	-4	16
20	-2	4	40	6	36
19	-3	9	26	-8	64
25	3	9	30	-4	16
			35	1	1
			45	11	121
198	0	80	374	0	380

$$\overline{x_1} = \frac{\sum x_1}{n_1} = \frac{198}{9} = 22; \qquad \overline{x_2} = \frac{\sum x_2}{n_2} = \frac{374}{11} = 34$$

$$S_1^2 = \frac{\sum (x_1 - \overline{x_1})^2}{n_1 - 1} = \frac{80}{8} = 10; \qquad S_2^2 = \frac{\sum (x_2 - \overline{x_2})^2}{n_2 - 1} = \frac{380}{10} = 38$$

Since $S_1^2 > S_2^2$, under H_0, the test statistic is

$$F = \frac{S_1^2}{S_2^2}$$
$$= \frac{38}{10} = 3.8$$

The critical value $F_{0.05}(10,8) = 3.35$

Since the calculated value of F is greater than the critical value therefore we reject the null hypothesis at 5% level and conclude that the variability of the time distribution in the two populations is not same.

Example 10.2: A plant has installed two machines producing polythene bags. During the installation, the manufacturer of the machine has stated that the capacity of the machine is to produce 20 bags in a day. Owing to various factors such as different operators working on these machines, raw material, etc. there is a variation in the number of bags produced at the end of the day. The company researcher has taken a random sample bags produced in 10 days for machine 1 and 13 days for machine 2, respectively.

The following data gives the number of units of an item produced on a sampled day by the two machines:

Machine I: 18 19 19 18 17 19 18 19 18 19

Machine II: 18 17 17 17 16 18 18 16 18 17 16 16 17

Solution: Let us set up null and alternative hypotheses as

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ i.e., there is no significant difference between the variances of the time distribution by the workers in performing a job by Method I and Method II.

Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Under the null hypothesis, the test statistic to be constructed is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) \quad provided\ S_1^2 > S_2^2$$

where S_1^2 and S_2^2 are unbiased estimates of the common population variance σ^2 defined by

$$S_1^2 = \frac{\sum (x_1 - \overline{x_1})^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\sum (x_2 - \overline{x_2})^2}{n_2 - 1}$$

We now calculate sample variances

x_1	$x_1 - 18.4$	$(x_1 - 18.4)^2$	x_2	$x_2 - 17$	$(x_2 - 17)^2$
18	-0.4	0.16	18	1	1
19	0.6	0.36	17	0	0
19	0.6	0.36	17	0	0
18	-0.4	0.16	17	0	0
17	-1.4	1.96	16	-1	1
19	0.6	0.36	18	1	1
18	-0.4	0.16	18	1	1
19	0.6	0.36	16	-1	1
18	-0.4	0.16	18	1	1
19	0.6	0.36	17	0	0
			16	-1	1
			16	-1	1
			17	0	0
184	0	4.4	221	0	8

$$\overline{x_1} = \frac{\sum x_1}{n_1} = \frac{184}{10} = 18.4; \quad \overline{x_2} = \frac{\sum x_2}{n_2} = \frac{221}{13} = 17$$

$$S_1^2 = \frac{\sum\left(x_1 - \overline{x_1}\right)^2}{n_1 - 1} = \frac{4.4}{9} = 0.488889; \qquad S_2^2 = \frac{\sum\left(x_2 - \overline{x_2}\right)^2}{n_2 - 1} = \frac{8}{12} = 0.66667$$

Since $S_2^2 > S_1^2$, under H_0, the test statistic is

$$F = \frac{S_2^2}{S_1^2}$$
$$= \frac{0.66667}{0.488889} = 1.363$$

which follows F-distribution with (12, 9) degrees of freedom.

The critical value $F_{0.05}(12,9) = 3.07$

Since the calculated value of F is less than the critical value therefore we may accept the null hypothesis at 5% level and conclude that there is no significant difference between the production capacity of the two machines.

Example 10.3: It is known that the mean diameters of rivets produced by two firms A and B are practically the same but the standard deviations may differ. For 22 rivets produced by firm A, the standard deviation is 2.9 mm, while for 16 rivets manufactured by firm B, the standard deviation is 3.8 mm. Compute the statistic you would use to test whether the products of firm A have the same variability as those of firm B and test its significance.

Solution: Given that

$$n_1 = 22, \qquad s_1 = 2.9\,mm.$$

$$n_2 = 16, \qquad s_2 = 3.8\,mm.$$

We state null and alternative hypotheses as

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ i.e., the products of both the firms A and B have the same variability.

Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Now, we have

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{22 \times (2.9)^2}{21} = 8.805$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{16 \times (3.8)^2}{15} = 15.393$$

Since $S_2^2 > S_1^2$, under H_0, the test statistic is

$$F = \frac{S_2^2}{S_1^2}$$
$$= \frac{15.393}{8.805} = 1.74821$$

which follows F-distribution with (15, 21) degrees of freedom.

The critical value $F_{0.05}(15, 21) = 2.20$

Since the calculated value of F is less than the critical value therefore we may accept the null hypothesis at 5% level and conclude that the hypothesis of equal variability may be accepted.

Example 10.4: Suppose a company manufacturing light bulbs is using two different processes A and B. The life of the light bulbs of process A has a normal distribution with mean μ_1 and standard deviation σ_1. Similarly the process B has also normal distribution with mean μ_2 and standard deviation σ_2. The data pertaining to the two processes are given below:

Sample A	Sample B
$n_1 = 16$	$n_2 = 21$
$\overline{x_1} = 1200hr.$	$\overline{x_2} = 1300hr.$
$s_1 = 60hr.$	$s_2 = 50hr.$

Test the hypothesis that the variability of the two processes is the same.

Solution: We state null and alternative hypotheses as

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ i.e., there is no significant difference in the variability of the two processes..

Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Now, we have

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{16 \times (60)^2}{16 - 1} = 3840$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{21 \times (50)^2}{21 - 1} = 2625$$

Since $S_1^2 > S_2^2$, under H_0, the test statistic is

$$F = \frac{S_1^2}{S_2^2}$$
$$= \frac{3840}{2625} = 1.46$$

which follows F-distribution with (15, 21) degrees of freedom.

The critical value $F_{0.05}(15, 20) = 2.20$

Since the calculated value of F is less than the critical value therefore we may accept the null hypothesis at 5% level and conclude that there is no significant difference in the variability of the two samples.

Example 10.5: Can the following two samples be regarded as coming from the same normal population?

Sample	Size	Sample mean	Sum of squares of deviations from the mean
1	10	12	120
2	12	15	314

Solution: A normal population has two parameters, viz., mean μ and standard deviation σ. To test whether two independent samples have been drawn from the same normal population, we have to test:

(i) the equality of population means

and (ii) the equality of population variances.

We set up the null hypothesis $H_0 : \mu_1 = \mu_2 \; and \; \sigma_1^2 = \sigma_2^2$ i.e., the two samples can be regarded as drawn from the same normal population.

To test the equality of population means we apply the $t-$test (sizes of both samples are small) and equality of populations variances can be performed with the help of F-test. Since test assumes, therefore first we apply F-test and then apply test.

