

स्वाध्याय

स्वमन्थन

स्वावलम्बन



PGBCH-106 Biostatistics

- Block - 1 : Research Methodology and Statistical Methods
Block - 2 : Probability Distribution Theory and Demography
Block - 3 : Tests of Significance



सेक्टर-एफ, शान्तिपुरम, फाफामऊ, प्रयागराज-211021

www.uprtou.ac.in

टोल फ्री नम्बर - 1800-120-111-333



कुलपति

उत्तर प्रदेश राजर्षि टण्डन मुक्त विश्वविद्यालय,
प्रयागराज

उत्तर प्रदेश सरकार का एकमात्र मुक्त विश्वविद्यालय

संदेश

जनसंख्या की दृष्टि से वृहत्तम्, उत्तर प्रदेश में उच्च शिक्षा से वंचित परन्तु शैक्षिक नवाचारों की सम्भावनाओं से परिपूर्ण, प्रदेश की जनसंख्या को गुणात्मक शिक्षा के अवसर सर्वसुलभ एवं सर्वव्यापी बनाने के संकल्प से तीर्थराज प्रयाग में उत्तर प्रदेश राजर्षि टण्डन मुक्त विश्वविद्यालय की **स्थापना 1999** में की गई थी। समाज का ऐसा वर्ग जो अपनी रोजमर्रा की अपरिहार्य व्यस्तताओं के चलते नियमित कक्षाओं में नहीं जा सकता, उसके घर तक अध्ययन सामग्री पहुँचाना एवं आवश्यकतानुसार परामर्श देना तथा उनकी अन्तर्निहित क्षमता को प्रस्फुटित करना विश्वविद्यालय की शीर्ष प्राथमिकता है। अपनी स्थापना काल से लेकर वर्तमान भूमण्डलीकरण, निजीकरण एवं उदारीकरण के इस उपयोगितावादी युग में सर्वांगीण व्यक्तित्व निर्माण के लक्ष्य को लेकर यह विश्वविद्यालय रोजगारपरक पाठ्यक्रमों के संचालन एवं क्रियान्वयन हेतु कृतसंकल्पित है।

सम्प्रति इस विश्वविद्यालय के बारह क्षेत्रीय केन्द्र क्रमशः प्रयागराज, वाराणसी, गोरखपुर, फैजाबाद, लखनऊ, बरेली, आगरा, मेरठ, झाँसी, कानपुर, नोएडा एवं आजमगढ़ में संचालित हैं। लगभग 1200 से अधिक संचालित अध्ययन केन्द्र के माध्यम से उच्च शिक्षा का प्रचार-प्रसार किया जा रहा है। विश्वविद्यालय के कार्यक्षेत्र की व्यापकता को देखते हुए, शिक्षार्थियों के हित में एक टोल फ्री नम्बर **1800-120-111-333** भी प्रारम्भ किया गया है।

विश्वविद्यालय में त्वरित गति से विकसित हो रही आधुनिक संचार प्रणाली का प्रयोग करते हुए '**वर्ष पर्यन्त प्रवेश**' प्रक्रिया के अन्तर्गत वर्ष में दो सत्रों जनवरी एवं जुलाई में ऑनलाइन प्रवेश लेने की व्यवस्था प्राविधानित है। शिक्षार्थी केन्द्रित व्यवस्था के अन्तर्गत प्रवेश, परामर्श, पाठ्यसामग्री, अधिन्यास, परीक्षा आदि विभिन्न शैक्षिक गतिविधियों को ऑनलाइन करके विश्वविद्यालय ने शिक्षा को शिक्षार्थी के द्वार तक पहुँचाने में सफलता अर्जित की है। शिक्षार्थियों की सुविधा को दृष्टिगत रखते हुए विश्वविद्यालय की वेबसाइट को फेसबुक, ट्विटर और यू-ट्यूब जैसे सोशल मीडिया उपकरणों से सम्बद्ध कर दिया गया है। छात्रों को ऑनलाइन पाठ्य सामग्री उपलब्ध कराने के लिए मोबाइल ऐप को विकसित करते हुए तथा MOOCS, SWAYAM आदि के माध्यम से यह विश्वविद्यालय वर्चुअल शिक्षा के स्वप्न को साकार करने की दिशा में शीघ्रता से आगे बढ़ रहा है।

शिक्षक और शिक्षार्थी के मध्य भौगोलिक दूरी की बाधा को न्यूनतम करने हेतु विश्वविद्यालय बहुआयामी संचार पद्धति का प्रयोग कर रहा है। शिक्षार्थियों तथा प्राध्यापकों की शोध के क्षेत्र में उन्नति और कैरियर संवर्धन हेतु विश्वविद्यालय में समसामयिक महत्वपूर्ण विषयों पर राष्ट्रीय संगोष्ठियों, कार्यशालाओं एवं व्याख्यान कार्यक्रमों का आयोजन निरन्तर किया जाता है। वर्तमान समय में विकास की वैकल्पिक विचारधारा के रूप में पंडित दीन दयाल उपाध्याय के विचारों की महत्ता एवं प्रासंगिकता को दृष्टिगत रखते हुए विश्वविद्यालय में '**दीन दयाल उपाध्याय शोध पीठ**' की स्थापना की गयी है। उल्लेखनीय है कि विश्वविद्यालय में अटल बिहारी बाजपेई सुशासन पीठ पहले से ही स्थापित है।

दूरस्थ व मुक्त शिक्षा प्रणाली में शिक्षार्थी सहायता सेवाओं की गम्भीरता को देखते हुए विश्वविद्यालय द्वारा विभिन्न अध्ययन केन्द्रों तथा क्षेत्रीय केन्द्रों पर संचालित गतिविधियों के निरीक्षण और अनुश्रवण हेतु प्रभावी एवं ठोस कदम उठाए जा रहे हैं। इस क्रम में हमारा प्रयास है कि हम शिक्षार्थियों से सीधे संवाद स्थापित कर उनकी जरूरतों को पहचानने और तदनुरूप अपनी शैक्षिक तथा प्रशासनिक व्यवस्थाओं को अधिक से अधिक अनुकूलतम बनाने की दिशा में कार्य कर सकें। देश की समसामयिक चुनौतियों एवं समस्याओं को देखते हुए वर्तमान सत्र से पूर्व में संचालित रोजगारपरक पाठ्यक्रमों को जारी रखते हुए नागरिकता संशोधन कानून एवं कोविड-19 पर जागरूकता पाठ्यक्रम विश्वविद्यालय में प्रारम्भ हो गया है।

शिक्षा को सामाजिक एवं राष्ट्रीय सरोकारों से जोड़ते हुए प्रति-वर्ष अन्तर्राष्ट्रीय योग दिवस, राजर्षि टण्डन जयन्ती स्मृति व्याख्यानमाला, पं. दीन दयाल उपाध्याय स्मृति व्याख्यानमाला, स्वैच्छिक रक्तदान शिविर, राष्ट्रीय तम्बाकू नियंत्रण जागरूकता कार्यक्रम, शिक्षक दिवस, हिन्दी दिवस और गाँधी जयन्ती पर विचार गोष्ठियों एवं व्याख्यान कार्यक्रमों के आयोजन की समृद्ध परम्परा है।

मैं विश्वविद्यालय की प्रगति में निरन्तर योगदान करने वाले शिक्षाविदों, क्षेत्रीय समन्वयकों तथा अध्ययन केन्द्रों के समन्वयकों, प्राचार्यों एवं समस्त विश्वविद्यालय परिवार के प्रति सम्मान एवं आभार ज्ञापित करती हूँ। विश्वविद्यालय को प्रगति के सोपानों पर अग्रसर करने में प्रत्यक्ष एवं परोक्ष रूप से सहयोग प्रदान करने वाले सभी सम्भ्रांत नागरिकों को मैं साधुवाद देती हूँ। मैं विश्वविद्यालय के केन्द्र बिन्दु तथा अध्ययन के प्रति निष्ठावान छात्र-छात्राओं के मंगलमय भविष्य की कामना करती हूँ।

प्रो० सीमा सिंह
कुलपति



॥ सरस्वती नः सुभगा मयस्कृत ॥

Uttar Pradesh Rajarshi Tandon
Open University

PGBCH-106

Biostatistics

Block-1	RESEARCH METHODOLOGY AND STATISTICAL METHODS	3-48
UNIT-1	Research Methods and Sampling Procedures	7
UNIT-2	Statistical Tools	13
Block-2	PROBABILITY-DISTRIBUTION THEORY AND DEMOGRAPHY	49-74
UNIT-3	Probability and Distribution Theory	53
UNIT-4	Demography	65
Block-3	TESTS OF SIGNIFICANCE	75-112
UNIT-5	Testing of Hypothesis	79
UNIT-6	Analysis of Variance	101



Uttar Pradesh Rajarshi Tandon
Open University

PGBCH-106

Biostatistics

BLOCK

1

RESEARCH METHODOLOGY AND STATISTICAL METHODS

UNIT-1

Research Methods and Sampling Procedures

UNIT-2

Statistical Tools

Course Design Committee

Dr. (Prof.) Ashutosh Gupta School of Science, UPRTOU Prayagraj	Director
Prof. Umesh Nath Tripathi Department of Chemistry Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur	Member
Prof. S.I. Rizvi Department of Biochemistry University of Allahabad, Prayagraj	Member
Prof. Dinesh Yadav Department of Biotechnology Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur	Member
Prof. Sharad Kumar Mishra Department of Biotechnology Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur	Member
Dr. Ravindra Pratap Singh Academic Consultant (Biochemistry) School of Sciences, UPRTOU Prayagraj	Member
Dr. Dharmveer Singh Academic Consultant (Biochemistry) School of Sciences, UPRTOU Prayagraj	Member/Secretary

Course Preparation Committee

Dr. Shruti Sr. Assistant Professor, School of Sciences U. P. Rajarshi Tandon Open University, Prayagraj	Writer
Prof. G. S. Pandey Department of Statistics, University of Allahabad, Prayagraj	Editor
Dr. Dharmveer Singh Academic Consultant (Biochemistry) School of Science, UPRTOU, Prayagraj	Course Coordinator

©UPRTOU, Prayagraj - 2021
ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the **Uttar Pradesh Rajarshi Tondon Open University, Prayagraj.**

Printed and Published by Dr. Arun Kumar Gupta Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2021.

Printed By : Chandrakala Universal Pvt. Ltd. 42/7 Jawahar Lal Neharu Road, Prayagraj 211002

BLOCK INTRODUCTION

The present SLM on **Bio Statistics** consists of three Blocks. *Block - 1 – Research Methodology and Statistical Methods* has two units; *Block - 2 – Probability - Distributions and Demography* has two units; and at the last *Block – 3 Tests of Significance and Analysis of Variance* has two units.

The **Block-1– Research Methodology and Statistical Methods** consists two units. The *first unit* of this block; named *Research Methods and Sampling Procedures*; describes the meaning and types of research, significance of research. It tells about the research problem and its selection with *Sampling Theory*; which discuss about the sampling, different types of sampling designs, simple random sampling, stratified sampling and cluster sampling with their applications.. The *second unit* of this block; named *Statistical Tools*; discusses about the measures of central tendency, measures of dispersion, measures of asymmetry, correlation and regression analysis, association of attributes and 3-sigma limits.

The **Block-2– Probability- Distributions Theory and Demography** is the second block having two units. The *first unit* of this block; named *Probability and Distribution Theory*; gives the Basic concepts of probability, definitions of probability, additive and multiplicative law of probability, conditional probability, Bayes' theorem, random variable and its types, probability mass function and probability density functions. In *Probability Distributions*; the concept of probability distribution, discrete and continuous probability distributions namely Binomial Distribution, Poisson Distribution, Geometric Distribution, Normal Distribution, Exponential Distribution have been also discussed in this unit along with their properties, applications and importance.. The *second unit* of this block; named *Demography*; gives knowledge about the vital statistics and demography, this also tells about the source of vital statistics and demographic data, rates, ratio, proportion, measures of fertility, measures of mortality, measures of morbidity and migration.

The **Block-3– Tests of Significance and Analysis of Variance** consists of two units. The *first unit* of this block; named *Testing of Hypothesis*; discuss about the hypothesis and its types, level of significance, critical region, p-value, types of errors, chi-square tests, t-tests and z-tests with their applications. The *second unit* of this block; named *Analysis of Variance*; discusses about the concept of analysis of variance and co-variance, basic principles of ANOVA and ANCOVA. (One Way, Two Way and Three Way Analysis).

Illustrations and examples on these topics have also been given.

At the end of every block/unit the summary, self assessment questions and further readings are given.

Unit-1 Research Methods and Sampling Procedures

Structure

- 1.1 Introduction
- 1.2 Objectives
- 1.3 Meaning and Types of Research and Research Design
- 1.4 Objectives and Characteristics of a Good Research
- 1.5 Research Problem
- 1.6 Sampling Procedures
- 1.7 Sampling Techniques
 - 1.7.1 Simple Random Sampling
 - 1.7.2 Stratified Sampling
 - 1.7.3 Cluster Sampling
- 1.8 Summary
- 1.9 Check Your Progress
- 1.10 Further Readings

1.1 INTRODUCTION

Now a day's research plays an important role in higher studies. Several research studies are published every year. But some are given more attention and some are less. A research which follows systematic and scientific procedures, gain more attention and found better results. This unit blows some light on all these.

1.2 OBJECTIVES

After studying this unit you will be able to

- Understand the meaning and types of research.
- Understand the significance of research, research problem and its selection.
- Understand the sampling procedures and sampling techniques.

1.3 MEANING AND TYPES OF RESEARCH AND RESEARCH DESIGN

Research is a method of seeking the new information, new thoughts, and new knowledge. Research is regularly used for solving problems and increasing the knowledge. This can be completed by theories and observations. Research is to be systematic, organized, objective and scientific.

The two basic research **types** are quantitative and qualitative research. But most of the time research can be separated into following different categories;

- Exploratory
- descriptive
- analytical
- causal
- Applied
- Fundamental
- Conceptual
- Empirical, etc

A **research design** is a overall strategic, planed framework for concluding the research scientifically through the collection, interpretation, analysis, and discussion of data. There are some major types of research design are as follows:

- General Structure and Writing Style
- Action Research Design
- Case Study Design
- Causal Design
- Cohort Design
- Cross-Sectional Design
- Descriptive Design
- Sampling Design
- Statistical Design
- Experimental Design

1.4 OBJECTIVE AND CHARACTERISTICS OF A GOOD RESEARCH

The **objective** of a research is to discover answer to question through the application of scientific and statistical procedures. The main aim of any research is to find out something new, which is hidden and also not been discovered as up till now.

A good research design should always accomplish the following main *characteristics*:

- It should be systematically, analytical, critical, methodical, and reflexive.
- It has a systematic statistically process of collecting and analyzing the data.
- It is methodological, empirical and replicable.
- It should be reliable.

1.5 RESEARCH PROBLEM

A *research problem* is a condition to be improved, a difficulty to be eliminated, or a troubling question that exists in scholarly literature, in theory, or in practical those points to the need for important concerned and focused study/research. In simple words we can say, a research problem refers to the difficulty which a researcher experiences in his theoretical or practical study. It has been seen that, the research problems are of three types: descriptive, relational, and causal/fundamental research. Thus a research problem helps researcher to invent the systematic chain of study. It also helps to avoid unnecessary steps during the research period. It is the beginning step in research study. It helps to understand the research procedure in a better scientific and systematic manner.

1.6 SAMPLING PROCEDURES

Before going to study of sampling theory, blow some lights on its terminology. It is known that *population* is a ‘well explained groups which is being studied’ and a *sample* is the “Small collection of the population” which has actually been observed. Before studying the theory of sampling, it is necessary to know about the *parameter* and *statistic*. The measures (like as mean, variance, correlation coefficient, etc.) of population are known *parameters* of the population. Same as these measures which drawn for sample is known as *statistic*. Therefore parameter and statistic are two most important elements of further study. It is known that a part of the population which is representative of the population is known as *sample* and the process of selection is known as *sampling* and also the relation between sample and population, with all statistical principles and theories are known as *sampling theory*. If population is very large, then it is single useful method which can be applicable. There are following types of sampling:

1.7 SAMPLING TECHNIQUES

Random and Non Random Sampling- A sampling procedure is said to be *random*, when it follow the scientific methods for selecting the samples. In this each unit of the population has an equal chance of selection in the sample.

A sampling procedure is said to be *non-random* when it do not chase any

scientific methods and has not an equal chance of selection. In other words the sample units are selected without use of randomization.

Note:

- Some sampling procedures of random Sampling are Simple Random Sampling, Stratified Sampling, Systematic Sampling, Cluster Sampling, and Multistage Sampling etc..
- Some sampling procedures of Non-random Sampling are Judgmental Sampling, Convenience Sampling, Quota Sampling, and Snowball Sampling etc..
- In the study of sampling theory, there are some possibilities to the occurrence of two types of errors, say *sampling errors* and *non sampling errors*.

In this unit we study only flowing random sampling procedures.

1.7.1 SIMPLE RANDOM SAMPLING (SRS)

A sampling in which, every unit of the population (here population must be finite and homogeneous) has equal independent chance of selection in sample. There are two types of simple random sampling, when

- The sample units are selected without replacement (no element can be selected more than once in the same sample), is known as *Simple random sampling without replacement (SRSWOR)*.
- The sample units are selected with replacement (an element may appear multiple times in the same sample), is known as *Simple random sampling with replacement (SRSWR)*. (Practically it is not used for further analysis).

There are some methods for selecting the SRS:

Lottery Method- This is most popular and the simplest method of selection of sample from population. In this method all objects of population are numbered on the slips (which are same in shape, size and colour). These slips are shuffled well and a blindfold selection is made.

Random Number Table Method- If the population size is very large then above method is impossible. Therefore random numbers table may be used. Some well known random number tables are, Tippet's random number table, Fisher and Yates table, Kendall and Smith's table, etc.

1.7.2 STRATIFIED SAMPLING

This method is useful when the nature of population is heterogeneous (i.e. not homogeneous). A sample which is drawn after stratification follows two steps mainly; at the first it is to divide the entire heterogeneous population into several non-overlapping homogeneous sub-populations/ groups, these homogeneous groups or classes known as *strata* and secondly from each *stratum*, the units are drawn by SRS, and lastly combining all units collect together. This technique

gives more representative sample than SRS in a large heterogeneous population. In other words, it gives a proportionate representative sample from each group is secured and it gives greater accuracy.

1.7.3 CLUSTER SAMPLING

In this sampling procedure, the whole population is divided into sub population (separate groups of units) say *clusters*, where each unit of the population belongs to one and only one cluster. A SRS is taken from each cluster. Whenever the elements within cluster are heterogeneous, it tends to provide best results. In this each cluster is the representative of the entire population. It is also called *Area Sampling*.

1.8 SUMMARY

This unit gives detailed knowledge about the meaning and types of research, significance of research. It tells about the research problem and sampling theory discuss about the sampling, different types of sampling designs, simple random sampling, stratified sampling and cluster sampling with their applications.