F-test:

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$

Given

$n_1 = 10, \quad n_2 = 12; \quad \overline{x_1} = 12, \quad \overline{x_2} = 15$

$\sum (x_1 - \overline{x_1})^2 = 120, \qquad \sum (x_2 - \overline{x_2})^2 = 314$

$\therefore S_1^2 = \dfrac{\sum (x_1 - \overline{x_1})^2}{n_1 - 1} \qquad \therefore S_2^2 = \dfrac{\sum (x_2 - \overline{x_2})^2}{n_2 - 1}$

$= \dfrac{120}{9} \qquad = \dfrac{314}{11}$

$= 13.33 \qquad = 28.55$

Since $S_2^2 > S_1^2$, under H_0, the test statistic is

$$F = \frac{S_2^2}{S_1^2}$$

$$= \frac{28.55}{13.33} = 2.14$$

which follows F-distribution with (10, 9) degrees of freedom.

The critical value $F_{0.05}(11,9) = 3.10$

Since the calculated value of F is less than the critical value therefore we may accept the null hypothesis at 5% level and the assumption of equality of population variances is valid.

Since $\sigma_1^2 = \sigma_2^2$, we can now apply $t-$test to test $H_0 : \mu_1 = \mu_2$

$t-$Test:

Under the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$, the test statistic is

$$t = \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$$

where $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x_1 - \overline{x_1})^2 + \sum (x_2 - \overline{x_2})^2\right]$$

$$= \frac{1}{20}[120 + 314]$$

$$= 21.7$$

$$\therefore \quad t = \frac{12 - 15}{\sqrt{21.7\left(\frac{1}{10} + \frac{1}{12}\right)}} = \frac{-3}{\sqrt{21.7 \times \frac{11}{60}}}$$

$$= -1.51$$

$$|t| = 1.51$$

The degrees of freedom = $n_1+n_2-2=20$

The critical value $t_{0.05}(20)$ = 2.086.

Since the calculated value of $|t|$ is less than the critical value of t at 5% level for 20 d.f., therefore we may accept the null hypothesis at 5% level of significance.

Since both the hypotheses $H_0 : \mu_1 = \mu_2$ and $H_0 : {\sigma_1}^2 = {\sigma_2}^2$ are accepted, therefore we may regard the given samples to have been drawn from the same normal population.

CHECK YOUR PROGRESS

Q 2: Determine the degrees of freedom of F -value for a sample of six observations in the numerator and four in the denominator? What is the critical value of for two-tailed at 5% level of significance?

Q 3: A random sample of eight observations from the first population resulted in a standard deviation of 10. A random sample of six observations from the second population resulted in a standard deviation of 7. At the 0.01 significance level, is there a significant difference in the variation of the two populations?

(a) State the null hypothesis and alternative hypothesis, (b) State the decision rule, (c) Compute the test statistic. (d) What is your decision regarding?

15.7 ANALYSIS OF VARIANCE

In unit 9, we have introduced hypothesis testing procedures to test the significance of difference between two sample means or equivalently to test if there is any significant difference between the two population means from which the samples have been drawn. For this, one test statistic employed was the t-test where we assumed that the two populations from which the samples were drawn had the same variance. But in real life, there may be situations when instead of comparing two sample means, a researcher has to compare three or more than three sample means (specifically, more than two). A researcher may have to test whether the three or more sample means computed from the three populations are equal. In other words, the null hypothesis can be that three or more population means are equal as against the alternative hypothesis that these population means are not equal. For example, suppose that a researcher wants to measure work attitude of the employees in four organizations. The researcher has prepared a questionnaire consisting of 10 questions for measuring the work attitude of employees. A five-point rating scale is used with 1 being the lowest score and 5 being the highest score. So, an employee can score 10 as the minimum score and 50 as the maximum score. The null hypothesis can be set as all the means are equal (i.e., there is no difference in the degree of work attitude of the employees) as against the alternative hypothesis that at least one of the means is different from the others (there is a significant difference in the degree of work attitude of the employees). In this situation, analysis of variance technique is used. "In its simplest form analysis of variance may be regarded as an extension or development of the $t-test$." Analysis of variance technique makes use of F distribution (F-statistic). Some more examples are presented below where we are required to test the equality of means of three or more populations. For example, whether:

(1) The average life of light bulbs being produced in three different plants is the same.

(2) All the three varieties of fertilizers have the same impact on the yield of rice.

(3) The level of satisfaction among the participants in all IIMs is the same.

(4) The impact of training on salesmen trained in three institutes is the same.

(5) The service time of a transaction is the same on four different counters in a service unit.

(6) The average price of different commodities in four different retail outlets is the same.

(7) Performance of salesmen in four zones is the same.

The word 'analysis of variance' is used since the technique involves first finding out the total variation among the observations in the collected data, then assigning causes or components of variation to various factors and finally drawing conclusions about the equality of means. Thus Analysis of variance or ANOVA can be defined as a technique of testing hypotheses about the significant difference in several population means. This statistical technique was developed by R.A.Fisher. The main purpose of analysis of variance is to detect the difference among various population means based on the information gathered from the samples (sample means) of the respective populations.

15.8 ASSUMPTIONS OF ANALYSIS OF VARIANCE

Analysis of variance is based on some assumptions.

(i) Each sample is a simple random sample.

(ii) Populations from which the samples are selected are normally distributed. However, in case of large samples this assumption is not required.

(iii) Samples are independent.

(iv) The population variances are identical, i.e., $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$.

Computation of Test Statistic:

The beauty of the technique of analysis of variance is that it performs the test of equality of more than two population means by actually analyzing the variance. In simple terms, ANOVA decomposes the total variation into two components of variation, namely, variation between the samples known

as the mean square between samples and variation within the samples known as the mean square within samples. The variance ratio denoted by F is given by:

$$F = \frac{Mean\ square\ between\ samples\ or\ groups}{Mean\ square\ within\ samples\ or\ groups}$$

If the calculated value of F is greater than the critical value of, we must reject the null hypothesis. In case the calculated value of is less than the critical value of, we are to retain or accept the null hypothesis.

Analysis of variance table (ANOVA):

The table showing the sources of variation, the sum of squares, degrees of freedom, mean squares and the formula for the F statistic is known as ANOVA table.

15.9 CLASSIFICATION OF ANALYSIS OF VARIANCE

ANOVA is mainly carried on under the following two classifications:

(i) One-way classification
(ii) Two- way classification

Variance and its different components may be obtained in each of the two types of classification by:

(a) Direct Method, (b) Short-cut Method and (c) Coding Method

15.9.1 ONE-WAY CLASSIFICATION

Many business applications involve experiments in which different populations (or groups) are classified with respect to only one attribute of interest such as (i) percentage of marks secured by students in a course, (ii) flavor preference of ice-cream by customers, (iii) yield of crop due to varieties of seeds, and so on. In all such cases observations in the sample data are classified into several groups based on a single attribute (i.e., criterion) and are termed one-way classification of sample data.

Under this one-way classification we set up the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ where $\mu_1, \mu_2, ..., \mu_k$ are the arithmetic means

of the populations from which the k samples are drawn.

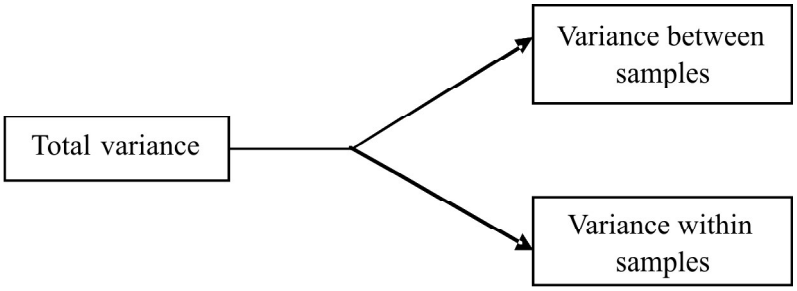
The alternative hypothesis is $H_1 : \mu_1 \neq \mu_2 \neq ..... \neq \mu_k$

After formulation of the null hypothesis and alternative hypothesis one needs to calculate the following by using any one of the above mentioned methods.

(i) Variance between the samples.
(ii) Variance within the samples.
(iii) Total variance by summing (i) and (ii) which may also be calculated directly for verification of calculations.

(iv) F-ratio (or F-statistic)

Thus in this process, the total variance can be divided into two additive and independent parts as shown:



After calculating the test statistic F, it should be compared with the critical value of F at a specified level of significance α for $(k-1, n-k)$ degrees of freedom and on the basis of this comparison; accordingly the decision to accept or reject the null hypothesis is taken.

(a) Direct Method

I. Calculation of variance between samples

It is the sum of squares of the deviations of the means of various samples from the grand mean. The procedure of calculating the variance between the samples is as shown below:

Observations	Number of samples				
	1	2	.....j	.....	k
1	x_{11}	x_{12}	..... x_{1j}	.....	x_{1k}
2	x_{21}	x_{22}	.....x_{2j}	.....	x_{2k}
.	.	.	.		.
i	x_{i1}	x_{i2}	.....x_{ij}	.....	x_{ik}
.	.	.			.
N	x_{n1}	x_{n2}	.....x_{nj}	.....	x_{nk}
Total	T_1	T_2	.....T_j	.....	T_k
A.M.	$\overline{x_1}$	$\overline{x_2}$	..... $\overline{x_j}$	.....	$\overline{x_k}$

Where

$$T_j = \sum_{i=1}^{n} x_{ij}\ , \qquad T = \sum_{j=1}^{k} T_j\ , \qquad \overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad \text{and} \quad \bar{\bar{x}} = \frac{1}{nk}\sum_{j=1}^{k} \overline{x_j}$$

The steps in calculating variance between samples are given by:

(i) In the first step, we need to calculate the mean of each sample i.e., $\overline{x_1}, \overline{x_2}, ..., \overline{x_k}$ of all the k samples.

(ii)Next, the grand mean is calculated by using the formula

$$\bar{\bar{x}} = \frac{\overline{x_1} + \overline{x_2} + ... + \overline{x_k}}{k} = \frac{T}{N}$$

where T =Grand total of all observations.

N =Total number of observations in all k samples.

(iii) In step 3, the difference between the mean of each sample and grand mean is calculated, that is we calculate $\overline{x_1} - \bar{\bar{x}}, \overline{x_2} - \bar{\bar{x}}, ..., \overline{x_k} - \bar{\bar{x}}$

(iv) In step 4, we square the deviations obtained in step (iii) and multiply by the number of items in the corresponding sample and then add the total. This total gives the sum of the squares of the deviations between the samples (or between the columns) and it is denoted by SSB or SSC

$$\text{Thus} \quad \text{SSB} = \sum_{j=1}^{k} n_j \left(\overline{x_j} - \bar{\bar{x}}\right)^2$$

where k is the number of groups or samples being compared, n_j the number of observations in group j, $\overline{x_j}$ the sample mean of group , and the grand mean.

(v) In the last step, the total obtained in step (iv) is divided by the degrees of freedom. The degrees of freedom is one less than the total number of samples. If there are k samples, the degrees of freedom will be $\nu = k-1.$ When the sum of squares obtained in step (iv) is divided by the number of degrees of freedom, the result is called mean square which is denoted by MSB or MSC. MSB indicates the degree of explained variance due to sampling fluctuations.

$$\text{Thus} \qquad \text{MSB}=\frac{SSB}{k-1}$$

II. Calculation of Variance within the samples:

This is usually referred to as the sum of squares within samples. The variance within samples measures the difference within the samples due to chance. This is usually denoted by SSE. The steps involved are the following:

(i) The first step is to calculate the mean values $\overline{x_1}, \overline{x_2}, ..., \overline{x_n}$ of all k samples,

(ii) Calculate the deviations of the various observations of k samples from the mean values of the respective samples,

(iii) Square all the deviations obtained in (ii) and find the total of these squared deviations. This total gives the sum of the squares of deviations within the samples or the sum of squares due to error. It is denoted by SSW or SSE.

$$\text{Thus} \qquad \text{SSW or SSE}=\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\overline{x_j}\right)$$

where x_{ij} is the ith observation in the jth group, $\overline{x_j}$ the sample mean of jth group, k the number of groups being compared, and N the total number of observations in all the groups.

(iv) As the last step, divide the total squared deviations obtained in step (iii)

by the degrees of freedom and obtain the mean square. The number of degrees of freedom can be calculated as the difference between the total number of observations and the number of samples. If there are N observations and k samples then the degrees of freedom is $\nu = N - k$.

Thus $$\text{MSW} = \frac{SSE}{N-k}$$

III. Calculate total sum of squares: The total variation is equal to the sum of the squared difference between each observation (sample value) and the grand mean $\bar{\bar{x}}$ and is often referred to as SST.

Thus SST can be calculated as:

SST=SSB+SSE

III. Calculation of the Test Statistic F:

When the null hypothesis is true, both mean squares MSB (MSC) and MSW (MSE) are the independent unbiased estimates of the same population variance σ^2. Hence, the test statistic is

$$F = MSB/MSW \text{ Or, } F = MSC/MSE$$

which follows F-distribution with degrees of freedom $(k-1, N-k)$

F is the ratio between the greater variance to the smaller variance. Generally, variance between the samples (MSB or MSC) is greater than the variance within the samples (MSW or MSE). But if MSW>MSB then the reverse ratio of F will be used i.e.,

$$F = MSW/MSB \quad or, F = MSE/MSC$$

IV. Conclusion:

To compare the calculated value of F with the critical value (tabulated) of F for $(k-1, N-k)$ degrees of freedom at the specified level of significance, usually 5% or 1% level. If the calculated value is greater than the tabulated value, the null hypothesis H_0 is rejected and conclude that all population means are not equal. Otherwise, the null hypothesis is accepted.

For analyzing variance in case of one way classification the following table known as the Analysis of Variance Table (or ANOVA Table) is constructed.

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)	Test Statistic
Between samples	SSB	$k-1$	MSB= $\frac{SSB}{k-1}$	$F=\frac{MSB}{MSW}$
Within samples	SSW	$N-k$	MSW= $\frac{SSW}{N-k}$	
Total	SST	$N-1$		

(b) Short-cut Method: The calculation of F-statistic (variance ratio) by using the direct Method is very time consuming. In practice, a short-cut Method based on sum of squares of the individual values (observations) is usually used. The computational work is much minimized in this method. The method is more convenient when some or all the sample means and the grand mean are fractional. The various steps involved in the calculations of variance ratio are the following:

(i) Calculate the grand total of all observations in samples, T

$$T=\sum x_1+\sum x_2+.....+\sum x_k$$

(ii) Calculate the correction factor $CF=\frac{T^2}{N}$; N= Total observations in samples.