1.9 CHECK YOUR PROGRESS

1. Discuss about the meaning and types of research?
2. Discuss about the research problem?
3. What do you mean by sampling procedures and its techniques?
4. Write a short note on:
 - a. Simple Random Sampling
 - b. Stratified Sampling
 - c. Cluster Sampling

1.9 FURTHER READINGS

1. Goon, Gupta & Dasgupta: Fundamentals of Statistics Vol. II the World Press Pvt. Ltd., Kolkata.
2. Kothari, C. R., Research Methodology: Methods and Techniques; New Age International Publishers, New Delhi.
3. Kothari, C. R., Quantitative Techniques; Vikas Publishing House, New Delhi.
4. Gupta S. C., Kapoor, V.K., Applied Statistics, S. Chand, New Delhi.

UNIT-2 STATISTICAL METHODS

Structure

- 2.1 Introduction
- 2.2 Objectives
- 2.3 Measures of Central Tendency
 - 2.3.1 Arithmetic Mean
 - 2.3.2 Geometric Mean
 - 2.3.3 Harmonic Mean
 - 2.3.4 Median
 - 2.3.5 Mode
 - 2.3.6 Percentile, Deciles and Quartiles to Measurement of Location
- 2.4 Measures of Dispersion
 - 2.4.1 Range
 - 2.4.2 Mean Deviation
 - 2.4.3 Variance and Standard Deviation
 - 2.4.4 Coefficient of Variation (CV)
- 2.5 Measures of Asymmetry
 - 2.5.1 Skewness and its measures
 - 2.5.2 Kurtosis
- 2.6 Measures of Relationship
 - 2.6.1 Karl Pearson coefficient of correlation
 - 2.6.2 Spearman coefficient of correlation
- 2.7 Regression Analysis
 - 2.7.1 Regression line X on Y
 - 2.7.2 Regression line Y on X
- 2.8 Association of Attributes
- 2.9 3-sigma limits
- 2.10 Summary
- 2.11 Check Your Progress
- 2.12 Further Readings

2.1 INTRODUCTION

Statistical Methods is a wide-ranging term which includes almost all the methods involved in the collection, processing, condensing and analyzing of data. The data collected from the field for a number of items vary greatly in their qualitative as well as quantitative nature. In this unit have been highlighted different measures of central tendency, measures of dispersion, measures of asymmetry, measures of relationships are covered.

2.2 OBJECTIVES

After studying this unit you will be able to

- Understand the meaning of central tendency of data and compute common measures of central tendency.
- Understand the meaning of dispersion of data and compute common measures of dispersion.
- Understand the meaning of symmetry of data and compute common measures of asymmetry.
- Understand the meaning of relationships of data and compute common measures of relationship.

2.3 MEASURES OF CENTRAL TENDENCY AND ITS TYPES

When the observations have a trend to cluster round a central value, and this feature of observations is called the “**Central Tendency**” and related statistical measure (which gives the point round which the observation has a tendency to cluster) is known as ‘**measure of central tendency**’. The central value of the variable in any series of observations is useful in finding the location of the distribution and so it is also called an **average**.

Seeing as each one of the given measures of central tendency has its own individual characteristics and properties, a decision must always be made as to which would be the most appropriate and useful in view of the nature of the statistical data and purpose of the inquiry. The qualities preferred in a measure should be rigidly defined, easily computed, capable of a simple interpretation, not influenced by one or two extremely large or small values and likely to fluctuate relatively little from one random sample to another (of the same size and from the same population).

The measures of central tendency or averages are of different types, but the most common in use are of three types:

1. Mean
2. Median
3. Mode

The mean is further classified as:

- (i) Arithmetic mean
- (ii) Geometric mean
- (iii) Harmonic mean

2.3.1 ARITHMETIC MEAN (FOR UNGROUPED DATA)

The arithmetic mean of a series of n observations $x_1, x_2, x_3, \dots, x_n$ is obtained by summing up the values of all the observations and dividing the total by the number of observations. Thus,

$$X = \frac{\text{Sum of observations (or values)}}{\text{Number of observations}}$$

$$= \frac{x_1 + x_2 + x_3 \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Where $\sum$ (sigma) stands for summations and x_i is the i^{th} value of the observation (variable).

Example

The rainfall record in a month of 10 regions of a State is given below. Compute the average rainfall of the month for the State.

Region	1	2	3	4	5	6	7	8	9	10
Rainfall	17.6	10.1	11.4	18.5	10.5	14.3	8.9	13.4	10.6	12.5

(in mm)

Solution :

$$\text{Mean } (X) = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{x_1 + x_2 + x_3 \dots + x_{10}}{10}$$

$$= \frac{17.6 + 10.1 + 11.4 + \dots + 10.6 + 12.5}{10} = \frac{127.8}{10} = 12.78 \text{ mm}$$

The computation of Arithmetic mean using short cut method is discussed below:

Short-Cut Method- This method is applied to avoid lengthy calculations.

Let $x_1 + x_2 + x_3 \dots + x_n$ be n individual reading on the variable and let A be the working mean. Let $d_1, d_2, d_3, \dots, d_n$ denote the differences between the working mean and individual values $x_1 + x_2 + x_3 \dots + x_n$ respectively. The mean $\bar{x}$ of X , in terms of the mean $\bar{d}$ of differences, is calculated as:

$$\begin{aligned}
 d &= \frac{d_1 + d_2 + d_3 \dots \dots + d_n}{n} \\
 &= \frac{(x_1 - A) + (x_2 - A) + (x_3 - A) + \dots \dots + (x_n - A)}{n} \\
 &= \frac{\sum_{i=1}^n x_i - nA}{n} = \frac{\sum_{i=1}^n X_i}{n} - A = \bar{X} - A
 \end{aligned}$$

Or,
$$\bar{X} = A + \bar{d} = A + \frac{1}{n} \sum_{i=1}^n d_i$$

An illustration is given below.

Example: Calculation the mean for the following scores: 60, 65, 74, 85, 95.

Solution:

Table

x_i (Scores)	$x_i - 74$
60	-14
65	-9
74	0
85	+11
95	+21
	+9

Then, mean score is

$$\bar{x} = 74 + (+9/5) = 74 + 1.8 = 75.8$$

Grouped data (Discrete Frequency Distribution)- In a discrete series, let the individual readings $x_1, x_2, x_3 \dots \dots x_n$ of the variable X occur (have frequencies) $f_1, f_2, f_3 \dots \dots, f_n$ times respectively. Then the mean of X is obtained by summing the product of individual readings with corresponding frequencies and dividing the total by the sum of frequencies, i.e.,

$$\bar{X} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 \dots \dots + x_n f_n}{f_1 + f_2 + f_3 \dots \dots + f_n} = \frac{\sum_{i=1}^n X_i f_i}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n X_i f_i}{N}$$

Where $N = \sum_{i=1}^n f_i$ is the total number of observations.

Short cut method- The arithmetic mean ($\bar{x}$) is then calculated with the help of following formula:

$$Mean = \bar{x} = A + \frac{1}{n} \sum_{i=1}^n f_i d_{xi}$$

Example- Below are given the number of children born per family in 735 families in a locality. Calculate the average number of children born per family in the locality.

Table

Number of children born per family	Number of families
0	96
1	108
2	154
3	126
4	95
5	62
6	45
7	20
8	11
9	6
10	5
11	5
12	1
13	1

Solution- Computation of the average number of children born per family:

Table

Number of children born per family (X)	Number of Families (f)	$X.f$ (xf)
0	96	$0 \times 96 = 0$
1	108	$1 \times 108 = 108$
2	154	$2 \times 154 = 308$
3	126	$3 \times 126 = 378$
4	95	$4 \times 95 = 380$
5	62	$5 \times 62 = 310$
6	45	$6 \times 45 = 270$
7	20	$7 \times 20 = 140$
8	11	$8 \times 11 = 88$
9	6	$9 \times 6 = 54$
10	5	$10 \times 5 = 50$
11	5	$11 \times 5 = 55$
12	1	$12 \times 1 = 12$
13	1	$13 \times 1 = 13$
Total	735	2166

Here $N = \sum f = 735$; $\sum Xf = 2166$

Average number of children born per family is given by

$$\text{mean}(x) = \frac{\sum xf}{\sum f} = \frac{2166}{735} = 2.9 \text{ children}$$

Grouped Data (Continuous Frequency Distribution)- When the measurement are given in the grouped form, the mean is computed by multiplying the various mid values m_i with their respective frequencies f_i where $i = 1, 2, \dots, n$ and dividing the product total by the sum of frequencies or total number of observations. If $m_1, m_2, \dots, m_n$ are the mid values or class intervals corresponding to frequencies $f_1, f_2, \dots, f_n$ then the mean is

$$\bar{X} = \frac{m_1 + m_2 f_2 + \dots + m_n f_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^n m_i f_i}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n m_i f_i}{N}$$

Where, n stand for the number of groups and N denotes the total number of observations.

Short Cut Method- If the class groups formed by individual reading are large, calculate arithmetic mean $\bar{x}$ by the following formula:

$$\text{mean} = (\bar{x}) = A + \left(\frac{\sum_{i=1}^n f_i d_i}{N} \right)$$

Step- Deviation Method- if $d'_i = \frac{(m_i - A)}{h}$, the formula for calculating arithmetic mean by this method is

$$\text{mean} = (\bar{x}) = A + \left(\frac{\sum_{i=1}^n f_i d_i}{N} \times h \right)$$

Where h is the size of the class intervals.

Example- Calculate mean from given table

Height (in inches)	0-10	10-20	20-30	30-40	40-50	50-60
No. of plants	22	10	8	15	5	6

Solution-

(a) By direct method

Height (in inches)	Mid Values (m)	No. of plants	$m.f$ (mf)
0-10	$\frac{0 + 10}{2} = 5$	22	$5 \times 22 = 110$
10-20	$\frac{10 + 20}{2} = 15$	10	$15 \times 10 = 150$
20-30	$\frac{20 + 30}{2} = 25$	8	$25 \times 8 = 200$
30-40	$\frac{30 + 40}{2} = 35$	15	$35 \times 15 = 525$
40-50	$\frac{40 + 50}{2} = 45$	5	$45 \times 5 = 225$
50-60	$\frac{50 + 60}{2} = 55$	6	$55 \times 6 = 330$
Total		66	1540

$$\text{Here } N = \sum mf = 1540$$

$$\text{mean } (\bar{x}) = \frac{\sum mf}{\sum f} = \frac{1540}{66} = 23.33 \text{ inches}$$

(b) Short cut method

Height (in inches)	Mid Values (<i>m</i>)	$d=m-A$	No. of plants	$d.f$ (df)
0-10	5	-30	22	$-30 \times 22 = -660$
10-20	15	-20	10	$-20 \times 10 = -220$
20-30	25	-10	8	$-10 \times 8 = -80$
30-40	35(A)	0	15	$0 \times 15 = 0$
40-50	45	10	5	$10 \times 5 = 50$
50-60	55	20	6	$20 \times 6 = 120$
Total			66	-770

$$\text{Here } N = \sum f = 66, \quad \sum df = -770$$

$$\text{mean } (\bar{X}) = A + \frac{\sum fd}{N}$$

$$= 35 + \frac{-770}{66} = 23.33 \text{ inches}$$

(c) By step- deviation Method

From the given data $h=10$

Height (in inches)	Mid Values (<i>m</i>)	$d'_i = \frac{(m_i - A)}{h}$	No. of plants	$d.f$ (df)
0-10	5	$\frac{5 - 35}{10} = -3$	22	$-3 \times 22 = -66$
10-20	15	$\frac{15 - 35}{10} = -2$	10	$-2 \times 10 = -20$

20-30	25	$\frac{25 - 35}{10} = -1$	8	$-1 \times 8 = -8$
30-40	35(A)	$\frac{35 - 35}{10} = 0$	15	$0 \times 15 = 0$
40-50	45	$\frac{45 - 35}{10} = +1$	5	$1 \times 5 = 5$
50-60	55	$\frac{55 - 35}{10} = +2$	6	$2 \times 6 = 12$
Total	-	66	66	-77

$$\text{Here } N = \sum f = 66, \quad \sum fd' = -77$$

$$\text{mean } (\bar{X}) = A + \frac{\sum fd'}{N} \times h = 35 + \frac{-77}{66} \times 10 = 23.33 \text{ inches}$$

Properties of Arithmetic Mean-

- 1. First Property of Mean-** The sum of the deviations about the arithmetic mean equals zero.

Mathematically

$$\sum [f_i(x_i - \bar{x})] = 0$$

This property says that if the mean is subtracted from each score, the sum of the differences will equal zero. The property results from the fact that the mean is the balance point of the distribution. The mean can be thought of as the fulcrum of a seesaw. When the score are distribution along the seesaw according to their values, the mean of the distribution occupies the position where the scores are in balance. This is known as first property of mean.

- 2. Second Property of Mean-** The sum to the squared deviations of all the scores about their arithmetic mean is minimum.

That is,

$$\sum [f_i(x_i - \bar{x})]^2 = \text{minimum}$$

This is an important characteristic and is used in many areas of statistics, particularly in regression analysis.

- 3. Combined (Additive) Property of Mean-** If $\bar{X}_1$ and $\bar{X}_2$ be the means of two series of sizes n_1 and n_2 respectively, then the mean of the combined series can be computed as:

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

In the same way, if $\bar{X}_1, \bar{X}_2, \bar{X}_3 \dots \dots \bar{X}_k$ be the means of k series of sizes $n_1, n_2, n_3 \dots \dots n_k$ respectively then the mean $\bar{X}$ of combined series is

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \dots \dots n_k\bar{X}_k}{n_1 + n_2 + \dots \dots n_k}$$

Example- The average ages of 250 males and 210 females in a village are 41.6 and 38.5 years respectively. Find the average age combining both males and females together.

Solution- Here (Combined average) is $N_1 = 250, \bar{X}_1 = 41.6$ years and $N_2 = 210, \bar{X}_2 = 38.5$ years. Therefore,

$$\begin{aligned}\bar{X} &= \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \frac{250 \times 41.6 + 210 \times 38.5}{250 + 210} = \frac{10400 + 8085}{460} \\ &= 40.18 \text{ years}\end{aligned}$$

2.3.2 GEOMETRIC MEAN

In case of finding the average or rates and ratios, geometric mean is more useful measure than others, e.g. in finding the population increase simple and compound interests etc.

Case 1- In case of ungrouped data it is obtained by multiplying together all the values of the variable and extracting the relevant root of the product. i.e. if $(x_1, x_2, x_3 \dots x_n)^{\frac{1}{n}}$ and n values of a variable under study, then the geometric mean (G.M.) is computed as:

$$G.M. = (x_1, x_2, x_3 \dots x_n)^{\frac{1}{n}} ; (x_i > 0)$$

To facilitate the computation one can make the use of logarithms as:

$$\begin{aligned}\log(G.M.) &= \frac{1}{n} \log(x_1, x_2, x_3 \dots x_n) \\ &= \frac{1}{n} [\log x_1 + \log x_2 + \log x_3 \dots + \log x_n] = \frac{1}{n} \left[\sum_{i=1}^n \log x_i \right]\end{aligned}$$

$$\text{So } G.M. = \text{Antilog } \frac{1}{n} \left[\sum_{i=1}^n \log x_i \right]$$

Thus the logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of individuals' measurements.

Case 2- In case of grouped data if $f_1, f_2, f_3 \dots f_n$ be the frequencies corresponding to the individual values $(x_1, x_2, x_3 \dots x_n)$ then G.M. is computed as:

$$G.M. = (x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \dots x_n^{f_n})^{\frac{1}{\sum_{i=1}^n f_i}} ; (x_i > 0)$$

$$\text{or } \log G.M. = \frac{1}{\sum_{i=1}^n f_i} [f_1 \log x_1 + \dots + f_n \log x_n] = \frac{1}{\sum_{i=1}^n f_i} \left[\sum_{i=1}^n f_i \log x_i \right]$$

$$\text{So } G.M. = \text{Antilog} \left\{ \frac{1}{\sum_{i=1}^n f_i} \left[\sum_{i=1}^n f_i \log x_i \right] \right\}$$

Here log G is the weighted mean of log x_i 's with weights $f_1, f_2, \dots, f_n$.

Additive Property of Geometric Mean- If G_1 and G_2 are the geometric means of two different series with respective sizes n_1 and n_2 , the combined Geometric mean G is

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

Example- The monthly average temperature of a station for five months is given as 16.2, 23.4, 20.6, 33.4 and 16.4 degree centigrade. Find the mean temperature of the station.

Solution- $G.M. = (16.2 \times 23.4 \times 20.6 \times 33.4 \times 16.4)^{1/5}$

or

$$\begin{aligned} \log G.M. &= \frac{1}{5} [\log 16.2 + \dots + \log 16.4] \\ &= \frac{1}{5} [1.2095 + 1.3692 + 1.3139 + 1.5238 + 1.2148] \\ &= \frac{1}{5} \times 6.6312 = 1.3262 \end{aligned}$$

so $G.M. = \text{Antilog} [1.3262] = 21.1934$ degree centigrade.

Example- From the following data, calculate the G.M.

Height (in inches)	0-10	10-20	20-30	30-40	40-50
No. of plants	14	23	27	21	15

Calculation-

Height (in inches)	Mid Values (x)	$\log$ (x)	No. of observations (f)	$f \log x$
0-10	5	0.69897	14	9.78558
10-20	15	1.1769	23	27.05007

20-30	25	1.39794	27	37.74438
30-40	35	1.54407	21	32.42547
40-50	45	1.65321	15	24.79815
Total	-		100	131.80365

Here $\sum f = N = 100$, $\sum f \log x = 131.80365$

$$G.M. = \text{Antilog} = \left\{ \frac{1}{\sum f} [\sum f \log x] \right\}$$

$$\text{Antilog} = \left[\frac{131.80365}{100} \right] = \text{Antilog} [1.318036] = 20.7987$$

2.3.3 HARMONIC MEAN

In problems such as work time and rate where the amount of work is held constant an average rate is required, the harmonic mean (HM) is utilized. It is defined as the reciprocal of the arithmetic mean of the reciprocals of the given individual readings i.e. H.M. of $x_1, x_2, \dots, x_n$ is defined as:

$$H.M. = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \quad (x_i > 0)$$

Where n is the number of observations.

Example- The H.M. of 2, 4, 6 is

$$H.M. = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{6}} = \frac{36}{11} = 3.37$$

In case of frequency distribution (grouped data), if $f_1, f_2, \dots, f_n$ be the frequencies corresponding to $x_1, x_2, \dots, x_n$ then H.M. is computed as:

$$H.M. = \frac{f_1 + f_2 + f_3 + \dots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \dots + \frac{f_n}{x_n}} = \frac{\sum_{i=1}^n f_i}{\sum_{i=1}^n \frac{f_i}{x_i}} \quad (x_i > 0)$$

Example Calculation-

Class group (X)	Mid Values (m)	No. of observations (f)	f / m
0-10	5	14	2.80
10-20	15	23	1.53

20-30	25	27	1.08
30-40	35	21	0.60
40-50	45	15	0.33
Total		100	6.34

Here $\sum f = N = 100$, $\sum f / m = 6.34$

$$H.M. = \frac{\sum f}{\sum \left(\frac{f}{m}\right)} = \frac{100}{6.34} = 15.77$$

2.3.4 MEDIAN

Median is another important and useful measure of central tendency. It has connotation of the middle most or most central value of a set of measurements. It is usually defined as the value which divides a distribution in such a manner that the number of items below it is equal to the number of items above it. The median is thus a positional average. Median is that variate value of the data or frequency distribution which divides it in two equal parts.