(iii) Find the sum of the squares of all observations in samples from each of k samples and subtract CF from this sum to obtain the total sum of the squares of deviations SST:

$$SST=\left(\sum x_1^2+\sum x_2^2+.....+\sum x_k^2\right)-CF$$

$$SSB=\frac{\left(\sum x_j\right)^2}{n_j}-CF$$ where n_j is the size of the j[th] sample.

And $SSE=SST-SSB$

(iv) MSB, MSW and F are obtained as in the Direct method and decision either to accept or to reject is exactly same as taken in case of Direct method.

(c) Coding Method:

When the magnitudes of the observations (values) in various samples are large then Short-cut-Method also becomes tedious. The variance ratio F (or the F statistic) has the property that it remains unchanged if all the values are multiplied or divided by a common factor or if a common factor is added to or subtracted from each value. Owing to this property big figures (values or observations) are reduced in magnitude by subtraction or division and the computational work is much simplified without altering the value of F. When the magnitude of data is reduced by subtractions or divisions, it is said that data have been coded to simplify calculations.

Example 10.6: The following data give the yield on 12 plots of land in three samples of 4 plots each, under three varieties of fertilizers A, B and C

A	B	C
25	20	24
22	17	26
24	16	30
21	16	20

Test whether there is any significant difference in the average yields of land under three varieties of fertilizers.

Solution: Here first we set up null and alternative hypotheses

Null hypothesis H_0 : There is no significant difference in the average yields under the three varieties.

Alternative hypothesis H_1 : There is significant difference in the average yield under the three varieties.

We calculate sample means, the variance between the samples and the variance within the samples by using the Direct Method.

	Sample I X_1	Sample II X_2	Sample III X_3
	25	20	24
	22	17	26
	24	16	30
	21	16	20
Total	92	72	100

$$\overline{X_1} = \frac{\sum X_1}{n_1} = \frac{92}{4} = 23 \quad \overline{X_2} = \frac{\sum X_2}{n_2} = \frac{72}{4} = 18 \quad \overline{X_3} = \frac{\sum X_3}{n_3} = \frac{100}{4} = 25$$

Grand mean ($\overline{\overline{X}}$) $= \dfrac{\overline{X_1} + \overline{X_2} + \overline{X_3}}{3}$

$$= \frac{23 + 18 + 25}{3} = \frac{66}{3} = 22$$

$$\text{SSB} = n_1\left(X_1 - \overline{\overline{X}}\right)^2 + n_2\left(X_2 - \overline{\overline{X}}\right)^2 + n_3\left(X_3 - \overline{\overline{X}}\right)^2$$

$$= 4(23 - 22)^2 + 4(18 - 22)^2 + 4(25 - 22)^2$$

=104

Degrees of freedom, $\nu_1 = k - 1 = 2$

Now MSW=Mean square between the samples

$$= SSB / \nu_1 = \frac{104}{12} = 52$$

Calculation for SSW

Sample I		Sample II		Sample III	
X_1	$\left(X_1 - \overline{X_1}\right)^2$	X_2	$\left(X_2 - \overline{X_2}\right)^2$	X_3	$\left(X_3 - \overline{X_3}\right)^2$
25	4	20	4	24	1
22	1	17	1	26	1
24	1	16	4	30	25
21	4	19	1	20	25
	10		10		52

SSW = Sum of squares within the samples

$$= \sum\left(X_1 - \overline{X_1}\right)^2 + \sum\left(X_2 - \overline{X_2}\right)^2 + \sum\left(X_3 - \overline{X_3}\right)^2$$

= 10+10+52=72

The degrees of freedom, $\nu_2 = N - k = 9$

MSW=72/9=8

Now we prepare the Analysis of Variance Table

Analysis of Variance Table (ANOVA)

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)	Test Statistic
Between samples	104	2	52	F=52/8=6.5
Within samples	72	9	8	
Total	SST=176	N-1=11		

The critical value of F for $\nu_1 = 2$ and $\nu_2 = 9$ at 5% level i.e., $F_{0.05}(2,9)$ is 4.26.

Since the calculated value of F (i.e., 6.5) is greater than the critical value $F_{0.05}(2,9)$, herefore we reject the null hypothesis at 5% level and conclude that the difference in the average yields under the three varieties is significant.

Solution of the above problem by Short cut Method:

After formulating the null and alternative hypotheses as in the Direct Method one can proceed with the Short cut Method as follows:

Calculation of T, SST, SSB, SSW

Sample		Sample I		Sample I	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
25	625	20	400	24	576
22	484	17	289	26	676
24	576	16	256	30	900
21	441	19	361	20	400
92	2126	72	1306	100	2552

T = Sum of all the values in the three samples

$$= \sum X_1 + \sum X_2 + \sum X_3 = 92 + 72 + 100 = 264$$

Correction factor = $\frac{T^2}{N} = \frac{(264)^2}{12} = 5808$

SST = Total sum of squares

$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2\right) - \frac{T^2}{N}$

= (2126+1306+2552)-5808 = 176

SSB = Sum of squares between the samples

$= \frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_3\right)^2}{n_3} - \frac{T^2}{N}$

= 2116+1296+2500 - 5808

= 104

Degrees of freedom = k-1=2

SSW = Sum of squares within the samples.

= SST-SSB = 176-104 = 72

Degrees of freedom = N – k=12-3=9

Now proceeding with the ANOVA table as in the Direct Method, one can arrive at the same conclusion.

Example 10.7: A study investigated the perception of corporate ethical values among individuals specializing in marketing. Using $\alpha = 0.05$ and the following data (higher scores indicate higher ethical values), test for significant differences in perception among three groups

Marketing Manager	Marketing Research	Advertising
6	5	6
5	5	7
4	4	6
5	4	5
6	5	6
4	4	6

Solution: We set up the null hypothesis H_0 that there is no significant difference in ethical values among individuals specializing in marketing.

Now we calculate sample means, the variance between the samples and the variance within the samples by using the Short cut Method.

Marketing Manager		Marketing Research		Advertising	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
6	36	5	25	6	36
5	25	5	25	7	49
4	16	4	16	6	36
5	25	4	16	5	25
6	36	5	25	6	36
4	16	4	16	6	36
30	154	27	123	36	218

T = Sum of all the values in the three samples

$$= \sum X_1 + \sum X_2 + \sum X_3 = 30 + 27 + 36 = 93$$

CF = Correction factor = $\frac{T^2}{N} = \frac{(93)^2}{18} = 480.50$

SST = Total sum of squares

$$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2\right) - \frac{T^2}{N}$$

= (154+123+218)-480.50 = 14.50

SSB = Sum of squares between the samples

$$= \frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_3\right)^2}{n_3} - \frac{T^2}{N}$$

$$= \frac{(30)^2}{6} + \frac{(27)^2}{6} + \frac{(36)^2}{6} - 480.50$$

= 150+121.5+216 -480.50

= 7

Degrees of freedom = k-1=2

SSW = Sum of squares within the samples.

= SST-SSB = 14.50-7 = 7.50

Degrees of freedom = N – k=18-3=15

Now we prepare the Analysis of Variance Table

Analysis of Variance Table (ANOVA)

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)	Test Statistic
Between samples	7	2	7/2=3.5	F=3.5/0.5=7
Within samples	7.5	15	7.5/15=0.5	
Total	14.5	N-1=17		

The critical value of F for $\nu_1 = 2$ and $\nu_2 = 15$ at 5% level i.e., $F_{0.05}(2,15)$ is 3.68.

Since the calculated value of F (i.e., 7) is greater than the critical value $F_{0.05}(2,15)$, therefore we reject the null hypothesis at 5% level and conclude that there is significant difference in ethical values among individuals specializing in marketing.

Example 10.8: Three samples of five, four and five scooters are drawn respectively from brands A, B and C manufactured by three companies. The mileage of these scooters (in kilometer) is given below:

Test whether there is any significant difference between the mean mileages of the three brands of scooters. Use coding method.

A	B	C
45	41	44
42	40	42
43	42	38
44	43	43
42		39

Solution: We set up the null hypothesis H_0 and the alternative hypothesis H_1 as

H_0: There is no significant difference between the mean mileages of the three brands of scooters.

H_1: There is significant difference between the mean mileages of the three brands of scooters.

Let us subtract 40 from each of the sample values and let X_1, X_2 and X_3 denote respectively the coded sample values in respect of brand A, B and C respectively. We make the following table:

Calculation of T, SST, SSB, SSW

Sample		Sample II		Sample III	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
5	25	1	1	4	16
2	4	0	0	2	4
3	9	2	4	-2	4
4	16	3	9	3	9
2	4	..	..	-1	1
16	58	6	14	6	34

Now T = Sum of all the values in the three samples.

$= \sum X_1 + \sum X_2 + \sum X_3 = 16 + 6 + 6 = 28$

CF = Correction factor = $\frac{T^2}{N} = \frac{(28)^2}{14} = 56$

SST = Total sum of squares

$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2\right) - \frac{T^2}{N}$

= (58+14+34)-56 =50

SSB = Sum of squares between the samples

$= \frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_3\right)^2}{n_3} - \frac{T^2}{N}$

$= \frac{(16)^2}{5} + \frac{(6)^2}{4} + \frac{(6)^2}{5} - 56$

= 51.2+9+7.2 -56

= 11.4

Degrees of freedom, ν_1 = k-1=2

SSW = Sum of squares within the samples.

= SST-SSB = 50-11.4 =38.6

Degrees of freedom, ν_2 = N – k=14-3=11

Analysis of Variance Table (ANOVA)

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)	Test Statistic
Between samples	11.4	2	11.4/2=5.7	F=5.7/3.51=1.624
Within samples	38.6	11	38.6/11=3.51	
Total	50	N-1=13		

The critical value of F for $\nu_1 = 2$ and $\nu_2 = 11$ at 5% level i.e., $F_{0.05}(2,11)$ is 3.98.

Since the calculated value of F (i.e., 1.624) is less than the critical value $F_{0.05}(2,11)$, therefore we may accept the null hypothesis at 5% level and conclude that there is no significant difference between the mean mileage of the three brands of scooters.

CHECK YOUR PROGRESS

Q 4. State whether the following statements are true or false:

(i) Analysis of variance is used to test the hypothesis that the means of several populations do not differ.

(ii) For analysis of variance, samples drawn from populations need not be independent.

(iii) The number of parts in which total variance in a one-way analysis of variance partitioned is 2.

(iv) The sum of squares within a one-way analysis of variance is SST+SSB.

(v) The test statistic used for equality of k population means is MSB/MSE.

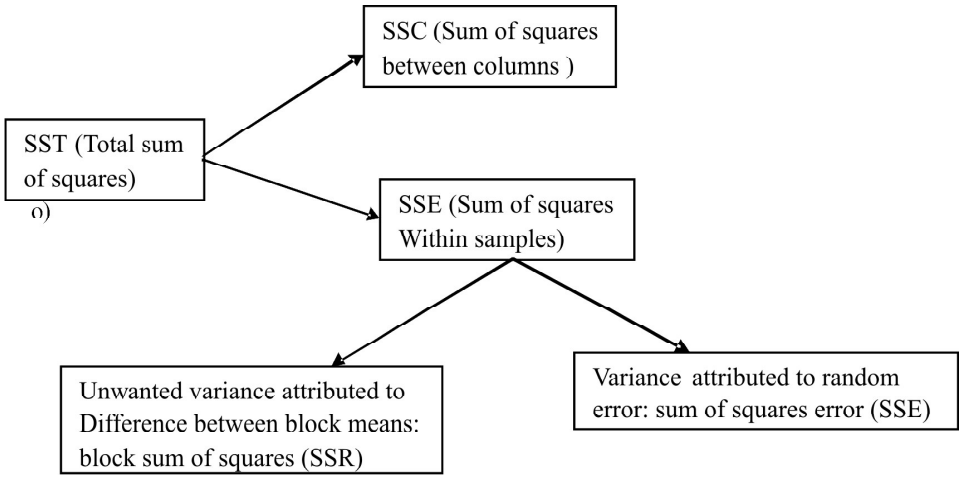
15.9.2 TWO-WAY CLASSIFICATION

In one-way ANOVA the total variation is partitioned into two components: variations between the samples or columns and variation within the samples due to random error. However, there might be a possibility that some of the variation left in the random error from one-way analysis of variation was not due to random error or chance but due to some other measurable factors. If this measurable factor is included in the MSE, it will

result in an increase in the MSE. Any increase in the MSE would result in a small value of F (MSE being a denominator in the F statistic formula), which would ultimately lead to the acceptance of the null hypothesis.

The two-way classification is a mere extension of the one-way ANOVA. It can be used to investigate two criteria (factors) of interest for testing the difference between sample means. For example, we can study the effect of the two factors namely different varieties of seeds and different varieties of fertilizers on the yield of a crop in several plots of land.

The total variance or the sum of squares of variation in a two-way classification consists of three parts: (i) SSC i.e., the sum of squares between the columns or variance between the columns, (ii) SSR i.e., the sum of squares between the rows or variance between the rows and (iii) SSE i.e., the sum of squares for the residuals or residual variance or the sum of squares due to error. Thus the partitioning of the total variation in the sample data is shown as



The total number of degrees of freedom = cr-1,

where c and r are the number of columns and the number of rows respectively.

Degrees of freedom for columns = c-1

Degrees of freedom for rows = r-1

Degrees of freedom for residuals= (cr-1)-(c-1)-(r-1)

= (c-1)(r-1)

The general ANOVA Table in case of two-way classification is given below:

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)
Between columns	SSC	$c-1$	MSC=SSC/(c-1)
Between rows	SSR	$r-1$	MSR=SSR/(r-1)
Residual	SSE	$(c-1)(r-1)$	MSE=SSE/(c-1)(r-1)
Total	SST	$N-1$	

The test statistic F for variance between columns is given by

$F = \frac{MSC}{MSE} = \frac{Variance\ between\ columns}{Residual\ variance}$, provided MSC > MSE

If the calculated value of F is greater than the critical value $F_{0.05}((c-1),(c-1)(r-1))$ then we reject the null hypothesis at 5% level and conclude that the difference between the column values is significant.

If MSE > MSC then we shall take $F = \frac{MSE}{MSC}$

Again the test statistic F for variance between rows is given by

$F = \frac{MSR}{MSE} = \frac{Variance\ between\ rows}{Residual\ variance}$ provided MSR > MSE

If MSE > MSR then we shall take $F = \frac{MSE}{MSR}$

If the calculated value of F is greater than the critical value $F_{0.05}((r-1),(c-1)(r-1))$ then we reject the null hypothesis at 5% level and conclude that the difference between the row values is significant.

Example 10.9: The following table gives the number of refrigerators sold by 4 salesmen in three months May, June and July:

Month	Salesmen			
	A	B	C	D
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

Is there any significant difference in the sales made by the four salesmen?
Is there a significant difference in the sales made during different months?