Calculation of Median (Ungrouped data)- Case 1 (n is odd): In case of ungrouped data when the number of observations are odd, then median is the middle value after the measurements have been arranged in ascending or descending order of magnitude, i.e. if there are n number of measurements and measurements are arranged in ascending or descending order of magnitude, the median of the measurements is $\left(\frac{n+1}{2}\right)^{th}$ measurement where n is an odd number.

Case 2 (n is even)- If the numbers of observations are even, median is defined as the mean of the two middle observations, when observations are arranged in ascending or descending order to magnitude i.e.

$$median = \frac{\left(\frac{n}{2}\right)^{th} value + \left(\frac{n+1}{2}\right)^{th} value}{2}$$

Example- Calculate median for the following data:

- (a) 68, 62, 75, 82, 68, 71, 68, 71, 62, 68, 74, 59, 74, 68, 60, 71, 59, 73, 73, 58.
- (b) 200, 150, 260, 285, 380, 305, 4989, 307, 1280, 233, 403

Solution-

- (a) To compute the median first we arrange the values in ascending order of magnitude as:

58, 59, 59, 60, 62, 62, 68, 68, 68, 68, 68, 71, 71, 71, 73, 73, 74, 74, 75, 82.

The number of observation n is even in this case, i.e., n=20

So

$$\begin{aligned} \text{median} &= \frac{\left(\frac{n}{2}\right)^{\text{th}} \text{ value} + \left(\frac{n+1}{2}\right)^{\text{th}} \text{ value}}{2} = \frac{10\text{th value} + 11\text{th value}}{2} \\ &= \frac{68 + 68}{2} = 68 \end{aligned}$$

(b) Let us first arrange the values in ascending order of magnitude as:

150, 200, 233, 260, 285, 305, 307, 380, 403, 1280, 4989

The number of observations n in this case is odd, i.e., n=11 so the median is the $\left(\frac{n+1}{2}\right)^{\text{th}}$ value i.e., $\left(\frac{11+1}{2}\right)^{\text{th}}$ or 6th value of the observation and thus underlined value. i.e., 305 is the median.

Calculation of Median (Grouped Data)- In case of discrete frequency distribution median can be obtained with the help of cumulative frequencies as follows:

- (i) First find N/2 where $N = \sum f$
- (ii) Find the cumulative frequency just greater than N/2.
- (iii) Corresponding value of X (i.e., of variable) is median.

Example Calculation-

Height (in metre) (X)	Number of units (f)	Cumulative frequency (f)
200	142	142
600	265	407
1000	560	967
1400	271	1238
1800	89	1327
2200	16	1343
Total	1343	

Here $f=N= 1343; \frac{N}{2} = 671.5$

The cumulative frequency just greater than 671.5 is 967 and corresponding to this cumulative frequency, the value of X is 1000 and thus the median height is 1000 meters.

Median (Continuous Grouped Data)- Median for such distribution is computed by the following formula

$$\text{Median (Md)} = l_m + \frac{\left(\frac{n}{2} - f_e\right)}{f_m} \cdot h$$

Where l_m is the lower limit f_m is the frequency of the median class, f_e is the cumulative frequency of the class, preceding the median class and h is the class width of the median class and $N = \sum f_i$

Example- Calculate median for the following grouped data.

Interval	35-45	45-55	55-65	65-75	75-85
Frequency	2	3	5	1	1
Cum. Freq.	2	5	10	11	12

The cum. Freq. is computed from the given freq. dist.

The median position = $(n+1)/2 = (12+1)/2 = 6.5$

Median lies between observation 55 and 65. Both of these observations fall in category 3, i.e., in class (35-65) with cumulative frequency of 10. Therefore,

$$\text{Median (Md)} = l_m + \frac{\left(\frac{n}{2} - f_e\right)}{f_m} \cdot h$$

Where $l_m=55$, $f_m=5$, $h=10$, $f_e=5$, $N=12$

$$\text{Median (Md)} = 55 + \frac{[(12/2) - 5]}{5} \times 10 = 55 + 2 = 57$$

2.3.5 MODE

Mode or modal value of a distribution is that the value which occurs most frequency. In case of frequency distribution the mode is that value which has maximum frequency. If two or more observations occur the same number of times then there is more than one mode and the distribution is called multi-model as against uni-model.

Calculation of Mode (Ungrouped Data)- Mode is defined as that variate value of the data or the frequency distribution which occurs most frequently.

Example- Find the modal temperature value from the values given

Solution-

(i) Putting data in array as:

58, 59, 59, 60, 62, 62, 68, 68, 68, 68, 68, 71, 71, 71, 73, 73, 74, 74, 75, 82.

Here mode = 68⁰ F.

(ii) Discrete series (converted to frequency distribution form)

Variable (X)	58	59	60	62	68	71	73	74	75	82
Frequency (f)	1	2	1	2	5	3	2	2	1	1

Here the value 68 occurs the maximum number of times, hence it is mode.

Discrete Series (Grouped Data)- In case of discrete frequency distribution, mode can be located by inspection of the distribution alone. The size having the maximum frequency will be reckoned as mode.

Continuous Series (Grouped)- Determine the value of mode by applying the following formula:

$$Mode = L + \left(\frac{f - f_P}{2f - f_P - f_S} \right) h$$

Where, L is the lower limit of the modal class; f is the frequency of the modal class; f_P is the frequency of the class preceding the modal class; f_S is the frequency of the class succeeding the modal class and h is the class width of the modal class.

Example- Compute the modal agricultural holding of the village from the data given in above example

$$Mode = L + \left(\frac{f - f_P}{2f - f_P - f_S} \right) \times h$$

if $L = 14.5$, $f = 150$, $f_P = 35$, $f_S = 70$ and $h = 5$.

$$\begin{aligned} Mode &= 14.5 + \left(\frac{150 - 35}{150 \times 2 - 35 - 70} \right) \times 5 \\ &= 14.5 + \frac{115}{195} \times 5 = 17.45 \text{ acres.} \end{aligned}$$

Relationship between mean, Median and Mode- For a symmetrical distribution mean, median and mode coincide and if the distribution is moderately asymmetrical, the mean, median and mode are approximately related by the formula:

$$\text{Mode} \cong 3 \text{ Median} - 2 \text{ Mean}$$

2.4 MEASURES OF DISPERSION

The measures of dispersion summarize the variability of the distribution. It is useful to know how similar or dissimilar scores are from the average score and from one another. We might like to know if scores cluster, they are more homogeneous. If scores are spread out widely then they are more heterogeneous. The measures of central tendency as discussed above only locate the center of a distribution but tell nothing about the degree of variability. In order to study the dispersion or variability in a distribution, we need alternative measures called the measures of dispersion. In central tendency is thought of as the point that best represents a central score in a distribution, the dispersion presents the other side of the coin.

Types of Measures of Dispersion- Dispersion is defined as the degree to which scores deviate from the central tendency (usually the mean) of the distribution. The statistical techniques that quantify this dispersion in a distribution are called measures of dispersion. Most commonly used measures of dispersion are range, average deviations, variance and standard deviation.

2.4.1 RANGE

Range is defined as the difference between the highest and the lowest scores in a distribution. Symbolically,

$$R = X_{\max} - X_{\min}$$

Where R is the range, $X_{\max}$ is the highest score, $X_{\min}$ is the lowest score.

Example- Find range for value: 87, 92, 47, 58, 87, 62, 73, 73, 61.

Solution- It is always a good idea to first rank the observation in ascending or descending order. In an ascending order the scores are: 47, 58, 61, 62, 73, 73, 87, 87, 92. A visual examination shows that $X_{\max} = 92$; $X_{\min} = 47$.

Therefore $R = 92 - 47 = 45$

2.4.2 MEAN DEVIATION

The formula for average deviation can be written as:

$$MD = \left(\sum |x_i - \bar{x}| \right) / n$$

Where $\bar{x}$ is the arithmetic mean; $x_i - \bar{x}$ is deviation of from $\bar{x}_i$ and $|x_i - \bar{x}|$ is the absolute value of the deviation which is always a positive number.

Example- The numbers of terms that five randomly selected Members of Parliament have served are: 3, 10, 12, 7, 8. Find the average deviation of these scores.

Solution- Make the following table containing the calculation.

Case number	Terms	$x_i - \bar{x}$	$ x_i - \bar{x} $
1	3	3-8=-5	5
2	10	10-8=2	2
3	12	12-8=4	4
4	7	7-8=-1	1
5	8	8-8=0	0
Total	40	0	12

Mean $\bar{x} = 40/5=8$

MD= 12/5=2.4 terms.

2.4.3 VARIANCE AND STANDARD DEVIATION

Variance may be defined as the mean squared deviation of scores around the mean. In the form of a formula, variance is given by:

$$S_x^2 = \sum \frac{(x_i - \bar{x})^2}{n} \quad (\text{for sample data})$$

$$\sigma_x^2 = \sum \frac{(x_i - \mu)^2}{N} \quad (\text{for population data})$$

Where S_x^2 = sample variance of variable X; σ_x^2 is population variance x_i is the value of X variable for i^{th} case; $\bar{x}$ is sample mean; μ is population mean; n is the sample size; and N is population size.

Standard Deviation- The square root of the variance is called the standard deviation (SD), represented by s or σ . That is Standard deviation = Square root of variance or $s = \sqrt{S^2}$

$$s = \sqrt{S^2} = \sqrt{\left[\sum \frac{(x_i - \bar{x})^2}{n - 1} \right]} = \sqrt{\sum f_i \frac{(x_i - \bar{x})^2}{n - 1}}$$

Similarly, population standard deviation $\sigma = \sqrt{\sigma^2}$

Effect of Change of Origin and Scale- This S.D. is

$$S_{xi} = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n - 1}}$$

Where $\bar{x} = \frac{1}{n} \sum x_i$

If $\delta_i = \frac{x_i - a}{h} \quad \forall i = 1, 2, \dots, n,$

$$S_{xi} = hS_{\delta i}$$

i.e., standard deviation is independent of change of origin but not independent of change of scale.

Example- Calculate the mean and S.D. for the following given table of marks distribution of 50 students

Marks	Students (f)	Mid value (x)	$x_i - 25$	$\delta_i = \frac{x_i - 25}{10}$	$f_i \delta_i$	$f_i \delta_i^2$
0-10	2	5	5-25=-20	-2	-4	8
10-20	10	15	15-25= -10	-1	-10	10
20-30	15	25	25-25=0	0	0	0
30-40	14	35	35-25=10	1	14	14
40-50	9	45	45-25=20	2	18	36
	50				18	68

$$\bar{x} = a + h \frac{\sum f_i \delta_i}{N} \quad N = \sum f_i$$

$$= 25 + \frac{10 \times 18}{50} = -25 + 3.6 = 28.6$$

$$S_{xi}^2 = h^2 S^2$$

$$S_{\delta i}^2 = \left(\frac{1}{N} \sum f_i \delta_i^2 \right) - \bar{x}_{\delta i}^2 = \left(\frac{1}{N} \sum f_i \delta_i^2 \right) - \left(\frac{1}{N} \sum f_i \delta_i \right)^2$$

$$= \frac{1}{50} \times 68 - \left(\frac{1}{50} \times 18 \right)^2 = 1.36 - (0.36)^2 = 1.36 - 0.1296 = 1.4896$$

$$S_{xi}^2 = 10 \times 1.49 = 14.896$$

Combined Variance Property- If n_1 and n_2 be the sizes of two series with respective means $\bar{X}_1$, $\bar{X}_2$ and respective variances then the standard deviations S_1^2 , S_2^2 of the combined series is denoted as S and defined as

$$S^2 = \frac{1}{n_1 + n_2} [n_1(S_1^2 + d_1^2) + n_2(S_2^2 + d_2^2)]$$

Where

$$d_1 = \bar{X}_1 - \bar{x}, \quad \bar{x}_1 = \frac{1}{n_1} \sum x_{i1}$$

$$d_2 = \bar{X}_2 - \bar{x}, \quad \bar{x}_2 = \frac{1}{n_2} \sum x_{2j}$$

$$\text{and } \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \quad (\text{combined mean})$$

another formula is

$$S^2 = \frac{1}{n_1 + n_2} \left[n_1 S_1^2 + n_2 S_2^2 + \left\{ \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^2 \right\} \right]$$

Where

$$S_1^2 = \frac{1}{n_1} \sum (x_{1i} - \bar{x}_1)^2, \quad S_2^2 = \frac{1}{n_2} \sum (x_{2j} - \bar{x}_2)^2$$

2.4.4 COEFFICIENT OF VARIATION (CV)

It is sometimes desirable to compare several groups with respect to their relative homogeneity in instances where the groups have very different means. Therefore it might be somewhat misleading to compare the absolute magnitudes of the standard deviations. One might expect that with a very large mean one would find a fairly large standard deviation. One might therefore be primarily interested in the size of the standard deviation relative to that of the mean. This suggests that we can obtain a measure of the relative variability by dividing the standard deviation by the mean. The result has been termed the coefficient of variation, denoted by CV. Thus

$$CV = \sigma / \bar{X}$$

Where σ is the SD and $\bar{X}$ is the mean.

The series which having lesser CV is more consistent than other

2.5 MEASURES OF ASYMMETRY

After studying the skewness and kurtosis, have an idea about the shape of the frequency curve of the distribution. A distribution is said to be symmetrical if the frequencies are on either side of the central value. It implies that both the right and left tails of the curve are exactly equal in shape and length. If a distribution is not symmetrical then it is called asymmetric or skewed in the direction of the extreme values, i.e., on the right – or on the left. Since extreme values give longer tail in its direction therefore, the distribution having longer right tail is called right skewed or positively skewed distribution. The left implies longer left tail. Thus a measure of skewness indicates the extent as well as direction of skewness of the distribution. The measure of kurtosis gives an idea whether the centre of the distribution is assuming flatness or peakedness similar to the hump of the normal probability curve or not. The measures of skewness are very useful in biological, chemical and physical laboratory works. They are also used in economic social statistics and medical statistics to study the behavior of the data.

2.5.1 SKEWNESS AND ITS MEASURES

Definition- By skewness of a frequency distribution we mean the degree of its departure from symmetry.

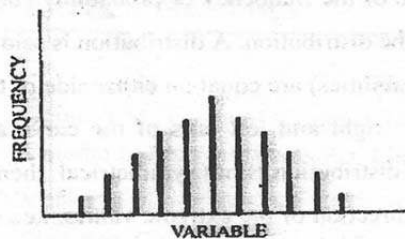


Fig. 2.1a A symmetrical distribution (discrete variable).

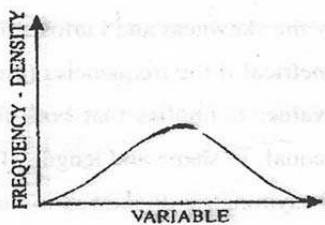


Fig. 2.1b A symmetrical distribution (continuous variable).

Figure 2.1a and 2.1b show two symmetrical distributions.

A distribution which is not symmetrical is called asymmetrical or skew. This skewness is said to be positive if the longer tail of the distribution is towards the higher values of the variable (Fig. 2.2a), negative if the longer tails is towards the lower values of the variable (Fig. 2.2b)

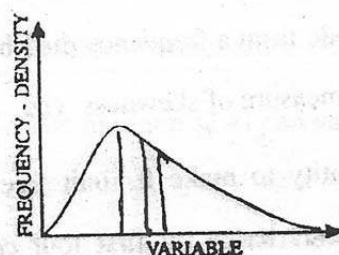


Fig. 2.2a A positively skew distribution.

Mode < Median < Mean

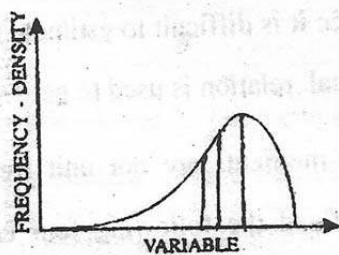


Fig. 2.2b A negatively skew distribution.

Mode < Median < Mean

For symmetrically shaped distribution: Mean = Median = Mode

For positively- skewed distribution: Mean > Median > Mode

For negatively skewed distribution: Mean < Median < Mode

Measures of Skewness: Pearsons Coefficient- An alternative measure of Skewness is obtained from the relative positions of the mean and the mode in a distribution.

In a symmetrical distribution, the mean, median and mode (assuming the distribution to be uni-modal) coincide. If the distribution is: skewed positively, then

$$\text{mean} > \text{median} > \text{mode.}$$

and if it is negatively skewed, then

$$\text{mean} < \text{median} < \text{mode.}$$

Hence the difference (mean mode) divided by the s.d., is taken as a measure of skewness.

$$Sk = \frac{\bar{x} - M_0}{s}$$

This is known as Pearson's first measure of skewness, provided $s > 0$.
another measure of skewness is.

$$Sk = \frac{3(\bar{x} - M_i)}{s}$$

Which is known as Pearson's second measure of skewness. If mean = median = mode then $S_k = 0$. The limits can vary between -3 to 3.

Measures of Skewness: Bowley's Coefficient- For a symmetrical distribution the lower and upper quartiles are equidistant from the median; for a positively skew distribution the lower quartile is nearer the median than the upper quartile is, while for a negative skew distribution the upper quartile is nearer.

Thus the new measure based on quartiles is

$$Sk = \frac{(Q_3 - M_i) - (M_i - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Md}{Q_3 - Q_1}$$

This is known as Bowley's measure of skewness. it has the limit -1 to 1.

2.5.2 KURTOSIS

Another method of describing a frequency distribution is to specify its degree of peakedness or kurtosis. Two distributions may have the same mean and the same standard deviation and may be equally skew, but one of them may be more peaked than the other.

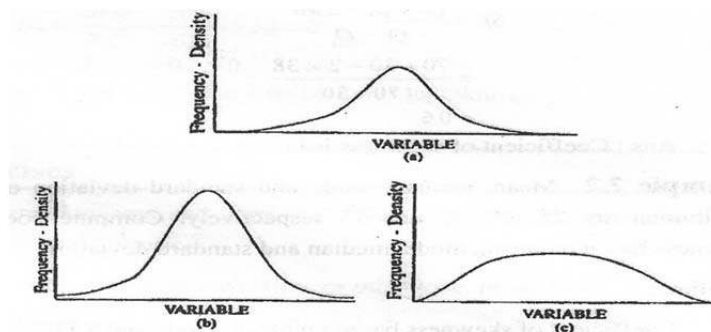


Fig. Three symmetrical distributions with different degrees of kurtosis : (a) mesokurtic, (b) leptokurtic, (c) platykurtic.

Concentration of values in the neighborhood of the central tendency and low tails, compared to a normal distribution with the same standard deviation. A normal curve is said to be mesokurtic (i.e. having medium kurtosis). A distribution with high peak is called leptokurtic, and one with flat peak is known as platykurtic.

Some Solved Examples of skewness and kurtosis-

Example- In a frequency distribution $Q_1=30$, $Q_3=70$ and median is 38, compute coefficient of skewness.