Solution: Let us take the null hypothesis that (a) there is no significant difference between sales made by the four salesmen and (b) there is no significant difference in the sales made during different months. The given data are coded by subtracting 40 from each observation. The coded data are given below:

Month	Salesman								Row sum
	$A(x_1)$	x_1^2	$B(x_2)$	x_2^2	$C(x_3)$	x_3^2	D (x_4)	x_4^2	
May	10	100	0	0	8	64	-1	1	17
June	6	36	8	64	10	100	5	25	29
July	-1	1	4	16	0	0	-1	1	2
Column sum	15	137	12	80	18	164	3	27	48

From the above table, we work out the requisite sums of squares for preparing the ANOVA Table as follows:

T = Sum of all observations in three samples of months = 48

CF = Correction factor = $\frac{T^2}{N} = \frac{(48)^2}{12} = 192$

SSC = Sum of squares between salesmen (Columns)

$$= \frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_3\right)^2}{n_3} + \frac{\left(\sum X_4\right)}{n_4} - \frac{T^2}{N}$$

$$= \frac{(15)^2}{3} + \frac{(12)^2}{3} + \frac{(18)^2}{3} + \frac{(3)^2}{3} - 192$$

= (75+48+108+3) - 192

= 42

Degrees of freedom, $df_c = c - 1 = 4 - 1 = 3,$

SSR = Sum of squares between months (rows)

$$= \frac{(17)^2}{4} + \frac{(29)^2}{4} + \frac{(2)^2}{4} - 192$$

= (72.25+210.25+1) - 192

= 91.5

Degrees of freedom, $df_r = r - 1 = 3 - 1 = 2$

SST = Total sum of squares

$$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2\right) - \frac{T^2}{N}$$

= (137+80+164+27) –192

= 216

Degrees of freedom, $df = N - 1 = 12 - 1 = 11$

Residual or Error Sum of Squares, i.e., $SSE = SST - (SSC + SSR)$

= 216 – (42+91.5)

= 82.5

Degrees of freedom, $df_e = (c-1)(r-1) = 3 \times 2 = 6$

Thus $$MSC = \frac{SSC}{(c-1)} = \frac{42}{3} = 14$$

$$MSR = \frac{SSR}{(r-1)} = \frac{91.5}{2} = 45.75$$

$$MSE = \frac{SSE}{(c-1)(r-1)} = \frac{82.5}{6} = 13.75$$

The above summarized results are presented in the following ANOVA Table as given below:

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squarcs(MS)	Test statistic
Between salesmen	42.0	3	14.00	$F_{treatment} = 14/13.75$ =1.018
Between months	91.5	2	45.75	$F_{block} = 45.75/13.75$
Residual error	82.5	6	13.75	=3.327
Total	216	11		

(a) The critical value of F at 5% level for (3, 6) degrees of freedom is 4.75.

Since the calculated value of $F_{treatment} = 1.018$ is less than the critical value, therefore we may accept the null hypothesis and conclude that sales made by the salesmen do not differ significantly.

(b) The critical value of F at 5% level for (2, 6) degrees of freedom is 5.14.

Since the calculated value of $F_{block} = 3.327$ is less than the critical value, hence we may accept the null hypothesis and conclude that sales made

during different months do not differ significantly.

Example 10.10: To study the performance of three detergents and three different water temperatures, the following 'whiteness' readings were obtained with specially designed equipment:

Water Temperature	Detergent A	Detergent B	Detergent C
Cold water	57	55	67
Warm water	49	52	68
IIot water	54	46	58

Perform a two-way s analysis of variance, using 5% level of significance.

Solution: Let us state the null hypothesis that there is no significant difference in the performance of three detergents due to water temperature and vice-versa.

The data are coded by subtracting 50 from each observation. The data in coded form are given below:

Month	Detergent						Row sum
	$A(x_1)$	x_1^2	$B(x_2)$	x_2^2	$C(x_3)$	x_3^2	
Cold water	7	49	5	25	17	289	29
Warm water	-1	1	2	4	18	324	19
Hot water	4	16	-4	16	8	64	8
Column sum	10	66	3	45	43	677	56

T= Sum of all observations in three samples of months = 56

CF = Correction factor = $\frac{T^2}{N} = \frac{(56)^2}{9} = 348.44$

SSC = Sum of squares between detergents (Columns)

$$= \frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_3\right)^2}{n_3} - \frac{T^2}{N}$$

$$= \frac{(10)^2}{3} + \frac{(3)^2}{3} + \frac{(43)^2}{3} - 348.44$$

= (33.33+3+616.33) – 348.44

= 304.22

Degrees of freedom, $df_c = c - 1 = 3 - 1 = 2$,

SSR = Sum of squares between water temperature (rows)

$$= \frac{(29)^2}{3} + \frac{(19)^2}{3} + \frac{(8)^2}{3} - 348.44$$

= (280.33+120.33+21.33) – 348.44

= 73.55

Degrees of freedom, $df_r = r - 1 = 3 - 1 = 2$

SST = Total sum of squares

$$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2\right) - \frac{T^2}{N}$$

= (66+45+677) – 348.44

= 439.56

Degrees of freedom, $df = N - 1 = 9 - 1 = 8$

$$\therefore SSE = SST - (SSC + SSR)$$

= 439.56 – (304.22+73.55)

= 61.79

Degrees of freedom, $df_e = (c-1)(r-1) = 2 \times 2 = 4$

Thus

$$MSC = \frac{SSC}{(c-1)} = \frac{304.22}{2} = 152.11$$

$$MSR = \frac{SSR}{(r-1)} = \frac{73.55}{2} = 36.775$$

$$MSE = \frac{SSE}{(c-1)(r-1)} = \frac{61.79}{4} = 15.447$$

The ANOVA Table is given below:

Sources of Variation	Sum of squares (SS)	Degrees of Freedom (*d.f.*)	Mean Squares(MS)	Test statistic
Between detergents(columns)	304.22	2	152.110	$F_{treatment} = 152.11/15.447$ =9.847
Between temperature (rows)	73.55	2	36.775	$F_{block} = 36.775/15.447$ =2.380
Residual error	61.79	4	15.447	
Total	439.56	8		

The critical value of F at 5% level for (2, 4) degrees of freedom is 6.94.

(a) Since the calculated value of $F_{treatment} = 9.847$ is greater than the critical value $F_{0.05}(2,4) = 6.94$, therefore we reject the null hypothesis and conclude that there is significant difference between the performance of the three detergents.

(b) Since the calculated value of $F_{block} = 2.380$ is less than the critical value, hence we may accept the null hypothesis and conclude that the water temperature do not make a significant difference in the performance of the detergent.

Example 10.11: The following table gives the number of subscribers added by four major telecom players in India, in the months of August, September, October and November 2005. The data are given in 000's and are rounded off to the nearest 100, and are thus in lakhs.

Months	Company			
	Bharati	BSNL	Tata Indicom	Reliance
August	6	6	2	5
September	7	6	2	3
October	7	6	6	4
November	7	8	7	4

(Source: Indianfoline.com on India Mobile Industry)

Find out (use 5% level of significance)

(i) If the four companies significantly differ in their performance.

(ii) Is there significant difference between the months?

Solution: First of all, we set up the null hypothesis that (a) there is no significant difference between the companies in their performance and (b) there is no significant difference between the months.