Solution- Coefficient of skewness based on quantities is

$$Sk = \frac{Q_3 + Q_1 - 2M_i}{Q_3 - Q_1} = \frac{70 + 30 - 2 \times 38}{70 - 30} = 0.6$$

Answer- Coefficient of skewness is 0.6

Example- Mean, median and mode and standard deviation of frequency distribution are 25, 27, 35 and 15 respectively. Compute coefficients of skewness based on mean, mode, median and standard deviations.

Solution- Coefficient of skewness based on mean, mode and S.D. is

$$Sk = \frac{\bar{x} - M_0}{s} = \frac{25 - 35}{15} = \frac{-10}{15} = -0.67$$

Coefficient of skewness based on mean, median and S.D. is

$$Sk = \frac{3(\bar{x} - M_i)}{s} = \frac{3(25 - 27)}{15} = \frac{-6}{15} = -0.4$$

2.6 MEASURES OF RELATIONSHIPS

Before reading this, blow some lights on Bivariate Distribution and scattered diagram- Distribution involving two discrete variables is called a Bivariate distribution. For example, The height and weights of the students of a class in school.

Let (x_i, y_i) , $i = 1, 2, 3, \dots, m$; $j = 1, 2, \dots, n$, be a bivariate distribution. If the pair (x_i, y_i) occurs f_{ij} times then f_{ij} is called the frequency of the pair (x_i, y_i) and $\sum_{i=1}^m \sum_{j=1}^n f_{ij} = N$ = the total frequency.

In general two variables are said to be related if they vary together in a systematic fashion. If two variables change in the same direction (either both increase or both decrease) they are said to be positively related. If the variables change in opposite direction (as one increases, the other decrease and vice versa), they are said to be negatively correlated. For example, The weight of an adult depends to some extent on the height.

Scatter Diagram and its Types- The simplest mode of diagrammatic representation of bivariate data is the use of *scatter diagram* (or dot diagram). A scatter diagram graphically summarizes the data on two quantitative variables by showing the joint distribution of the values on two variables. Each point in a scatter plot represents individual's scores on the two variables represented by the two axes of the graph.

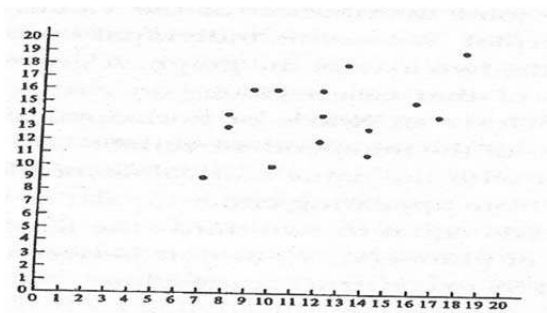
Example- A researcher is interested in studying the relationship between two variables the data are given below:

Case:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X:	7	12	8	12	14	9	18	14	8	12	17	10	16	10	13

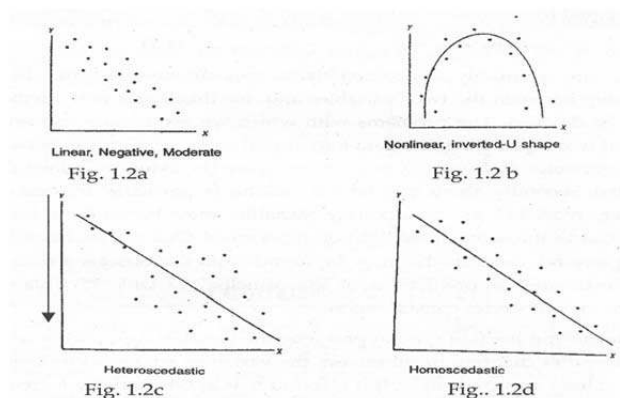
Y:	9	16	14	12	11	16	19	13	13	14	14	16	15	10	18
----	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----

For these data, make a scattered diagram.

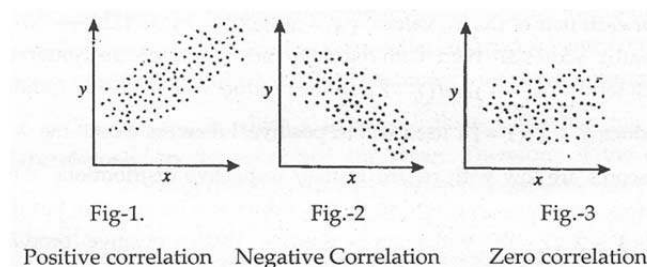
Solution- A scattered diagram of above data



Some more types of Scattered Diagrams are



Sometimes from the scatter diagram, the variables are found to linearly related, at least approximately. If it is found that as one variable increases the other also increases, in general or on the average there will be said to be *positive correlation* between them. Following scatter diagrams shows the nature of correlation between two variables x and y.



2.6.1 KARL PEARSON'S CORRELATION COEFFICIENT

The Pearson's correlation coefficient r_{xy} or product moment correlation coefficient is,

$$r_{xy} = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$$

$$r_{xy} = \frac{\frac{1}{N} \sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\frac{1}{N} \sum (X_i - \bar{X})^2 \right] \left[\frac{1}{N} \sum (Y_i - \bar{Y})^2 \right]}} = \frac{\sum [(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{(X_i - \bar{X})^2} \sqrt{(Y_i - \bar{Y})^2}}$$

$$r_{xy} = \frac{\sum [x_i y_i / N] - \bar{X}\bar{Y}}{\{[\sum_i X_i^2 / N] - \bar{X}^2\}^{1/2} \{[\sum_i Y_i^2 / N] - \bar{Y}^2\}^{1/2}}$$

$$= \frac{N \sum_i X_i Y_i - (\sum_i X_i)(\sum_i Y_i)}{\{N \sum_i X_i^2 - (\sum_i X_i)^2\}^{1/2} \{N \sum_i Y_i^2 - (\sum_i Y_i)^2\}^{1/2}}$$

Example Solution- The necessary information is presented in the following table.
The first two columns contain the given X and Y score.

X_i	Y_i	$X_i Y_i$	X_i^2	Y_i^2
7	9	63	49	81
12	16	192	144	256
8	14	112	64	196
12	12	144	144	144
14	11	154	196	121
9	16	144	81	256
18	19	342	324	361
14	13	182	196	169
8	13	104	64	169
12	14	168	144	196
17	14	238	289	196
10	16	160	100	256
16	15	240	256	225
10	10	100	100	100
13	18	234	169	324
Total	180	210	2577	3050

Putting $N=15$, $\sum X=180$, $\sum Y=210$, $\sum X^2=2320$, $\sum Y^2= 3050$, $\sum XY=2577$ in equation (1.8) and (1.1), we get

$$r_{xy} = \frac{N(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)}{\sqrt{[N(\sum X_i^2) - (\sum X_i)^2] \{N(\sum Y_i^2) - (\sum Y_i)^2\]}}$$

$$r = r_{xy} = r_{yx} = \frac{15(2577) - (180)(210)}{\sqrt{[(15 \times 2577 - 180^2)(15 \times 2320 - 210^2)]}} = 0.43$$

as a measure of degree of relationship between linearly related variable X & Y.

Properties of Correlation Coefficient- The properties of correlation coefficient r are:

- (1) Correlation coefficient is a pure number having no unit.
- (2) The limits of correlation coefficient r lies between -1 to +1, that is $-1 \leq r \leq 1$
- (3) It is not affected by linear transformation of variables, that is, if $u=X-A$, $v=Y-B$, Then $r_{xy}=r_{uv}$ and is independent of A and B.
- (4) If correlation coefficient between X and Y be r and regression coefficient b_{yx} and b_{xy} , then $r = \sqrt{b_{yx}b_{xy}}$
- (5) If two variables x and y are independent then $r = 0$, but converse is not true.

Interpretation of values -1, +1 and 0 or r

- (1) If $r = +1$, it means that r.v. ensure perfect linear relationship between X and Y. In this case all points into scatter diagram lie on a straight line, extending from left bottom to the right bottom, if there is an increase in X then there will be proportional increase in Y.
- (2) If $r = -1$, it means that r.v. Y decreases with the increase in X and vice versa. There is perfect negative relationship between the variables. All the points in the scatter diagram (i.e. on a straight line extending from left top to right bottom).
- (3) The value $r = 0$ confirms the lack of linear relationship between two variables. All the points are scattered on the graph and hardly any these points lie in a straight line.

2.6.2 SPERMAN'S RANK CORRELATION COEFFICIENT

First, let us suppose that there is no tie, i.e., no two individuals are ranked equal in either variable. The ranks x 's and y 's take values 1, 2, 3,..... n in some order. Then spearman correlation coefficient is pronounced as Roh and defined as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

This is Spearman's formula for the rank correlation coefficient. Since the rank correlation coefficient is simple product moment correlation coefficient between two series of ranks, it always lies between -1 to +1.

Rank Correlation Coefficient for Tied Ranks- If some of the individuals have the same rank in ranking they are said to be tied. If the same rank is allocated to m individuals, then we have a tie of length m.

So the Spearman's rank correlation coefficient in the case of tied ranks becomes

$$\rho = \frac{\frac{n^2 - 1}{12} - \frac{T_x + T_y}{2} - \frac{1}{2n} \sum d_i^2}{\left(\frac{n^2 - 1}{12} - T_x\right)^{1/2} \left(\frac{n^2 - 1}{12} - T_y\right)^{1/2}}$$

$$\rho = 1 - \frac{6 \left[\frac{1}{n} \sum d_i^2 + T_x + T_y \right]}{(n^2 - 1)}$$

Example- Ten hand writing were ranked by two judges in a completion. The ranking are given below. Calculate Spearman's rank correlation coefficient.

Hand writing										
	A	B	C	D	E	F	G	H	I	J
Judge-I	3	8	5	4	7	10	1	2	6	9
Judge-II	6	4	7	5	10	3	2	1	9	8

Solution- The differences of ranks between the ranks $d_i = x_i - y_i$ of two judges for ten observations are -3, 4, -2, -1, -3, 7, -1, 1, -3, 1

Hence

$$\sum d_i^2 = 9 + 16 + 4 + 1 + 9 + 49 + 1 + 1 + 9 + 1 = 100$$

Thus Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 100}{10(10^2 - 1)} = 1 - \frac{600}{900} = 0.394 \text{ Answer}$$

Example- Two supervisors ranked 12 workers working under them in order of efficiency as follows.

No. of workers	1	2	3	4	5	6	7	8	9	10	11	12
Supervisor-I	5	6	1	2	3	8.5	8.5	4	7	11	10	12
Supervisor-II	5.5	5.5	2	2	2	9	7	4	8	10.5	12	10.5

Calculate rank correlation coefficient.

Solution- In the ranking, of first supervisor, there 15 one tie of length 2.

Thus

$$T_x = \frac{1}{12} \frac{2^3 - 2}{12} = 0.0417$$

In the ranking of second supervisor, there are three ties of length 2, 3 and 2 respectively. Thus

$$T_y = \frac{1}{12} \left[\frac{2^3 - 2}{12} + \frac{3^3 - 3}{12} + \frac{2^3 - 2}{12} \right] = 0.25$$

$$\sum d_i^2 = .25 + .25 + 1 + 0 + 1 + .25 + 2.25 + 0 + 1 + .25 + 4 + 2.25 = 12.50$$

$$\begin{aligned} \rho &= \frac{\frac{n^2 - 1}{12} - \frac{T_x + T_y}{2} - \frac{1}{2n} \sum d_i^2}{\left(\frac{n^2 - 1}{12} - T_x \right)^{1/2} \left(\frac{n^2 - 1}{12} - T_y \right)^{1/2}} \\ &= \frac{\frac{12^2 - 1}{12} - \frac{0.0417 + 0.25}{2} - \frac{12.50}{2 \times 12}}{\sqrt{\frac{12^2 - 1}{12} - 0.0417} \sqrt{\frac{12^2 - 1}{12} - 0.25}} \\ &= \frac{11.25}{\sqrt{11.8750} \times \sqrt{11.6667}} = 0.956 \quad \text{Answer.} \end{aligned}$$

2.7 REGRESSION ANALYSIS

The regression analysis is used for Prediction, which is an objective to all sciences, including social and behavior sciences. Prediction is based on relationship between or among variables. There are some types of regression analysis:

Simple regression- The regression analysis confined to the study of only two variables at a time is called the simple regression.

Multiple Regression- The regression analysis for studying more than two variables at a time is known as multiple regression.

Linear Regression- If the regression curve is a straight line, then there is a linear regression between the variables under study. In other words, in linear regression the relationship between the two variables X and Y is linear.

Non linear Regression- If a curve or regression is not a straight line, then it is called a non-linear or curvilinear regression.

Lines of Regressions- A line of regression is the line which gives the best estimate of one variable for any given value of the other variable.

2.7.1 LINE OF REGRESSION OF X ON Y.

It is the line which gives the best estimate for the values of X for a specified value of Y.

$$\text{It is given by, } X = \bar{X} + r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

Where $\bar{X}$, $\bar{Y}$ are means of X series and Y series respectively and σ_x , σ_y are S.D. of X and Y series respectively and r is the correlation coefficient between X and Y.

2.7.2 LINE OF REGRESSION OF Y ON X.

It is the line which gives the best estimates for the values of Y for any specific values of X.

Regression equation of Y on X is given by:

$$Y = \bar{Y} + r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

It can also be put in the form

Regression Coefficients- The regression coefficient of Y on X is $b_{yz} = r \frac{\sigma_y}{\sigma_x}$ and that of x on y is $b_{xy} = r \frac{\sigma_x}{\sigma_y}$

Some Important Properties Relating to Regression Coefficient-

1. Numerically the correlation coefficient is the geometric mean of the two regression coefficients. As regards the sign of r, it is the same as the common sign of the two regression coefficients.

$$|r| = \sqrt{b_{yx} \times b_{xy}}$$

2. If one of the regression coefficients is greater than unity, then the other is less than unity.

$$b_{yx} \leq \frac{1}{b_{xy}} < 1$$

3. Arithmetic mean of the regression coefficient is greater than the correlation coefficient.
4. Regression coefficients are Independent of change of origin but not of scale.
5. The sign of correlation is same as that of regression coefficients,

Example- Find both the regression equations from the following data:

$$\sum X = 60 \qquad \sum Y = 40 \qquad \sum XY = 1150$$

$$\sum X^2 = 4160 \quad \sum Y^2 = 1720 \quad N = 10$$

Solution- The regression coefficient b_{xy} and b_{yx} given by:

$$b_{xy} = \frac{\sum XY - \frac{\sum X - \sum Y}{N}}{\sum Y^2 - \frac{(\sum Y)^2}{N}} = \frac{1150 - \frac{60 \times 40}{10}}{1720 - \frac{(40)^2}{10}} = \frac{910}{1560} = 0.58$$

And

$$b_{yx} = \frac{\sum XY - \frac{\sum X - \sum Y}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}} = \frac{910}{4160 - \frac{(60)^2}{10}} = 0.24$$

Also

$$\bar{X} = \frac{\sum X}{N} = \frac{60}{10} = 6; \quad \bar{Y} = \frac{\sum Y}{N} = \frac{40}{10} = 4$$

Now, the regression equation of Y on X is-

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\Rightarrow Y - 4 = 0.24(X - 6) \text{ or } Y = 0.24X + 2.56.$$

The regression of X on Y is-

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$\Rightarrow X - 6 = 0.58(Y - 4) \text{ or } X = 0.58Y + 3.68.$$

Example- From the following results, obtain the two regression equations and estimate the Y when the X is 29 . And the X, when the Y is 600:

	Y	X.
Mean	508.4	26.7
S.D.	36.8	4.6

Coefficient of correlation between yield and rainfall is + 0.52.

Solution-

Regression of y on x is given by $y - \bar{y} = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$

Or

$$y - 508.4 = 0.52 \times \frac{36.8}{4.6}(x - 26.7) \Rightarrow y - 508.4 = 0.52 \times 8(x - 26.7)$$

$$\Rightarrow y - 508.4 = 4.16x - 111.072 \quad \Rightarrow y = 4.16x + 397.328$$

When $x = 29$, $y = 4.16 \times 29 + 397.328 = 517.968$

Regression of x on y is given by $x - \bar{x} = r \frac{\sigma_y}{\sigma_x} (y - \bar{y})$

Or

$$\begin{aligned} x - 26.7 &= 0.52 \times \frac{4.6}{36.8} (y - 508.4) = 0.065(y - 508.4) \\ &= 0.065 y - 33.046 \end{aligned}$$

Or $x = 0.065 y - 6.346$

When $y = 600$, $x = 0.065 \times 600 - 6.34 = 32.654$

2.8 ASSOCIATION OF ATTRIBUTES

In real life we often study of characteristics/ qualitative data the numerical value are not possible. Simply we can observe the presence or absence of these specific characteristics (or properties) over a group of individuals or units under observation. Such characteristics are termed as attributes. Hence attribute is a qualitative character which cannot be measured numerically and one simply observes the presence or absence of it. Honesty, beauty, preferences, likings, colors, blindness, smoking etc are few examples of attributes. If an attribute is classified in two groups it is called a dichotomous attribute whereas if it is classified in many categories it is called manifold.

Combinations, Classes and Class Frequencies of Attributes (Dichotomous Classification)- Different attributes, their subgroups and combinations are called classes and the numbers of observation assigned to them are called respective class frequencies.

Example- Consider a dichotomous attribute “Smoking” which may be divided into two categories

1. Smokers (A)
2. Nonsmokers (α)

Let us further consider one more dichotomous attribute “Literacy” which is also classification into two categories

1. Literate (B)
2. Illiterate (β)

Both of these attributes are called dichotomous as they are divided into two subgroups and any individual will belong to only one of these categories either A or α and B or β .

Symbolically, we may write,

(AB) to denote the number of individuals who are smokers and literate.

$(A\beta)$ to denote the number of individuals who are smokers and illiterate.

(α, B) to denote the number of individuals who are nonsmokers and illiterate.

(α, β) to denote the number of individuals who are non smokers and illiterate.

Thus in general (AB) stands for the frequency of the individuals or units which possess the attribute A and B simultaneously. Similar interpretations exists (α, β) for etc. Total number of observation is denoted by N.

Obviously we have,

$$N = (A) + (\alpha) = (B) + (\beta)$$

$$N = (AB) + (A\beta) + (A\beta) + (\alpha\beta)$$

Order of Class and Class Frequencies- Total number of observation N is taken as one class of order “Zero”. (A), (B), (C), etc will be each of order one, $(AB) + (A\beta) + (A\beta) + (\alpha\beta)$ etc. will be each of order three and so on. A class or class frequency (ABC...M, r letters) containing r attributes in it said to be of order “r”.

With n attributes there are in all 2^n positive class frequencies.

With n attributes the number of negative class frequencies will be $(2^n - 1)$, because except N.

Consistency of Data- A data is said to be consistent if all of its class frequencies have been appeared to have been observed within one and the same population. Class frequencies of a consistent data will not have any mutual contradiction rather they will support each other.

Example- If, in a series of houses actually invaded by small pox 70% of the inhabitants are attacked and 85% have been vaccinated. What is the lowest % of the vaccinated that have been attacked?

	(B) Vaccinated	(β) Not vaccinated	
(A)Attacked	AB	$(A\beta)$	70
(α) Not attacked	(βB)	$(\alpha\beta)$	30
	85	15	100

Solution- Given that, $(A)=70$, $(B) = 85$, $(\alpha) = 30$, $(\beta) = 15$, $N = 100$

Since, $(AB) \geq 0$ and $(\alpha\beta) \geq 0 \Rightarrow (1-A)(1-B)N \geq 0$

or $(1 - A - B - AB)N \geq 0$

or $N - (A) - (B) + (AB) \geq 0$

or $(AB) \geq (A) + (B) - N$

or $(AB) \geq (70) + (85) - 100$

or $(/ \setminus B) \geq 55$

So, out of 85 vaccinated at least 55 are being attacked.

Hence at least $\frac{55}{85} \times 100 = 64.7\%$ vaccinated have been attacked.

Contingency Tables- Let there be two attributes divided into two categories i.e. we are considering dichotomous classification attributes. In this case there are 9 class frequencies. These class frequencies in usual notations may be written in a 2×2 table as follows:

Attributes	A	α	Total
B	$(AB) = a$ (say)	$(\alpha B) = b$ (say)	(B)
β	$(A\beta) = c$ (say)	$(\alpha\beta) = d$ (say)	(β)
Total	(A)	(α)	N

This table is known as 2×2 contingency table. The contingency table need not be 2×2 only.

Independence and Association of Attributes- Consider two dichotomous attributes A and B.

So.

$$(A) = (AB) + (A\beta)$$

$$(\alpha) = (\alpha B) + (\alpha\beta)$$

$$(B) = (AB) + (\alpha B) \quad (1)$$

$$(\beta) = (A\beta) + (\alpha\beta) \quad (2)$$

$$N = (AB) + (A\beta) + (\alpha B) + (\alpha\beta)$$

Further,

$$N = (A) + (\alpha) \quad (3)$$

and

$$N = (B) + (\beta) \quad (4)$$

For A and B to be independent we must have

$$\frac{(AB)}{(A)} = \frac{(\alpha\beta)}{(\alpha)} = \frac{(B)}{N}$$

Thus, $(AB) = \frac{(A) \cdot (B)}{(N)}$ then A and B are said to be independent.

If $(AB) > \frac{(A) \cdot (B)}{N}$, then A and B are said to be positively associated.

If $(AB) < \frac{(A).(B)}{N}$, then A and B are said to be negatively associated.

In general, if attributes are not independent they are said to be associated.

2.9 3 - SIGMA LIMITS

Three-sigma limits (3-sigma limits) is a statistical calculation that refers to data within three standard deviations from a mean. Three-sigma limits are used to set the upper and lower control limits. Mathematically it is defined as $\bar{x} \pm 3\sigma$. Here $\bar{x} + 3\sigma$ is upper limit and $\bar{x} - 3\sigma$ is lower limit.

2.10 SUMMARY

This unit gives detailed knowledge about Measures of central tendency, measures of dispersion, measures of symmetry, measures of relationships, predictions, association of attributes and 3-sigma limits, with numerical examples.

2.11 CHECK YOUR PROGRESS

1. A sample of 12 fathers and their eldest sons gave the following data about their heights in inches:

Father:	65	63	67	64	68	62	70	66	68	67	69	71
Son:	68	66	68	65	69	66	68	65	71	67	68	70

Calculation coefficient of rank correlation.

2. Calculate mean, median, mode and variance.

<i>Class group (X)</i>	<i>No. of observations(f)</i>
0-10	14
10-20	23
20-30	27
30-40	21
40-50	15
Total	100

3. Given that $(A) = (\alpha) = (B) = (\beta) = \frac{1}{2}$ show that $(AB) = (\alpha\beta)$.

4. What are skewness and kurtosis? Give some suitable measures for skewness
5. Calculate Karl Pearson's coefficient of correlation for the data given below:

Independent Variable X	3	7	5	4	6	8	2	7
Dependent Variable Y	7	12	8	8	10	13	5	10

6. Find the two lines of regression from the following data:

Age of husband	25	22	28	26	35	20	22	40	20	18
Age of wife	18	15	20	17	22	14	16	21	15	14

Hence estimate:

- i. the age of husband when the age of wife is 19 and
- ii. The age of wife when the age of husband is 30.

2.12 FURTHER READINGS

1. Goon, Gupta & Dasgupta: Fundamentals of Statistics Vol. I the World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kendall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Ltd.
3. C.E. Weatherburn: Mathematical Statistics. Kenney, J.F and Keeping, E.S.: Mathematics of Statistics, Part I Van Nostrand, 1954 and Affiliated East- West Press.
4. Mills, F.C.: Statistical Methods (Ch.5) H. Holt, 1955.
5. Arora P. N., Arora Sumit. Arora A; Comprehensive Statistical Methods, S. Chand, New Delhi
6. Gupta S. C., Kapoor,V.K., Fundamentals of Mathematical Statistics, S. Chand, New Delhi.



॥ सरस्वती नः सुभगा मयस्करत् ॥

Uttar Pradesh Rajarshi Tandon
Open University

PGBCH-106

Biostatistics

BLOCK

2

PROBABILITY-DISTRIBUTION THEORY AND DEMOGRAPHY

UNIT-3

Probability and Distribution Theory

UNIT-4

Demography

Course Design Committee

Dr. (Prof.) Ashutosh Gupta School of Science, UPRTOU Prayagraj	Director
Prof. Umesh Nath Tripathi Department of Chemistry Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur	Member
Prof. S.I. Rizvi Department of Biochemistry University of Allahabad, Prayagraj	Member
Prof. Dinesh Yadav Department of Biotechnology Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur	Member
Prof. Sharad Kumar Mishra Department of Biotechnology Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur	Member
Dr. Ravindra Pratap Singh Academic Consultant (Biochemistry) School of Sciences, UPRTOU Prayagraj	Member
Dr. Dharmveer Singh Academic Consultant (Biochemistry) School of Sciences, UPRTOU Prayagraj	Member/Secretary

Course Preparation Committee

Dr. Shruti Sr. Assistant Professor, School of Sciences U. P. Rajarshi Tandon Open University, Prayagraj	Writer
Prof. G. S. Pandey Department of Statistics, University of Allahabad, Prayagraj	Editor
Dr. Dharmveer Singh Academic Consultant (Biochemistry) School of Science, UPRTOU, Prayagraj	Course Coordinator

©UPRTOU, Prayagraj - 2021
ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the
Uttar Pradesh Rajarshi Tandon Open University, Prayagraj.

BLOCK INTRODUCTION

The present SLM on **Bio Statistics** consists of three Blocks. *Block - 1 – Research Methodology and Statistical Methods* has two units; *Block - 2 – Probability - Distributions and Demography* has two units; and at the last *Block – 3 Tests of Significance and Analysis of Variance* has two units.

The **Block-1– Research Methodology and Statistical Methods** consists two units. The *first unit* of this block; named *Research Methods and Sampling Procedures*; describes the meaning and types of research, significance of research. It tells about the research problem and its selection with *Sampling Theory*; which discuss about the sampling, different types of sampling designs, simple random sampling, stratified sampling and cluster sampling with their applications.. The *second unit* of this block; named *Statistical Tools*; discusses about the measures of central tendency, measures of dispersion, measures of asymmetry, correlation and regression analysis, association of attributes and 3-sigma limits.

The **Block-2– Probability- Distributions Theory and Demography** is the second block having two units. The *first unit* of this block; named *Probability and Distribution Theory*; gives the Basic concepts of probability, definitions of probability, additive and multiplicative law of probability, conditional probability, Bayes' theorem, random variable and its types, probability mass function and probability density functions. In *Probability Distributions*; the concept of probability distribution, discrete and continuous probability distributions namely Binomial Distribution, Poisson Distribution, Geometric Distribution, Normal Distribution, Exponential Distribution have been also discussed in this unit along with their properties, applications and importance.. The *second unit* of this block; named *Demography*; gives knowledge about the vital statistics and demography, this also tells about the source of vital statistics and demographic data, rates, ratio, proportion, measures of fertility, measures of mortality, measures of morbidity and migration.

The **Block-3– Tests of Significance and Analysis of Variance** consists of two units. The *first unit* of this block; named *Testing of Hypothesis*; discuss about the hypothesis and its types, level of significance, critical region, p-value, types of errors, chi-square tests, t-tests and z-tests with their applications. The *second unit* of this block; named *Analysis of Variance*; discusses about the concept of analysis of variance and co-variance, basic principles of ANOVA and ANCOVA. (One Way, Two Way and Three Way Analysis).

Illustrations and examples on these topics have also been given.

At the end of every block/unit the summary, self assessment questions and further readings are given.

UNIT-3 PROBABILITY AND DISTRIBUTION THEORY

Structure

- 3.1 Introduction
- 3.2 Objectives
- 3.3 Definition and laws of Probability
 - 3.3.1 Additive law of Probability
 - 3.3.2 Multiplicative law of Probability
- 3.4 Conditional Probability
- 3.5 Baye's Theorem
- 3.6 Random Variable, Probability Function and its Types
 - 3.6.1 Discrete R.V. and Probability Mass Function
 - 3.6.2 Continuous R.V. and Probability Density Function
- 3.7 Probability Distributions
 - 3.7.1 Discrete Distributions
 - 3.7.1.1 Binomial Distribution
 - 3.7.1.2 Poisson Distribution
 - 3.7.1.3 Geometric Distribution
 - 3.7.2 Continuous Distributions
 - 3.7.2.1 Normal Distribution
 - 3.7.2.2 Exponential Distribution
- 3.8 Summary
- 3.9 Check Your Progress
- 3.10 Further Readings

3.1 INTRODUCTION

In various fields of social, biological physical sciences etc., we come across with experiments and phenomenon in which some kind of uncertainty is involved. This uncertainty leads to the study of Probability. This unit introduces the basic idea of such approach to probability will be explained and their distributions will be discussed.

3.2 OBJECTIVES

After going through this unit you shall be able to:

- Understand the probability and its laws/properties.
- Understand probability function and its types.
- about probability distributions

3.3 DEFINITION AND LAWS OF PROBABILITY

Before study the definition of Probability, the information of these given terms are important.

- The experiment is termed as *Trial* and the results are *Events* or *Cases*, e.g. tossing of a coin is a trial and getting head or tail is event.
- The total number of all possible results of any trial is *Exhaustive cases*, e.g. in above example there are only two exhaustive cases.
- When the happening of one event affects the happening of another, is *Mutually Exclusive*. In other words both events cannot occur at the same time, e.g. in above example head or tail is mutually exclusive events.
- When the result of one event does not affected by the other, is *Independent events*, e.g. if a coin tossed twice, the result of both throw is not affected to each other.
- when the happening of every event is same, called *Equally Likely events* e.g. in above example the occurrence of head or tail is equally likely.

Definition- If an event can occur in n total number of exhaustive/ mutually exclusive/equally likely cases and m of them are in favorable to the happening of a particular event, say A , then the probability of occurrence of the event A is given by

$$P(A) = p = \frac{\text{Favourable number of cases}}{\text{Exhaustive number of cases}} = \frac{m}{n}$$

p is known as probability of success and probability of failure is q .

$$q = 1 - p = 1 - \frac{m}{n} = \frac{n-m}{n}$$

$$\text{Hence } p + q = 1 \quad \text{And} \quad 0 \leq p, q \leq 1$$

Hence the mathematical value of p is lies between zero to unity (one). It may not be negative. If the mathematical value of p is 0, then related event is called impossible event and if the mathematical value of p is 1, then related event is called sure possible event.

Laws of Probability- On the basis of the above definition of the probability some rules are important to know:

3.3.1 ADDITIVE LAW OF PROBABILITY

If $A_1, A_2, \dots, A_n$ are the mutually exclusive events then

$$P(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

If $A_1, A_2, \dots, A_n$ are the equally likely events then the sum of their probabilities is unity. i.e.

$$P(A_1) + P(A_2) + \dots + P(A_n) = 1$$

In other words

For two mutually exclusive events A and B, $(A \cap B = \emptyset)$.

$$P(A \cup B) = P(A) + P(B)$$

For pair wise mutually exclusive events $A_1, A_2, \dots, (A_i \cap A_j = \emptyset \forall i \neq j)$,

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

3.3.2 MULTIPLICATIVE LAW OF PROBABILITY

If $A_1, A_2, \dots, A_n$ are the independent events then

$$P(A_1 \text{ and } A_2 \text{ and } \dots \text{ and } A_n) = P(A_1) \times P(A_2) \times \dots \times P(A_n)$$

This is called **Multiplicative law of Probability**.

3.4 CONDITIONAL PROBABILITY

Definition- Let A, B be two events defined on the same sample space Ω . Then conditional probability of B given A denoted by $P(B|A)$, is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad \text{provided } P(A) \neq 0.$$

The conditional probability satisfies all the rules of a probability.

Definition- Two events A and B are said to be independent whenever

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

Note- If A and B are independent events then $P(B|A) = P(B)$ and $P(A|B) = P(A)$, provided

$$P(A) > 0, P(B) > 0.$$

3.5 BAYE'S THEOREM

Let $B_1, \dots, B_k$ be a partition of the sample space Ω and A be an event ($A \subset \Omega$), Then

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^k P(A|B_j)P(B_j)}$$

Proof- We have

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{P(A)}$$

Where the last equality follows from the multiplicative law of probability. Further, from the previous result, we have,

$$P(A) = \sum_{j=1}^k P(A|B_j)P(B_j)$$

Substituting the value of $P(A)$, we obtain

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^k P(A|B_j)P(B_j)} \quad \text{Hence the result}$$

Note- Here the probabilities $P(B_j)$'s are known as “**a priori (or prior) probabilities**”. They exist before we gain any information about A (the result of an experiment). The probabilities $P(A|B_j)$'s are known as “**likelihoods**”. They indicate how likely event A to occur under the information that B_j occurs. The probabilities $P(B_j|A)$'s are known as “**a posteriori (or posterior) probabilities**”. They are determined after the results of the experiment are known.

Example- A bag contains 10 coins out of which five coins of type I are unbiased and remaining five coins are biased. Among the biased coins, four coins of type II have probabilities of head $1/3$ are remaining one coins of type III have probability of head $9/10$. A coin is selected at random and tossed three times. If we get heads in a row, what is the probability that the coin is of type III.

Solution- Let

B_1 : Event the coin is of type I

B_2 : Event the coin is of type II

B_3 : Event the coin is of type III

A : Event that we get three heads in a row.

Given that $P(B_1) = 0.5$, $P(B_2) = 0.4$, $P(B_3) = 0.1$

$$P(A|B_1) = (1/2)^3, P(A|B_2) = (1/3)^3, P(A|B_3) = (9/10)^3$$

Then the probabilities that the coin is of type III given that three heads have been obtained in a row, i.e., $P(B_3|A)$, is given by

$$P(B_3|A) = \frac{P(A|B_3) P(B_3)}{P(A|B_1) P(B_1) + P(A|B_2) P(B_2) + P(A|B_3) P(B_3)}$$

$$= \frac{\left(\frac{9}{10}\right)^3 \times 0.1}{\left(\frac{1}{2}\right)^3 \times 0.5 + \left(\frac{1}{3}\right)^3 \times 0.4 + \left(\frac{9}{10}\right)^3 \times 0.1} = 0.46$$

3.6 RANDOM VARIABLES AND PROBABILITY FUNCTION

A random variable is a variable whose value is unknown or a function that assigns value to each of an experiment's result. It is used to measure outcomes of random occurrences in probability theory.

There are two types of random variables; discrete and continuous.

3.6.1 DISCRETE RANDOM VARIABLE AND PROBABILITY MASS FUNCTION

A random variable X is said a discrete random variable if it takes finite or countable infinite number of values. Thus the range of a discrete random variable has countable number of points.

Let X be a discrete random variable with $x_1, x_2, \dots, x_i$, possible outcomes than $P(x_i) = P(X = x_i)$; $i = 1, 2, \dots$. Satisfying the following conditions:

- (i) $p(x_i) \geq 0$ for all $i = 1, 2, \dots$
- (ii) $\sum_{i=1}^n p(x_i) = 1$

Then the function $p(x)$ satisfying the above conditions is called the **probability mass function** (pmf) of the random variable X .

The collection of pairs $(x_i, p(x_i))$; $i = 1, 2, \dots$ is called the probability distribution of X . the cdf of X is given by

$$F(x) = \sum_{i, x_i \leq x} p(x_i)$$

3.6.2 CONTINUOUS RANDOM VARIABLE AND PROBABILITY DENSITY FUNCTION

A random variable X is said a continuous random variable if it takes infinite number of values. Thus the range of a continuous random variable has uncountable number of points.

For continuous random variables, there exists a function $f(x)$ called the **probability density function** (pdf) of X , such that

- (i) $f(x) \geq 0$ for all x .
- (ii) $\int_{-\infty}^{\infty} f(x) dx = 1$

- (iii) For real number a, b with $-\infty < a < b < \infty$, the probability that X lies in interval $(a, b]$ is given by

$$P(a < X \leq b) = \int_a^b f(x)dx$$

Since for a continuous random variable X , the probability of a single point $P(X = x) = 0$ for all x , we have

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b)$$

The cdf of X is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x')dx'$$

Conversely, the pdf of X , in terms of cdf, is given by

$$f(x) = \frac{d}{dx} F(x)$$

3.7 PROBABILITY DISTRIBUTION

There are two types of probability distributions: discrete and continuous.

3.7.1 DISCRETE DISTRIBUTIONS

The distribution functions which are based on discrete random variables is known as discrete distributions. Here we study about the binomial distribution, Poisson distribution and Geometric distribution.

3.7.1.1 BINOMIAL DISTRIBUTION

Definition- The binomial distribution with parameter n, p distribution of random variable X for which,

$$P[X = x] = \binom{n}{x} p^x q^{n-x} \quad (x = 0, 1, 2, \dots, n)$$

$$\binom{n}{x} p^x q^{n-x} = \frac{n!}{x! (n-x)!} p^x q^{n-x}$$

Note-

- The mean of binomial distribution is np and variance is npq .
- The sum of independent binomial variates is not a binomial variate. In other words, binomial distribution not holds the additive or reproductive property.

Example- Ten coins are throwing simultaneously. Find the probability of getting at least eight heads.

Solution- p = probability of getting a head = $1/2$

q = probability of not getting a head = $1/2$

The probability of getting x heads in a throw of 10 coins is

$$p(x) = {}^{10}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x} = {}^{10}C_x \left(\frac{1}{2}\right)^{10} \cdot x = 0, 1, 2, \dots, 10$$

$\therefore$ Probability of getting x heads is given by

$$P[x \geq 8] = P(8) + P(9) + P(10)$$

$$= \left(\frac{1}{2}\right)^{10} \left[{}^{10}C_8 + {}^{10}C_9 + {}^{10}C_{10} \right] = \frac{45 + 10 + 1}{1024} = \frac{56}{1024} = \frac{7}{128}$$

Example 2- The probability that a person recovers from a serious disease is 0.30. Find the probability that at least one of 5 persons admitted to a hospital will survive.

Solution- If X denotes the number of persons recovered from disease. Then we have to find $P[X \geq 1]$. X has binomial distribution with $n = 5$, $p = 0.30$.