Now we prepare the following table for calculation:

Months	Company				Total	Average
	Bharati	BSNL	Tata Indicom	Reliance		
August	6	6	2	5	19	4.75
September	7	6	2	3	18	4.50
October	7	6	6	4	23	5.75
November	7	8	7	4	26	6.50
Total	27	26	17	16	Grand total=86	
Average	6.75	6.5	4.25	4	Grand mean=5.375	

From the above table, we work out the requisite sums of squares for preparing the ANOVA Table as follows:

T = Sum of all observations in three samples of months $= \sum\sum x_{ij} = 86$

CF = Correction factor = $\frac{T^2}{N} = \frac{\left(\sum\sum x_{ij}\right)^2}{16} = \frac{(86)^2}{16} = 462.25$

SST = Total sum of squares

$= \sum\sum x_{ij}^{\ 2} - CF$

$= (6^2 + 6^2 + \ldots\ldots\ldots + 4^2) - 462.25$

$= 514 - 462.25$

$= 51.75$

Degrees of freedom, $df = N - 1 = 16 - 1 = 15$

SSR = Sum of squares between months (rows)

$$= \frac{(19)^2}{4} + \frac{(18)^2}{4} + \frac{(23)^2}{4} + \frac{(26)^2}{4} - 462.25$$

$= 472.5 - 462.25$

$= 10.25$

Degrees of freedom, $df_r = r - 1 = 4 - 1 = 3$

SSC = Sum of squares between Companies (Columns)

$$= \frac{(27)^2}{4} + \frac{(26)^2}{4} + \frac{(17)^2}{4} + \frac{(16)^2}{4} - 462.25$$

$= 487.5 - 462.25$

$= 25.25$

Degrees of freedom, $df_c = c - 1 = 4 - 1 = 3,$

SSE=Residual or Error Sum of Squares

i.e., $SSE = SST - (SSC + SSR)$

= 51.75 – (10.25+25.25)

= 16.25

Degrees of freedom, $df_e = 15 - 3 - 3 = 9$

The above obtained summarized results are presented in the following ANOVA able:

Source of Variation	Sum of Squares	Degrees of Freedom(d.f.)	Mean Squares(M.S.)	F (Calculated)	$\alpha = 5\%$ (Tabulated)
Between Months	10.25	3	3.42	3.42/1.81 = 1.89	3.86
Between Companies	25.25	3	8.42	8.42/1.81 =4.65	3.86
Error	16.25	9	1.81		
Total	51.75	15			

With calculated and tabulated values of F we can derive following conclusions:

(a) Since the calculated value of $F_{Months} = 1.89$ for months is less than the critical value $F_{0.05}(3,9) = 3.86$, therefore we may accept the null hypothesis and conclude that there is no significant difference between the months in terms of total additional consumers.

(b) Since the calculated value of $F_{Companies} = 4.65$ for companies is greater than the critical value $F_{0.05}(3,9) = 3.86$, therefore we reject the null hypothesis and conclude that there is a significant difference between the companies in terms of adding subscribers.

CHECK YOUR PROGRESS

Q 5: The two null hypotheses to the following two-way ANOVA table are

H_{01} : The treatments means are equal i.e., $\mu_1 = \mu_2$

H_{02} : The block means are equal i.e., $\mu_1 = \mu_2 = \mu_3$

Source	SS	df	MS	F
Treatment	204.167			
Blocks	76.000	2	38.00	
Error		2		
Total	289.500	5		

(a) Complete the ANOVA table.

(b) Give your decision regarding the two sets of hypothese.



15.10 LET US SUM UP

The major characteristics of the F distribution are:

(1) It is never negative.

(2) It is continuous.

(3) It is positively skewed.

(4) It is based on two sets of degrees of freedom.

The distribution is used to test whether two population variances are the same. In this situation, the sampled populations must follow the normal distribution.

The distribution can also be used as given below:

(ii) F test is used to test the significance of an observed sample multiple correlation.

(iii) It is used to test the significance of an observed sample correlation ratio.

(iv) It is applied for testing the linearity of regression.

(v) It is used to test the equality of several population means.

This unit also focuses on a procedure called the analysis of variance, or ANOVA. It is a technique used to test simultaneously whether the means of several populations are equal. It uses the distribution as the distribution of the test statistic.

As an example of analysis of variance, a test could be conducted to

resolve whether there is any difference in effectiveness among five fertilizers (say) on the weight of popcorn ears. This type of analysis termed as one-factor ANOVA because we are able to draw conclusions about only one factor, known as treatment. On the other hand, if we want to draw conclusions about the simultaneous effects of more than one factor or variable, we use the two- factor ANOVA technique. Both the one –factor and two factor tests use thedistribution as the distribution of the test statistic.

15.11 FURTHER READINGS

1) Srivastava, T.N., Rego, S. (2008). Statistics for Management. New Delhi. Tata McGraw Hill Education Private Limited.

2) Sharma, J.K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.

3) Hazarika, P.L. (2016). Essential Statistics For Economics And Business Studies. New Delhi. Akansha Publishing House.

4) Lind, D.A., Marshal, W.G., Wathen, S.A. (2009) Statistical Techniques in Business and Economics. New Delhi. Tata McGraw Hill Education Private Limited.

5) Bajpai, N. (2014). Business Statistics. New Delhi. Pearson Education Ltd.

15.12 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q No 1: (i) True, (ii) False, (iii) True.

Ans. to Q No 2: The degrees of freedom of F -value are 5 in the numerator and 3 in the denominator.

The critical value of with (5, 3) degrees of freedom for two-tailed at 5% level of significance = 9.01

Ans. to Q No 3: (a) Null hypothesis is: There is no significant difference in the variation of the two populations i.e.,.

Alternative hypothesis is: There is significant difference in the variation of the two populations i.e.,

(b) Reject if 10.5, the critical value of with (7, 5) degrees of freedom for two-tailed at 5% level of significance.

(c) The test statistic $F = \frac{S_1^2}{S_2^2} = \frac{10^2}{7^2} = 2.04$

(d) H_0 may be accepted.

Ans. to Q No 4: (i) True, (ii) False, (iii) True, (iv) False, (v) True.

Ans. to Q No 5: (a) With the given information the ANOVA table is completed as follows:

Source	SS	df	MS	F
Treatment	204.167	1	$\frac{204.167}{1} = 204.167$	$\frac{204.167}{4.667}$ =43.75
Blocks	76	2	38	$\frac{38}{4.667}$ =8.14
Error	289.5-(204.167+76) =9.333	2	$\frac{9.333}{2} = 4.667$	
Total	289.5	5		

(b) Now, since $F_{Treatment} > F_{0.05}(1,2) = 18.5$, the critical value of F therefore reject H_{01}.

Again, since $F_{blocks} < F_{0.05}(2,2) = 19.0$, the critical value of thereforemay be accepted.

15.13 POSSIBLE QUESTIONS

1. Define F-distribution.
2. Write a note on properties and applications of F-distribution.
3. What is Analysis of variance? What are the chief assumptions of ANOVA?
4. Tests for breaking strength were carried out on two lots of 5 and 9 steel wires respectively. The variance of one lot was 230 and that of the other was 492. Is there a significant difference in their variability?
5. Two random samples drawn from normal population are:

Sample I : 20 16 26 27 23 22 18 24 25 19

Sample II : 27 33 42 35 32 34 38 28 41 43 30 37

Obtain estimates of the variances of the population and test whether the two population have the same variance?