We have

$$P[X \geq 1] = 1 - P[x = 0]$$

$$= 1 - {}^5C_0 (0.30)^0 (0.70)^5 = 1 - 0.168 = 0.832$$

Example- Comment on the following:

The mean of a binomial distribution is 4 and variance is 5.

Solution- If the given binomial distribution has parameter n and p then we have

$$\text{mean} = np = 4 \quad \text{and} \quad \text{variance} = npq = 5$$

Dividing eqn. we get

$$q = 5/4$$

which is impossible, since probability cannot exceed unity. Hence given statement is wrong.

Example- The mean and variance of binomial distribution are 3 and $3/4$ respectively. Find $P[X \geq 1]$.

Solution-

$$\text{Mean} = E(X) = np = 3$$

$$\text{Var}(x) = npq = 3/4$$

$$\text{Dividing we get } q = 1/4$$

$$\therefore p = \frac{3}{4}$$

$$\text{Now we have } np = 3$$

$$\Rightarrow \frac{3}{4}n = 3$$

$$\Rightarrow n = \frac{4}{3} \times 3 = 4$$

$$\therefore P[X \geq 1] = 1 - P[X = 0] = 1 - q^n$$

$$= 1 - \left(\frac{1}{4}\right)^4 = 1 - \frac{1}{256} = \frac{255}{256} = 0.996$$

3.7.1.2 POISSON DISTRIBUTION

The probability distribution having $p(x) = p(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$, Where $\lambda > 0$ is a constant; as its probability mass function is known as Poisson distribution with parameter λ .

This distribution was discovered by Simen Denis Poisson in 1837. Poisson distribution is a limiting case of binomial distribution under the following conditions:

- (i) n , the number of trials is indefinitely large, i.e., $n \rightarrow \infty$.
- (ii) p , the constant probability of success for each trial is indefinitely small i.e., $p \rightarrow 0$.
- (iii) $np = \lambda$ (say), is finite. Thus $= \frac{\lambda}{n}$ and $q = 1 - \frac{\lambda}{n}$, where λ is a positive real number.

Note-

- The mean and variance of Poisson Distribution is λ .
- Sum of independent Poisson variates is also a Poisson variate

The following are some examples of Poisson Variate:

- (i) The number of defective screws per box of 100 screws.
- (ii) The number of printing mistakes at each page of a book.
- (iii) The number of deaths in a district in one year by rare disease.
- (iv) The number of air accidents in some unit of time.
- (v) The number of suicides reported in a particular city.
- (vi) The number of cars passing through a certain street in time.

3.7.1.3 GEOMETRIC DISTRIBUTION

Definition- A random variable X is said to have the geometric distribution with parameter p , $0 < p < 1$, if its p.m.f. is given by

$$P[X = x] = \begin{cases} q^{x-1}p, & x = 0, 1, 2, 3 \dots \\ 0 & \text{other wise} \end{cases}$$

Note-

- The mean = $1/p$ and $V(X) = q/p^2$, hence for geometric distribution variance is greater than mean.

3.7.2 CONTINUOUS DISTRIBUTION

The distribution functions which are based on continuous random variables is known as continuous distributions. Here we study about the normal distribution and exponential distribution.

3.7.2.1 NORMAL DISTRIBUTION

It has been observed that most business and economic variables result in continuous data whose behavior is often best described by a bell-shaped continuous curve. Since most populations on these variables normally yield a bell-shaped curve, such a curve has come to be universally known as a normal curve. Accordingly, the probability distribution described by normal curve is called the normal (probability) distribution.

The normal distribution is defined by the probability density function.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \text{ for } -\infty \leq x \leq \infty \quad \dots \dots \dots (3.1)$$

Where μ , σ and σ^2 are two parameters of the distribution.

A continuous r.v. X having above pdf is called normal random variable. We write it as $X \sim N(\mu, \sigma^2)$ μ and σ^2 are the parameters of the distribution. We shall see μ that is mean and σ^2 is variance of the distribution.

This curve is bell-shaped and symmetric about mean μ . As the curve is symmetrical about the ordinal at $x = \mu$, The curve is concave down wards in the centre μ , but after $x = \mu \pm \sigma$ is becomes concave upwards. The two points of inflexion are given by $x = \mu \pm \sigma$,

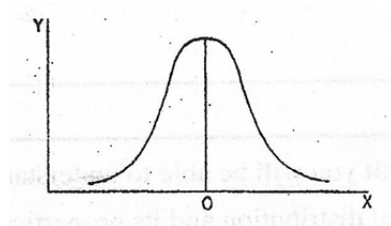


Figure: 3.1 Normal Distribution

It shows the deviations of the values of normal variables X from its mean μ . The larger these deviations, the higher are the value of standard deviations σ (or variance σ^2) which is the denominator in the exponent.

The shape of the curve shows that the observations occur most frequently in the neighborhood of the mean and their frequency decreases as they move away from the mean. From (3.1) it is seen that the distribution is symmetrical about the point $x = \mu$. $f(\mu + a) = f(\mu - a) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp(-\sigma^2 a^2)$.

Whatever μ may be. Hence μ is mean as well as median of the distribution. Again μ is also mode of the distribution, since $f(\mu) = 0$ and $f^x \mu < 0$. Also $\exp\left(-\frac{a^2}{\sigma^2}\right)$ decreases monotonically as a^2 increases from zero, i.e., as a deviates from zero in either direction.

Thus the mean, median and mode of the distribution coincide at μ . Thus the mode μ of the distribution also coincides with mean and median μ .

Approximately 68% of the area under the curve lies in the region $[\mu - \sigma, \mu + \sigma]$, approximately 95% in $[\mu - 2\sigma, \mu + 2\sigma]$ and almost all in the region $[\mu - 3\sigma, \mu + 3\sigma]$. In fact approximately 0.27% area lies outside region $[\mu - 3\sigma, \mu + 3\sigma]$. This is Area Property.

Normal Distribution with mean μ and variance σ^2 is also denoted by $N(\mu, \sigma^2)$ or $N(\mu, \sigma)$. Mean, Median and Mode of this distribution, coincide at μ .

Standard Normal Distribution- If $X \sim N(\mu, \sigma^2)$ then its pdf is given by equ. (3.1) using transformed

$$z = \frac{x - \mu}{\sigma}$$

We obtain the pdf of z as

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad \text{for } -\infty < z < \infty \quad (3.1a)$$

Obviously $E(z) = 0$ and $\text{var}(Z) = 1$

Hence, z is standard normal variable and the pdf given by equ. (3.1a) is called standard normal distribution.

Properties of Normal Distribution-

- The normal distribution is symmetrical about $X = \mu$ i.e., there is no skewness in normal distribution
- The normal distribution is mesokurtic i.e., Kurtosis of a normal distribution vanishes; it is neither platykurtic nor leptokurtic.
- Any linear function of normal variate is also normally distributed.

Properties of a Normal Curve-

- (i) Normal curve is symmetric; bell shaped and unimodal.
- (ii) The mean, median and mode all coincide.

3.7.2.2 EXPONENTIAL DISTRIBUTION

Exponential distribution plays important role in statistics. It has been used as potential model for lifetimes of many things. Exponential distribution is also used for continuous waiting time random variable of various events whereas geometric distribution is used for discrete waiting time random variable.

Definition- A continuous random variable X is said to follow an exponential distribution if it assumes non negative values with probability density function (p.d.f.) given by

$$f(x, \theta) = \begin{cases} \theta e^{-\theta x}, & x \geq 0, \theta > 0 \\ 0, & \text{otherwise} \end{cases}$$

Here θ is the parameter of exponential distribution

Note-

- Mean and variance of the exponential distribution is

$$\left. \begin{array}{l} \text{Mean} = \frac{1}{\theta} \\ \text{Variance} = \frac{1}{\theta^2} \end{array} \right\} \dots \dots \dots (4.15)$$

3.8 SUMMERY

This unit gives a complete idea about probability, random variables, its types, probability distributions, discrete and continuous distributions.

3.9 CHECK YOUR PROGRESS

1. Discuss about the mean and variance of Binomial Distribution.
2. Is mean and variance of Poisson distribution is equal.
3. In geometric distribution, is mean is greater than variance?
4. Draw the curve of normal distribution.
5. Find the variance of exponential distribution?

3.10 FURTHER READINGS

1. Cramer H, Mathematical Methods of Statistics, Princeton University Press, 1946 and Asia Publishing House, 1962.
2. Hogg R.V. and Craig A.T., Introduction to Mathematical Statistics, Macmillan, 1978.
3. Prazen E., Modern Probability Theory and its Applications, John Wiley, 1960 and Wiley Eastern 1972.

4. Rao C.R., Linear Statistical Inference and Its Applications, John Wiley, 1960 and Wiley Eastern 1974.
5. Rohtagi V.K. (1984), An Introduction to Probability Theory and Mathematical Statistics, John Wiley, 1976 and Wiley Eastern 1985.
6. Vikas S.S., Mathematical Statistics, John Wiley, 1962 and Toppan.
7. Arora P. N., Arora Sumit. Arora A; Comprehensive Statistical Methods, S. Chand, New Delhi
8. Gupta S. C., Kapoor,V.K., Fundamentals of Mathematical Statistics, S. Chand, New Delhi.

UNIT-4 DEMOGRAPHY

Structure

- 4.1 Introduction
- 4.2 Objectives
- 4.3 Rates Ratio and Proportion
- 4.4 Fertility and its Measures
 - 4.4.1 Crude Fertility Rate
 - 4.4.2 Gross Fertility Rate
 - 4.4.3 Age Specific Fertility Rate
 - 4.4.4 Total Fertility Rate
- 4.5 Mortality and its measures
 - 4.5.1 Crude Death Rate
 - 4.5.2 Specific Death Rate
 - 4.5.3 Age Specific Death Rate
 - 4.5.4 Infant Mortality Rate
- 4.6 Measurement of Population Growth
 - 4.6.1 Crude Rate of Natural Increase
 - 4.6.2 Pearle's Vital Index
 - 4.6.3 Gross Reproduction Rate
 - 4.6.4 Net Reproduction Rate
- 4.7 Summary
- 4.8 Check Your Progress
- 4.9 Further Readings

4.1 INTRODUCTION

The events happening in human's life are called *vital events* as births, deaths, migration, marriage, divorce, adoptions, separations, etc. and the methods

and techniques which used in the analysis of such data are termed as *vital statistics*. It is used by the any government for population estimation and projection and trends of population and for developing the overall planning and evaluation of economic and social development programmes. For collecting such data the registration method, census enumeration and sample survey etc. are some popular methods.

4.2 OBJECTIVES

After going through this unit you shall be able to:

- Understand the concept of demography.
- Understand measures of fertility and mortality.

4.3 RATES, RATIO AND PROPORTION

A ratio compares the extent of two quantities. When the quantities have different units, then this ratio is called a rate. Mathematically, If a and b are two quantities of the same kind (in same unit) then the fraction a/b is called the ratio of a to b = a/b or $a:b$.

A proportion is a statement of equality between two ratios. Mathematically, a, b, c and d are said to be in proportion if $a:b=c:d$

i.e. if $a/b=c/d$ i.e. if $ad=bc$

this proportion is also written as $a:b::c:d$

4.4 FERTILITY AND MEASURES OF FERTILITY

Fertility is a natural capability to deliver the children (live births). Generally this study is related with the effects of contraception, post partum abstinence and breastfeeding on fertility and the birth interval also play role in many other population measures.

Some measurements of fertility are as follows:

4.4.1 CRUDE BIRTH RATE (CBR)

It is simple, easy to calculate and easy to understand. But it is not a reliable estimate, because it considers overall population, male female in all age group which is not true. The mathematical formula is

$$CBR = \left(\frac{\text{Total number of live births}}{\text{Total population}} \right) \times 1000$$

4.4.2 GENERAL FERTILITY RATE (GFR)

It is simple, easy to calculate and easy to understand. In order to overcome of the drawback of CBR, it considers the females in the reproductive age. The

mathematical formula is

$$G.F.R. =$$

$$\left\{ \frac{\left(\begin{array}{l} \text{Total number of live births which} \\ \text{occurred among the population} \\ \text{of a region during a given year} \end{array} \right)}{\left(\begin{array}{l} \text{Mid – year female population of} \\ \text{reproductive ages 15 to 49 in the} \\ \text{given area} \\ \text{during the same year} \end{array} \right)} \right\} \times 1000$$

4.4.3 AGE SPECIFIC FERTILITY RATE (ASFR)

In order to overcome of the drawback of GFR, it considers the live births with respect to the females in the different age group of reproductive age. The mathematical formula is

$$A.S.F.R. =$$

$$\left\{ \frac{\left(\begin{array}{l} \text{Annual live births which} \\ \text{occurred to a specified age group} \\ \text{of the female population of a region} \\ \text{during a given year} \end{array} \right)}{\left(\begin{array}{l} \text{Mid – year annual female population} \\ \text{of same specified age – group} \\ \text{in the given area} \\ \text{during the same year} \end{array} \right)} \right\} \times 1000$$

4.4.4 TOTAL FERTILITY RATE (TFR)

It is the most important measurement of fertility. The total fertility rate is the sum of the age specific fertility rates from a given age to the last point of child bearing age of a female. Symbolically,

$$TFR = \sum_{15}^{49} ASFR$$

4.5 MORTALITY AND MEASUREMENTS OF MORTALITY

Population size may decrease due to death. The numbers of deaths in a given period are known as mortality. Some measurements of mortality are as follows:

4.5.1 CRUDE DEATH RATE (CDR)

It is simple, easy to calculate and easy to understand. But it is not a

reliable estimate, because it ignores age and sex of the population. Mathematically it is defined as

$$CDR = \frac{\text{Total Annual Deaths}}{\text{Total Annual Population}} \times 1000$$

The crude death rate can be calculated for males and females separately. It is usually lies between 8 and 30 per 1000. The female rate is generally lower than the male rate.

4.5.2 SPECIFIC DEATH RATE (SDR)

It is find the mortality experience in different sections of the population such as infants, mothers, male etc. The formula is:

$$SDR = \left(\frac{\text{Total deaths which occurred in among a specific group of the population of a given area during a given period}}{\text{Mid year population of the specific group of the population in the same area during the same period}} \right) \times 1000$$

4.5.3 AGE SPECIFIC DEATH RATE (ASDR)

In order to overcome of the drawback of CDR, it considers the age of the population. The mathematical defined as

$$ASDR = \left\{ \frac{\left(\text{Annual deaths which occurred to a specified age group of the population of a region during a given year} \right)}{\left(\text{Mid – year annual population of same specified age – group in the given geographical area during the same year} \right)} \right\} \times 1000$$

4.5.4 INFANT MORTALITY RATE (IMR)

It is a death rate of kids who are less than one year of age. It is an important indicator of overall physical health of community. If any community has high IMR it means there is a need of more medical care nutrition etc.

$$IMR = \frac{\text{No. of deaths under 1 year of age}}{\text{No. of live birth}} \times 1000$$

4.6 MEASUREMENT OF POPULATION GROWTH

The fertility and mortality rates separately both are not enough to give the idea about the rate of population growth. So, for obtaining the rate of growth of population these measures are useful.

4.6.1 CRUDE RATE OF NATURAL INCREASE (CRNI)

It is simple and easy to understand. Formula is

$$CRNI = CBR - CDR$$

4.6.2 CRUDE RATE OF NATURAL INCREASE (CRNI)

It is throws light on the likely growth of population. Formula is

$$PVI = \frac{CBR}{CDR} \times 100$$

4.6.3 GROSS REPRODUCTION RATE (GRR)

It is the sum of the age sex specific fertility rates (ASSFR) calculated from female births for each year of reproductive period.

$$GRR = \sum_{15}^{49} \left\{ \frac{\left(\begin{array}{c} \text{Annual live female births} \\ \text{which occurred to a} \\ \text{specified age group of} \\ \text{the female population} \end{array} \right)}{\left(\begin{array}{c} \text{population of same specified} \\ \text{Mid - year annual female} \\ \text{age group} \end{array} \right)} \right\} \times 1000$$
$$GRR = \sum_{15}^{49} ASSFR$$

The GRR is computed by the following formula also:

$$GRR = \frac{\text{No. of female births}}{\text{Total no. of births}} \times TFR$$

The value of GRR is always lies between 0 to 5.

4.6.4 GROSS REPRODUCTION RATE (GRR)

GRR excludes the effect of the mortality on the birth rate. But NRR consider the mortality also. Symbolically,

$$NRR = \sum_x \{ (female\ ASSFR) \times (survival\ factor) \} \times 1000$$

NRR lies between from 0 to 5 and $NRR \leq GRR$.

If NRR is equals to 1 the population is constant means all female babies exactly replaced her mother up to the reproductive period. If NRR is less than 1 indicates the declining population and if it is greater than 1 shows the increasing population. It is used for population projection.

4.7 MORBIDITY

Morbidity is the condition of being ill, diseased, or unhealthy. The example of an acute illness can be the flu, a broken arm, or a heart attack. Chronic illnesses are more like diabetes, or cancer. A person can live for several years with one or more morbidities which refer to having a disease or a symptom of disease, or to the amount of disease within a population. Morbidity also refers to medical problems caused by a treatment.

Example- Determination of Gross and Net Reproduction Rates for India, 1993

(1) Age in Year	(2) Age-specific fertility rate	(3) Female life-table stationary population	(4) Col. (2) x col. (3)
15-19	0.0696	4180	290.9
20-24	0.2346	4123	967.3
25-29	0.1897	4063	770.8
30-34	0.1143	4001	457.3
35-39	0.0611	3934	240.4
40-44	0.0285	3860	110.0
45-49	0.0101	3763	38.0
			2,874.7
T otal	0.7079	-	

The sex ratio at birth for the country may be supposed to be 105 males to 100 females. Hence the above table, we get

$$GRR = 5 \times 0.7079 \times \frac{100}{205} = 1.73$$

$$\text{and } NRR = \frac{2874.7}{1000} \times \frac{100}{205} = 1.40$$

Example- The procedure for computing the GRR and NRR is also shown in the following Table:

Methods for Computing GRR and NRR, U.P. 1971

Age group	Female population (1)	Live Birth (2)	ASFRs (per women) (3)=(2)/(1)	Female live birth (4)	Age specific maternity rates (5)=(4)/(1)	Age-specific survival rates* (5Lx/10) ($e_x^s=68$) (6)	Expected Female births per women (7)= (5)x(6)
15-19	79865	8813	0.1103	4293	0.0538	0.90530	0.0487
20-24	63315	7620	0.2783	8582	0.1355	0.90048	0.1220
25-29	51680		0.2812	7104	0.1370	0.89472	0.1226
30-34	44440	4585	0.2303	4986	0.1122	0.88799	0.0996
35-39	38795	4035	0.1951	3686	0.1950	0.88009	0.0836
40-44	32250		0.058	1344	0.0417	0.87013	0.0363
45-49	26720	569	0.0143	186	0.0070	0.85705	0.0060
		760					
		81					
Total	-	1963	1.1951	30181	0.5822	-	0.5188
TFR	-		5,9755	-	2.9110	-	-
GRR*	-		-	-		-	2.5940
NRR	-		-	-		-	-

Here

$$NRR = \frac{1}{10} \sum_{x=15}^{49} L_x f_x^w = (.5188) \times 5 = 2.60$$

* Probability of surviving from birth to the midpoint of the age group,

$$** GRR = TFR \times \frac{1}{1 + S.R. at birth}$$

$$= 5.9755 \times .4878 = 2.91$$

Or

$$GRR = \sum_{x=15}^{45} 5f_x^w = (.5822) \times 5 = 2.91$$

$$NRR = \sum_{x=15}^{45} 5f_x^w \times \frac{5l_x^w}{l_0} = (.5188) \times 5 = 2.91$$

4.7 SUMMERY

In this unit we studied about the vital statistics and its all measures

4.8 CHECK YOUR PROGRESS

1. The quinquennial fertility rates (computed on the basis of female births alone) for Kerala 1961 are shown in the following table, together with the survival factor for each 5 year age-group (which is the probability for a newborn female to survive till the mid point of the age group and is approximately equal to $f_5 L_x / 5^f l_0$):

Age	Fertility rate (female births)	Survival factor
15-19	0.0106	0.968
20-24	0.0660	0.968
25-29	0.0673	0.964
30-34	0.0410	0.957
35-39	0.0214	0.953
40-44	0.0065	0.944
45-49	0.0006	0.929

Compute the GRR and NRR for Kerala for 1961 on the basis of the above data.

2. The number of births occurring in Assam in 1988 is shown here classified according to age of mother, together with the female population in each age-group of the child bearing period:

Age	Female Population	Number of births to mothers in the age group
15-19	200	4,000
20-24	173	26,000
25-29	161	32,000
30-34	160	23,000
35-39	155	11,000
40-44	125	2,000
45-49	87	125
Total		

The total population of Assam in 1988 was 4,000.5 thousands.

Determine (a) the crude birth rate, (b) the general fertility rate, (C) the age specific rates and (d) the total fertility rate for 1988.

4.9 FURTHER READINGS

1. Goon A.N., Gupta M.K. & Das Gupta B Applied Statistics Vol. II The World Press Pvt. Ltd., Kolkata.
2. Gupta S.C. and Kapoor V. K., Applied Statistics, S Chand, New Delhi.
3. Ramkumar, Demography



॥ सरस्वती नः सुभगा मयस्करत् ॥

Uttar Pradesh Rajarshi Tandon
Open University

PGBCH-106

Biostatistics

BLOCK

3

TESTS OF SIGNIFICANCE

UNIT-5

Testing of Hypothesis

UNIT-6

Analysis of Variance

Course Design Committee

Dr. (Prof.) Ashutosh Gupta School of Science, UPRTOU Prayagraj	Director
Prof. Umesh Nath Tripathi Department of Chemistry Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur	Member
Prof. S.I. Rizvi Department of Biochemistry University of Allahabad, Prayagraj	Member
Prof. Dinesh Yadav Department of Biotechnology Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur	Member
Prof. Sharad Kumar Mishra Department of Biotechnology Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur	Member
Dr. Ravindra Pratap Singh Academic Consultant (Biochemistry) School of Sciences, UPRTOU Prayagraj	Member
Dr. Dharmveer Singh Academic Consultant (Biochemistry) School of Sciences, UPRTOU Prayagraj	Member/Secretary

Course Preparation Committee

Dr. Shruti Sr. Assistant Professor, School of Sciences U. P. Rajarshi Tandon Open University, Prayagraj	Writer
Prof. G. S. Pandey Department of Statistics, University of Allahabad, Prayagraj	Editor
Dr. Dharmveer Singh Academic Consultant (Biochemistry) School of Science, UPRTOU, Prayagraj	Course Coordinator

©UPRTOU, Prayagraj - 2021
ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the
Uttar Pradesh Rajarshi Tandon Open University, Prayagraj.

BLOCK INTRODUCTION

The present SLM on **Bio Statistics** consists of three Blocks. *Block - 1 – Research Methodology and Statistical Methods* has two units; *Block - 2 – Probability - Distributions and Demography* has two units; and at the last *Block – 3 Tests of Significance and Analysis of Variance* has two units.

The **Block-1– Research Methodology and Statistical Methods** consists two units. The *first unit* of this block; named *Research Methods and Sampling Procedures*; describes the meaning and types of research, significance of research. It tells about the research problem and its selection with *Sampling Theory*; which discuss about the sampling, different types of sampling designs, simple random sampling, stratified sampling and cluster sampling with their applications.. The *second unit* of this block; named *Statistical Tools*; discusses about the measures of central tendency, measures of dispersion, measures of asymmetry, correlation and regression analysis, association of attributes and 3-sigma limits.

The **Block-2– Probability- Distributions Theory and Demography** is the second block having two units. The *first unit* of this block; named *Probability and Distribution Theory*; gives the Basic concepts of probability, definitions of probability, additive and multiplicative law of probability, conditional probability, Bayes' theorem, random variable and its types, probability mass function and probability density functions. In *Probability Distributions*; the concept of probability distribution, discrete and continuous probability distributions namely Binomial Distribution, Poisson Distribution, Geometric Distribution, Normal Distribution, Exponential Distribution have been also discussed in this unit along with their properties, applications and importance.. The *second unit* of this block; named *Demography*; gives knowledge about the vital statistics and demography, this also tells about the source of vital statistics and demographic data, rates, ratio, proportion, measures of fertility, measures of mortality, measures of morbidity and migration.

The **Block-3– Tests of Significance and Analysis of Variance** consists of two units. The *first unit* of this block; named *Testing of Hypothesis*; discuss about the hypothesis and its types, level of significance, critical region, p-value, types of errors, chi-square tests, t-tests and z-tests with their applications. The *second unit* of this block; named *Analysis of Variance*; discusses about the concept of analysis of variance and co-variance, basic principles of ANOVA and ANCOVA. (One Way, Two Way and Three Way Analysis).

Illustrations and examples on these topics have also been given.

At the end of every block/unit the summary, self assessment questions and further readings are given.

UNIT-5 TESTING OF HYPOTHESIS

Structure

- 5.1 Introduction
- 5.2 Objectives
- 5.3 Hypothesis and its Type
- 5.4 Critical Region
- 5.5 Types of Error and Level of Significance
- 5.6 Test of Significance
- 5.7 Test of Significance Based on Chi-Square Test
- 5.8 Test of Significance Based on t-test
- 5.9 Test of Significance Based on z-test
- 5.10 Summary
- 5.11 Check Your Progress
- 5.12 Further Readings

5.1 INTRODUCTION

There are two major areas of statistical inference namely the estimation of parameter and the testing of hypotheses. Our present aim is to introduce to concepts involved in the development of general methods for testing of hypotheses. Some illustrations are taken from population having the common known distributions. In all the problems of statistical inference there is generalization of the results or conclusions of a sample(s) from the population to the population itself. The error is possible. We shall explain the two kinds of error in the context. In testing to hypotheses a decision is taken on the basis of a samples (s) whether to accept or to reject a specified value $H_0: \theta = \theta_0$ or a set of specified values $H_0: \theta \in \theta_0$ where $Q_0 \subset \theta$. On the basis of the results of sample (s) from the population $f(x; \theta)$, $\theta \in \theta$ where θ may be vector and θ is the parameter space of θ , for example in exponential distribution $f(x; \theta) = \theta e^{-\theta x}$, $x > 0, \theta > 0$; θ is a single parameter and but in normal density $f(x; \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

Where ,

$$-\infty < x < \infty, \sigma^2 > 0, -\infty < \mu < \infty$$

Unknown parameter μ and σ^2 are unknown the $\theta = (\mu, \sigma^2)$ and its parameter space is $\theta: (-\infty < \mu < \infty, \sigma^2 > 0)$.

In the present discussion the sample size n is considered fixed in the

advance, The case of sequential analysis, discussed by A. Wald, where are is used are not considered.

A pharmaceutical concern may be interested to find if a new drug is really effective for treatment of an element say cancer or in reducing blood pressure or inducing sleep; One may want whether a new foodstuff is really effective in increasing weight; or which of the two brands of particular product say, food stuff, fertilizers, etc. is more effective; such practical problems may be quoted, where the modern probability theory plays a vital role in decision making and the branch of statistic which helps us in arriving at a criterion for such decision known as testing of hypothesis. The theory of testing of hypothesis was first given a mathematical sound footings by J. Neyman and E.S. Pearson through series of papers. They dealt with the statistical techniques to arrive at decision in certain situations where there is an element of doubt, on the basis of a sample whose size is fixed in advance. There is another technique known as sequential testing pro-founded by Abraham Wald where the sample size is not fixed I advance but is regarded as a random variables.

5.2 OBJECTIVES

After studying this unit you shall be able to:

- Understand testing of hypothesis.
- Decide a null hypothesis and alternative hypothesis.
- Understand the meaning of level of significance, size of a test and power of a test.
- We shall define some terms associated with testing of hypothesis.

5.3 HYPOTHESES AND ITS TYPE

A statistical hypothesis is some assumption or statement about a population, or probability distribution characterizing the given population. It is frequently denoted by H.

A hypothesis is not accepted without being supported by evidence from the population. A hypothesis needs to be verified and is therefore, put to test; and based on the evidences provided by a random sample (which are set of independent observations) from the population a decisions is taken to accept or reject it. For example if r.v. $X \sim N(\mu, 25)$ then the statement that the mean of the population is greater than 20, is a statement about the population mean with known variance $\sigma^2 = 25$ and therefore is a hypothesis. We write $H: \mu > 20$.

Simple and Composite Hypotheses- A hypothesis is known as *simple hypothesis* if it completely specifies the population; otherwise it is known as a *composite hypothesis*.

For example, in sampling from a normal population $N(\mu, \sigma^2)$, the hypothesis.

(a) $H_0: \mu = \mu_0, \sigma^2 = \sigma_0^2$

is a simple hypothesis. It specifies values to both parameter μ and σ^2 ; and therefore, it completely specifies the distribution. On the other hand each of the following hypotheses is composite hypothesis:

(ii) $H_0: \mu = \mu_0$ (No statement about σ^2)

(iii) $H: \sigma^2 = \sigma_0^2$ (μ is not specified)

(i) $H: \mu < \mu_0, \sigma^2 = \sigma_0^2$

(ii) $H: \mu > \mu_0, \sigma^2 = \sigma_0^2$

(iii) $H: \mu = \mu_0, \sigma^2 > \sigma_0^2$

and so on

Null hypothesis and Alternative Hypotheses

a statistical hypothesis is said to be *Null hypothesis* which is put to test for possible rejection under the assumption that it is true; it is denoted by H_0 .

For example in sampling from normal population $N(\mu, \sigma^2)$ the hypothesis $H_0: \mu = \mu_0$ is a null hypothesis if it is to be tested. It is said to be a null hypothesis since it states that there is no difference between $H_0: \mu$ and μ_0 .

It is very important to state the alternative hypothesis H_1 explicitly in respect to any null hypothesis H_0 because the acceptance or rejection of H_0 is meaningful only if it is being tested against the rival hypothesis H_1 .

The concept of simple and composite hypothesis applies also to alternative hypothesis. For example in comparing the mean effect on the yield of soyabean of two fertilizer say A and B, we may formulate the null and alternative hypothesis as $H_0: \mu_A = \mu_B$ against $H_1: \mu_A < \mu_B$. H_0 is a simple hypothesis and H_1 is a composite hypothesis.

If we want to test the null hypothesis H_0 that the population $N(\mu, \sigma^2)$ has specified mean μ_0 let be σ^2 known then

The null hypothesis $H_0: \mu = \mu_0, \sigma^2 > 0$

and alternative hypothesis $H_0: \mu > \mu_0, \sigma^2 < 0$

5.4 CRITICAL (REJECTION) REGION

Mathematically,

Let the population be $X \sim f(x; \theta)$ $\theta \in Q$ where Q is the parameter space of the parameter θ . Let $\underline{x}: (x_1, x_2, \dots, x_n)$ be n independent sample observations corresponding to a random sample $\underline{X}: (X_1, X_2, \dots, X_n)$ of size n from the population. The n -dimensional space S which is the aggregate of all sample points $\underline{x}: (X_1, X_2, \dots, X_n)$ is called a *sample space* and is denoted by S .

The test for a hypothesis divides the whole sample S into two disjoint (mutually exclusive) regions; one region A for acceptance of hypothesis H and another region R (or C) for rejection of hypothesis H .

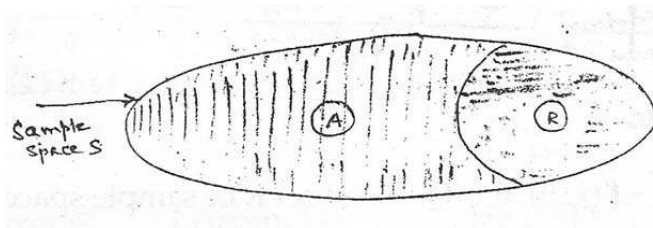


Figure 3.1 Sample Space, Acceptance (A) and Rejection (R)(or critical) region

Thus the test for hypothesis H is;

Rejected H_0 if $(x_1, x_2 \dots x_n) \in R$,

Accepted H_0 if $(x_1, x_2 \dots x_n) \in A$

A statistic associated with the test is called a test statistic.

Let $X \sim f(x; \theta)$ $\theta \in Q$. A subset R of sample space S, such that if R then H_0 is rejected (with probability 1) is called the *critical region (or rejection region)* C of the test, where

$$C = \{\underline{x} \in S: H_0 \text{ is rejected if } \underline{x} \in R\}$$

The complementary set A or R is said to *acceptance region* of the test.

5.5 TYPES OF ERROR AND LEVEL OF SIGNIFICANCE

The decision of the test for hypothesis H is taken on the basis of the information of a sample from the population. As such there is an element of risk – the risk of taking wrong decisions. In any test procedure, there are four possible mutually exclusive and exhaustive decisions:

- (i) Reject H_0 when actually H_0 is not true (false)
- (ii) Accept H_0 when it is true
- (iii) Reject H_0 when it is true
- (iv) Accept H_0 when it is false

The decisions in (i) and (ii) are correct while the decisions (iii) and (iv) are wrong decisions. These decisions may be expressed in the following dichotomous table.

True state in The nature	Decision		
		H_1 True	H_1 False
	H_0 True	Wrong (Type I error)	Correct
	H_0 False (H_1 True)	Correct	Wrong (Type II error)

Thus in testing hypothesis may lead to following two kinds of errors.

An error of type I is made if the null hypothesis H_0 is rejected when H_0 is true; and the error of Type II is made if the null hypothesis H_0 is accepted when H_0 is false.

Type I Error = [Reject H_0 | H_0 is true]

Type II Error = [Accept H_0 | H_0 is false]

= [Accept H_0 | H_0 is true]

= [Reject H_1 | H_0 is true]

The probabilities of type I and Type II errors are denoted by α and β respectively.

Definition- The size of a Type I error is the probability of type I error α similarly, the size of a type II error is the probability of type II error β

Thus, **Level of significance-**

α = Probability of Type I error

= Prob. [Reject H_0 | H_0]

The level of significance is the maximum of probability of the type I error with which one is prepared to reject H_0 when H_0 is true. It is also called the size of critical region.

5.6 TEST OF SIGNIFICANCE

Suppose that the problem is to test the hypothesis that the mean μ of the normal population $N(\mu, \sigma^2)$ with known variance σ^2 is different from μ_0 . As explained above the null hypothesis H_0 and the alternative hypothesis H_1 will be set up as follows:

$$H_0: \mu = \mu_0 \text{ against alternative } H_1 = \mu \neq \mu_0$$

A test of significance for hypothesis H_0 : is a procedure to assess the difference between the sample statistic and the value of parameter given by H_0 or differences between two independent statistic to be significant or to reject or accept H_0 at the given level of significance α .

The procedure to be adopted for test of significance is outlined below-

- Propose the null hypothesis H_0 and alternative hypothesis H_1 .
- Fix a level of significance α for the test and a sample size n .
- Then choose a statistic $T(x)$ whose sampling distribution is known under H_0 .
- Keeping the value of α in mind decide upon those values of the test statistic (i.e. rejection region) that lead to its acceptance. In other words, define the test for H_0 Vs. H_1 at level α .
- Now draw a random sample of size n from the population and compute the value of the test statistic.

- Finally on the basis of the value of the test statistic take the decision to accept or reject H_0 .

5.7 TEST OF SIGNIFICANCE BASED ON CHI-SQUARE TEST

(a) *To test the signification of variance from a normal population.*

Let $X_1, X_2, \dots, X_n$ be random sample of size n taken from $N(\mu, \sigma^2)$ and we wish to test $H_0: \sigma^2 = \sigma_0^2$.

If μ is known then H_0

$$\frac{\sum (X_i - \mu)^2}{\sigma_0^2}$$

Follows Chi-square distribution with n degree of freedom.

If μ is not known then under H_0

$$\frac{\sum (X_i - \bar{X})^2}{\sigma_0^2}$$

Follows Chi-square distribution with $(n-1)$ degree of freedom;

$\bar{X} = \frac{1}{n} \sum X$ Is the sample mean.

(b) *To Test Goodness of Fit.*

Here we test the null hypothesis that data follow a particular probability distribution (like Binomial, Poisson, Normal etc.).

Let $O_1, O_2, \dots$ be the observed frequencies and $e_1, e_2, \dots$ are the expected frequencies corresponding to these observed frequencies then

$$\sum \frac{(O_i - e_i)^2}{e_i} = \sum \left(\frac{O_i^2}{e_i} \right) - N \quad \text{where } N = \sum O_i = \sum e_i$$

follows a chi-square distribution with certain say v degrees of freedom. If expected frequencies are less than 5, then pooling is done.

Example- Five dice were thrown 192 times and the number of times 4, 5 or 6 were as follows:

No. of dice throwing 4,5,6	5	4	3	2	1	0	Total
Observed frequency: O_i	6	46	70	48	20	2	192

Calculation the value of Chi-Square on the hypothesis that dice were unbiased and hence test whether the data are consistent with the hypothesis.

Solution- Probability of throwing 4, 5, 6 is $3/6 = 1/2 = p$

Therefore from binomial distribution the theoretical frequencies of getting 5,4,3,2,1,0 successes with 5 dice re respectively the successive terms of :

$$N(p+q)^5 = 192(1/2+1/2)^5$$

Which are as follows: 6, 30, 60, 30, 66, 36 respectively

Table showing observed and expected frequency after pooling the frequency which is less than 5.

No. of dice throwing 4,5,6	5	4	3	2	1or 0	Total
Observed frequency: O_i	6	46	70	48	22	192
Expected frequency: e_i	6	30	60	60	36	192

$$\begin{aligned}
 \text{So } \chi^2 &= \sum_i \left\{ \frac{(O_i - e_i)^2}{e_i} \right\} \\
 &= \frac{(6-6)^2}{6} + \frac{(46-30)^2}{30} + \frac{(70-60)^2}{60} + \frac{(48-40)^2}{60} \\
 &\quad + \frac{(22-36)^2}{36} \\
 &= 0 + 8.53 + 1.66 + 2.4 + 5.44 = 18.03
 \end{aligned}$$

There are $n=6$ cells and $k=1$ cell, i.e., “0” is polled with cell “1”, therefore the degree of freedom of is

$$V = n-1-k, = 6-1-1=4.$$

The tabulated value of Chi-square (χ^2) on 4 degrees of freedom and at 5 % level of significance is 9.488. Since the calculated value of χ^2 is greater than the tabulated value hence null hypothesis be rejected. Therefore the observed frequency distribution is not consistent with the hypothesis.