6. An agriculture research organization wants to study the effect of four types of fertilizers on the yield of a crop. It divided the entire field into 24 plots of land and used fertilizer at random in 6 plots of land. Part of the calculations are given below:

Source	SS	D.F.	MS	F
Fertilizers	2940	3	.....	5.99
Within groups	....	.....	.....	
Total	6212	24		

(a) Fill in the blanks in the ANOVA table.

(b) Test at 5% level of significance, whether the fertilizers differ significantly.

7. There are three main brands of a certain powder. A sample of 120 packets sold is examined and found to be located among four groups A, B, C and D, and brands I, II and III as shown below:

Brand	Group			
	A	B		
I	0	4	8	15
II	5	8	13	6
III	18	19	11	13

Is there any significant difference in brand preferences?

8. A tea company appoints four salesmen A, B, C and D, and observes their sales in three seasons—summer, winter and monsoon. The figures (in lakhs) are given below:

Season	Salesmen				
	A	B	C	D	Total
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Total	90	93	81	96	360

(a) Do the salesmen significantly differ in performance?

(b) Is there significant difference between the seasons?

*****



REFERENCES

1. Babbie, E. 2001. *The practice of social research* (9th ed.). Belmont, CA: Wadsworth.
2. Bachman, R. & Schutt, R. 2007. *The practice of research in criminology and criminal justice* (3rd ed.). Thousand Oaks, CA: Sage.
3. Corley, K. & Gioia, D. 2011. Building theory about theory building: What constitutes a theoretical contribution? *Academy of Management Review*, *36*, 12-32.
4. Cozby, P. 2009. *Methods in behavioral research*. Boston: McGraw Hill Higher Education.
5. C.R. Kothari 2004 Research Methodology Methods and Techniques, 2nd Edition: New Delhi, New Age International Publisher
6. Dubin, R. (1978). *Theory building* (Rev. ed.). NY: Free Press
7. Guthrie Gerard, *Basic Research Methods: An Entry to Social Science Research*. New Delhi: Sage publication. India. 2010.
8. Gupta S.P. 2010. Statistical Methods. New Delhi: Sultan Chand & Sons.
9. Gupta S.P. 2010. Statistical Method. New Delhi: Sultan.
10. Goode, W. J and Hatt, P.K. 2006. Methods in Social Research. Delhi: Surajeet Publication.
11. Gelso, C. (2006). Applying theories to research: The interplay of theory and research in science. In F.T. Leong & J.T. Austin (Eds.), *The psychology research handbook*. Thousand Oaks, CA: Sage. Retreived from http://www.sage-ereference.com/hdbk_ psychrsch/Article_n32
12. Hay, J. & Lee, S. (2009). Theory building through qualitative research: Marshalling opportunities to advance cancer screening efforts. *Health Education & Behavior*, *36*, 145-149. doi: 10.1177/10908109340515
13. Heinen, J. (1985). A primer on psychological theory. *The Journal of Psychology*, *119*, 413-421.
14. Kerlinger, R. (1986). *Foundations of behavioral research*. NY: Holt, Rinehart, & Winston.
15. Jahoda, Marie, Morton Deutsch, and Stuart W. Cook. 1951. Research Methods in Social Relations. New York: Dryden.
16. Kothari C.R. *Research Methods and Techniques.* New Delhi: New Age International Publisher. 2004
17. Lincoln, Y., & Guba, E. (2000). Paradigmatic controversies, contradictions, and emerging confluences. In N. Denzin &Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed.) (pp. 163–188). London: Sage.

18. Lynham, S. (2002). The general method of theory-building research in applied disciplines. *Advances in Developing Human Resources*, *4*, 221-241.
19. Locke, E. (2007). The case for inductive theory building. *Journal of Management*, *33*, 867-890. doi: 10.1177/0149206307307636
20. L. Choudhury 2000 An introduction to Statistics, Volume One: Guwahati Kitapghar
21. Medhi, J. 2006. Statistical Methods : An Introductory text. New Age International (p) Ltd.
22. Morrow, R. (1994). *Critical theory and methodology*. Thousand Oaks, CA: Sage.
23. Quinney, R. (1970). *The social reality of crime*. Boston: Little, Brown.
24. Rynes, S. & Gephart, R. (2004). From the editors: Qualitative research and the academy of management journal. *Academy of Management Journal*, *47*, 454-462.
25. Stinchcombe, A. (1968). *Constructing social theories*. Chicago: University of Chicago Press.
26. Sutton, R. & Staw, B. (1995). What theory is not. *Administrative Science Quarterly*, *40*, 371–384.
27. Smith, J., Bekker, H., & Cheater, F. (2011). Theoretical versus pragmatic design in qualitative research. *Nurse Researcher*, *18*, 39-51.
28. Southern, S. & Devlin, J. (2010). Theory development: A bridge between practice and research. *The Family Journal: Counseling and Therapy for Couples and Familes*, *18*, 84-87.doi: 10.1177/1066480709358422
29. Thurstone, L. L, .and E. J. Chave. 1929. The Measurement of Attitudes Chicago: University of Chicago Press.
30. Torraco, R. (2002). Research methods for theory building in applied disciplines: A comparative analysis. *Advances in Developing Human Resources*, *4*, 355-376.
31. Young, P.V. *Scientific Social Surveys and Research*. New Delhi: Prentice Hall, 2009
32. Young, P. V. 1992. Scientific social surveys and research, New Delhi: Prentice Hall of India.
33. Zikmund, W., Babin, B., Carr, J., & Griffin, M. (2008). *Business research methods* (8th ed.). Mason, OH: Cengage.

Name of the Paper: **Course Code:**

Centre for Internal Quality Assurance (CIQA)
Krishna Kanta Handiqui State Open University
City Office, Housefed Complex, Guwahati- 06

Learner's Feedback on Course

Dear Learner,

Regarding the course as mentioned above, we would like to know your opinions and comments so as to improve the quality of self learning materials in future. Please respond to the following statements by ticking the number you feel most reflect your opinion. After completion of the additional comments, please detach the page and send/mail the same to us at the address given below.

The Director, Centre for Internal Quality Assurance, KKHandiqui State Open University
Housefed Complex, Dispur, Guwahati- 781006
(E-mail id: ciqa@kkhsou.in)

1. Approximately how many hours did you spend for studying the units in the course? | | | |
2. Please give your opinions (by √ mark) to the following items based on your reading of the block:

Sl. No.	Statements	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
I.	The SLMs of the course fulfil the learning objectives set out in the beginning of each unit					
II.	The units of the SLM could satisfy my academic needs and expectations					
III.	The Unit writers have excellent knowledge about the course contents					
IV.	Language and contents of the units were presented at a level which I could readily understand					
V.	Ample opportunity for participation in the activities provided in the units					
VI.	Used enough Illustrations (Diagrams, tables etc.) for conceptual clarity					
VII.	Quality of content is engaging, relevant, and up-to-date					
VIII.	The self check questions are very helpful					
IX.	The Possible/model questions and the answers to check my progress have benefited me a lot					

Additional Comments: *(Please feel free to provide your open comments)*

1. Which aspects of the SLM, according to you, worked well?

   ---

2. What sort of changes/improvements do you feel KKHSOU could implement to improve the overall quality of the SLM?

Thank you for taking the time to complete this form.

******