(c) Testing of independence or Association between two (attributes) (characters):

If a character (factor, attribute) A is classified into $A_1, A_2, A_i, \dots, A_r$ classes and second character (factor/ attribute) B into $B_1, B_2, \dots, B_i, \dots, B_c$ classes and if O_{ij} is the observed frequency due to A_i class of A and B_i class B which are shown in the following table:

Character B

Character A		B_1	B_2	B_j	B_c	R_1
	A_1	O_{11}	O_{12}	O_{1j}	O_{1c}	R_1
	A_2	O_{21}	O_{22}	O_{2j}	O_{2c}	R_2
	A_i	O_{i1}	O_{i2}	O_{ij}	O_{ic}	R_i

	A _r	O _{r1}	O _{r2}	O _{tkj}	O _{rc}	R _r
	Total	C ₁	C ₂	C _j	C _{rc}	N

Where,

$$\sum_{i=1}^r R_i = \sum_{j=1}^c C_j = N \text{ grand total}$$

Then this table is called r x c contingency table.

We are interested in testing the null hypothesis.

H₀: Character (attributes) A and B are independence, i.e., there is no associated between two character A and B.

Let e_{ij} denotes the expected frequency due to ith class of A and j-th class of character B, i= 1, 2,.....r, j= 1,2,.....,c, Then

$$e_{ij} = \frac{R_i \times C_j}{N} = N \left(\frac{R_i}{N} \right) \left(\frac{C_j}{N} \right)$$

Thus, the expected frequency of any cell is equal to the product of the class totals of the two classes to which the cell belongs divided by the total number of observations.

Hence to test the above null hypothesis, we use Chi-square as given by:

$$\sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - e_{ij})^2}{e_{ij}} = \sum_i \sum_j \frac{O_i^2}{e_{ij}} - N$$

It is a Chi statistic with (r-1) (c-1) degrees of freedom.

The calculated value of is compared against the table value of χ^2 on (r-1) (c-1) degrees of freedom and 5% probability level. If calculated value of χ^2 is greater than its table value then the null hypothesis H₀ is rejected, otherwise H₀ will be accepted.

Rejecting the null hypothesis means that there is association between two factors (attributes/ character).

Example- From a village 200 persons were randomly selected and data about their income and education achievement were recorded, which are given in the following table.

Education

		High	Medium	Low	Total
Income	High	60	20	20	R ₁ =100

	Low	20	20	60	R ₂ =100
Total		e ₁ =80	e ₂ =40	e ₃ =80	N= 200

Test whether education depends upon income.

Null Hypothesis

Against the alternative

H₁; There is association between education and income.

Expected frequencies

$$e_{11} = \frac{100 \times 200}{200} = 40, \quad e_{12} = \frac{100 \times 40}{200} = 20, \quad e_{13} = \frac{100 \times 80}{200} = 40, \quad e_{21} = \frac{100 \times 80}{200} = 40, \quad e_{22} = \frac{100 \times 40}{200} = 20, \quad e_{23} = \frac{100 \times 80}{200} = 40$$

Table of expected frequencies:

Education

		High	Medium	Low	Total
Income	High	40	20	40	100
	Low	40	20	40	100
Total		80	40	80	200

$$\begin{aligned} \chi^2 &= \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{(60 - 40)^2}{40} + \frac{(20 - 20)^2}{20} + \frac{(20 - 40)^2}{40} + \frac{(20 - 40)^2}{40} + \frac{(20 - 20)^2}{20} \\ &\quad + \frac{(60 - 40)^2}{40} \\ &= \frac{20 \times 20}{40} + 200 + \frac{20 \times 20}{40} + \frac{20 \times 20}{40} + 0 + \frac{20 \times 20}{40} \\ &= 10 + 0 + 10 + 10 + 0 + 10 = 40 \end{aligned}$$

$$\text{d.f.} = (r-1)(c-1) = (2-1)(3-1) = 1 \times 2 = 2$$

The tabulated value of χ^2 and 2 d.f. and at 5% level of significance is 5.991.

Since the calculated value of χ^2 is greater than the table value of χ^2 on 2 degrees of freedom and at 5% probability level so our null hypothesis will be rejected. Therefore, it can be concluded from the above data that there between education and income.

5.8 TESTS OF SIGNIFICANCE BASED ON T-TEST

t- test is based on t- distribution and used for testing; the significance of mean from a normal population, the significant different between two population means of normal population, the significance of correlation coefficient from a bivariate normal population and for the significance of regression coefficient.

(i) Testing the significance of mean from a normal population.

Suppose $x_1, x_2, x_3, \dots, x_n$ is a random sample of size n from a normal population with mean μ and variance σ^2 (σ^2 not known) and one wishes to test. H_0 :

Under $H_0: \mu = \mu_0$

$$t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{S} \text{ follows}$$

t distribution with $(n-1)$ degrees of freedom. Here, $\bar{x}$ is the sample mean and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Where S^2 is an unbiased estimate of σ^2

The calculated of $|t|$ is compared against the tabulated value of t . If the calculated value of $|t|$ is greater than the table value of t on $(n-1)$ degrees of freedom and at α % probability level, the above null hypothesis will be rejected at α -level of significance otherwise, H_0 will be accepted.

Example- Ten rice plants were randomly selected from a small research plot having 100 plants. The height of these plants was recorded to study the effect of a bio-fertilizer on the growth behavior of plants which are given below-

Height in cm: 80,76,78,84,82,83,77,80,81,79

In the light of the above data, test whether the average height of plants in the population is 82.5 cm.

X height in cm	$X_1 - \bar{X}$	$(X_1 - \bar{X})^2$	$H_0: \mu = 82.5$ $S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ $= 60/9 = 6.666$ $S = \sqrt{6.666}$
80	0	0	
76	-4	16	
78	-2	4	
84	+4	16	
82	+2	4	
83	+3	9	
77	-3	9	

80	0	0	S= 2.58= 2.58
81	+1	1	
79	-1	1	
Total 800 = $\sum X_i$	-	60 = $\sum (X_i - \bar{X})^2$	

$$\bar{X} = \frac{\sum X_i}{n} = \frac{800}{10} = 80, \quad t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{(80-82)\sqrt{10}}{2.58} = \frac{2.5 \times 3.16}{2.58} = -3.07$$

or $|t| = 3.07$

Table value of t on 9 d.f. and at 5% level of significance is 2.262.

Since the calculated value of $|t|$ is greater than the table value of t on 9 degrees of freedom and at 5% probability level, so our null hypothesis H_0 will be rejected. Therefore it can be concluded from the given data that population mean is significantly different from = 82.5 cm., in other words the average height of plants in the population cannot be regarded as 82.5 cm from which a random sample of 10 plants have been selected with sample mean 80 cm.

(ii) Testing the significant different between two population means

Suppose $x_1, x_2, x_3, \dots, x_{n_1}$, is a random sample from 1st Normal population with mean μ_1 and variance σ_1^2 and another independent random sample $y_1, y_2, \dots, y_{n_2}$ from 2nd normal population with mean and variance. It is further assumed that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say) unknown

We wish to test the null hypothesis $H_0 = : \mu_1 = \mu_2$

Under H_0

$$\frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Follows t-distribution with $(n_1 - n_2)$ degrees of freedom. Where

$$S^2 = \frac{\sum_{i=1} (x_i - \bar{x})^2 + \sum_{i=1} (y_i - \bar{y})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Is an unbiased estimated of σ^2 , The calculated value of $|t|$ is compared against tabulated value of t on $(n_1 + n_2 - 2)$ d.f. and at $\alpha\%$ level of significance. If cal. $|t| > t_{\alpha, n_1 + n_2 - 2}$ then the null hypothesis H_0 is rejected otherwise H_0 is accepted. The above procedure is called as **two sample t-test**.

Example- Ten red plants were randomly selected from 1st plot and 8 yellow plants were randomly selected from second rose plot. The height of these selected were separately recorded and are given below in cm.

Heights of red rose plants in cm = x_i	$X_1 - \bar{X}$	$(X_1 - \bar{X})^2$	Heights of yellow rose plants in	$y_1 - \bar{y}$	$(y_1 - \bar{y})^2$
--	-----------------	---------------------	----------------------------------	-----------------	---------------------

			cm = y _i		
60	0	0	62	0	0
64	4	16	60	-2	4
61	1	1	63	+1	-
56	-4	10	64	+2	4
59	-1	1	61	-1	1
62	2	4	63	+1	1
58	-2	4	63	+1	1
60	0	0	61	-1	1
63	3	9			
57	-3	9			
$\sum X_i=600$		$\sum (X_1 - \bar{X})^2=60$	$\sum y_i=496$		$\sum (X_1 - \bar{X})^2=12$

Discuss the suggestion whether there is significant different between the mean height of red and yellow row plants.

Ho: There is no different between the mean heights of red and yellow rose plants.

$$\bar{x} = \frac{\sum x_i}{n_1} = \frac{600}{10} = 60 \quad \bar{y} = \frac{\sum y_i}{n_1} = \frac{496}{8} = 62$$

$$S^2 = \frac{\sum_{i=1} (x_1 - \bar{x})^2 + \sum_{i=1} (y_1 - \bar{y})^2}{n_1 + n_2 - 2} = \frac{60+12}{10+8-2} = \frac{72}{16} = 4.5$$

$$S^2 = 4.5, \quad S = \sqrt{4.5} = 2.12$$

Therefore,

$$t = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{60 - 62}{2.12 \sqrt{\frac{1}{8} + \frac{1}{10}}}$$

$$t = \frac{-2}{2.12 \times 3.5} = \frac{-2}{0.742} = -2.7 \quad |t| = 2.7$$

The calculated value of |t| is greater than the table value of t on 16 d.f. and at 5% probability level, so our null hypothesis is rejected. Therefore it can be concluded that our null hypothesis is rejected, hence there is significant different between the mean heights or red and yellow rose plants.

(iii) Paired t-test

Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be a random sample of size n drawn from a bivariate normal population $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ we wish to test

$$H_0: \mu_1 = \mu_2$$

In this case, under H_0 :

$$\frac{\bar{d}\sqrt{n}}{S_d}$$

Follows t-distribution with $(n-1)$ degrees of freedom.

Where

$$d_i = x_i - y_i, \bar{d} = \frac{1}{n} \sum d_i, S_d^2 = \frac{1}{n} \sum (d_i - \bar{d})^2$$

Example of paired observations-

If we want to compare the effects of two drugs D_1 and D_2 for the same disease, then first of all drug D_1 should be administered to a set of certain patients ($i=1, 2, 3, \dots, n$) and its effect should be recorded. After a reasonable interval of time, the drug D_2 should be administered to the same set of patients and its effect should be recorded. The observations so obtained are said to be paired.

Example- The scores of 10 cadets before and after training gave below:

Cadet No.	1	2	3	4	5	6	7	8	9	10
Score before training x_i	8	3	2	4	6	8	5	8	7	9
Score after training y_i	7	5	7	5	3	9	5	10	6	10

Based on the above data can you say whether training is effective in improving the performance of cadets.

Solution- Since observation under x and y are paired, so to test the null hypothesis.

H_0 : Training is not effective in improving the performance of cadets i.e.,

$H_0: \mu_1 = \mu_2$ against the alternative $H_1: \mu_1 \neq \mu_2$ we use the paired t-test and test statistic.

$$t = \frac{\bar{d}}{S_d/\sqrt{n}}$$

$d_i = x_i - y_i$	-1	2	5	1	-2	1	0	2	-1	1	$\sum d_i = 7$
d_i^2	1	4	25	1	4	1	0	4	1	1	$\sum d_i^2 = 47$

$$\bar{d} = \frac{\sum d_i}{n} = \frac{7}{10} = 0.7$$

$$S_d^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} = \frac{\sum d_i^2 - n\bar{d}^2}{n-1} = \frac{47 - 10 \times 4.9}{9} = \frac{47 - 49}{9}$$

$$= \frac{4.21}{9} = 2.15$$

Here,

$$S_d = \sqrt{4.65} = 2.15$$

We have

$$t = \frac{\bar{d}\sqrt{n}}{S_d} = \frac{0.7 \times \sqrt{10}}{2.15} = \frac{0.7 \times 3.16}{2.15} = 1.28$$

Table value of t (5%) on 9 d.f. is 2.262

The calculated value of t is less than the table value of t on 9 degrees of freedom and at 5% probability level. Therefore, null hypothesis may be accepted and it can be concluded that training is not very effective in improving the performance of cadets.

iii) Test of significance of correlation coefficient

Suppose a pair of random sample $(x_1, y_1); (x_2, y_2) \dots (x_n, y_n)$ is drawn from a bivariate normal population with population correlation p .

Let the sample correlation between x and y be r

Then to test the null hypothesis:

$H_0: p=0$ (i.e. population correlation effective is zero) against the alternative $H_1: p \neq 0$.

Under H_0 ; $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ follows a t-distribution with $(n-2)$ degrees of freedom.

Example- A random sample of size 18 form a bivariate normal population gave a correlation coefficient 0.6. Does it indicate the existence of correlation in the population?

$H_0: p=0$ Here, we wish to test $H_0: p=0$ Vs $H_1: p \neq 0$.

The value of $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.6\sqrt{18-2}}{\sqrt{1-.36}} = \frac{0.6 \times 4}{.80} = \frac{2.4}{.8} = 3.0$$

The tabulated value of t on 16 degrees of freedom at 5% level of significance is 2.12.

Since the calculated value of t is greater than the table value of t on 16 d.f. and at 5% level of significance so our null hypothesis H_0 is rejected. Therefore, it can be

concluded that the sample correlation coefficient $r=0.6$ is significant and it indicates the existence of correlation in the population.

(iv) To test the significance of Regression Coefficient:

Suppose $(x_1, y_1); (x_2, y_2) \dots (x_n, y_n)$ is a random sample from a bivariate normal population with regression coefficient β , of y and x

We know that the regression equation of y on x from the sample is: $y - \bar{y} = b_{yx}(x - \bar{x})$.

$\bar{x}, \bar{y}$ are sample means and b_{yx} is the sample regression coefficient of y or x .

So that the estimated value of y corresponding to given x_i is

$$Y_i = \bar{y} + b_{yx}(X_i - \bar{X})$$

We wish to test the null hypothesis:

$H_0: \beta = \beta_0$ where β_0 is known

Under H_0 :

$$\frac{(b_{yx} - \beta_0)\sqrt{n-2} \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}$$

Follows t - distribution with $(n-2)$ degrees of freedom.

5.9 TEST OF SIGNIFICANCE BASED ON Z-TEST

Testing significance of mean- Let $x_1, x_2, x_3, \dots, x_n$ be random sample of size n taken from $N(\mu, \sigma^2)$. The sample mean $\bar{x} = \frac{1}{n} \sum x_i$ follows normal distribution with mean μ and variance σ^2/n . For large samples, (i.e. **where sample size is more than 30**), it is true even if population is not normal. Thus,

$$Z = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma}$$

be have large samples. However, if the parent population is normal then the result is true even for small samples.

With an increased sample size, the sample variance $S^2 = \frac{1}{n} \sum (x - \bar{x})^2$ can safely be taken as an approximation to population variance $\sigma^2 =$ (if not known). Thus without any significant error approximation in large samples we have,

$$Z = \frac{(\bar{x} - \mu)\sqrt{n}}{S}$$

So if one wishes to test $H_0: \mu = \mu_0$ in case of large samples with unknown population variance one may use the test statistic

$$Z = \frac{(\bar{x} - \mu_0)\sqrt{n}}{S}$$

For known variance one uses

$$Z = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s}$$

To reject or accept the above null hypothesis, the calculated value of Z is compared with 1.96 which is table value of a standard normal variate at 5% probability level for two tail test and 1.645 for one sided test.

If the calculated value of $|Z|$ is greater than 1.96, then we reject our null hypothesis H_0 and otherwise it is accepted.

Example- A random sample of 400 male students is found to have a mean height of 168 cm. Can it be reasonable regarded as a sample from a population with mean height = 167.8cm. and standard deviation 3.25cm?

Here $\bar{x} = 168$, $\mu_0 = 167.8\text{cm}$, $\sigma = 3.25$, $n = 400$.

So

$$Z = \frac{(\bar{x} - \mu_0)}{\frac{\sigma}{\sqrt{n}}} = \frac{168 - 167.8}{\frac{3.25}{\sqrt{400}}} = \frac{0.2 \times 20}{3.25} = \frac{4.00}{3.25} = 1.23$$

The calculated value of is smaller than the table value of $Z = 1.96$ at 5% probability level, so our null hypothesis will be accepted. Hence there is no significant difference between sample mean and population mean. Therefore it can be reasonably regarded as a random sample from a population with mean 167.8cm and $\sigma = 3.25$

Testing Equality of Means- Suppose $x_1, x_2, \dots, x_n$ is an random sample from a normal population with μ_1 as mean and variance σ_1^2 and another random sample $y_1, y_2, \dots, y_{n_2}$ from second normal population with mean μ_2 and σ_2^2 . It is assumed that and are large two random samples are independent. Then we wish to test the null hypothesis. n_1 and n_2 are large and two random samples are independent. Then we wish to test the null hypothesis.

$H_0: \mu_1 = \mu_2$

Now we know that

$$\bar{x} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \bar{y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Hence,

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ is a standard normal variate (S.N.V.)}$$

So, if $H_0: \mu_1 = \mu_2$ is true then

$$Z = \frac{(\bar{x} - \bar{y})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ is a standard normal variate (S.N.V.)}$$

Hence if calculated value of $|Z|$ is greater than 1.96 which is table value of Z at 5% probability level, then our null hypothesis H_0 will be rejected, otherwise accepted. The table value of Z at 1% level of significance is 2.58.

Example- A random sample of 150 village was taken from district A and the average population per village of was found to be 440 and standard deviation 32. Another random population of 250 villages from the same district gave an averaged population 480 per village with standard

Here $n_1 = 150$ $\bar{x} = 440$ $s_1 = 32$

$n_2 = 250$ $\bar{y} = 480$ $s_2 = 56$

We want to test the null hypothesis

$H_0: \mu_1 = \mu_2$

Under H_0 ,

$$\begin{aligned} Z &= \frac{(\bar{x} - \bar{y})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(440 - 480)}{\sqrt{\frac{(32)^2}{150} + \frac{(56)^2}{250}}} \\ &= \frac{-40}{\sqrt{6.83 + 12.542}} = \frac{-40}{\sqrt{19.372}} = -\frac{40}{4.42} = -9.09 \end{aligned}$$

Hence $|Z| = 9.09$

Since the calculated value of $|Z|$ is greater than the table value of $Z = 2.58$ at 1% level of significance. Therefore the difference between the means is highly significant.

Testing Significance of Proportion- If a random sample of size n is drawn from a population with population proportion P . then we wish to test the null hypothesis

$H_0: P = P_0$ where P_0 is a particular specified value of P

Standard error of sample proportion is

$$SE(p) = \sqrt{\frac{PQ}{n}} \text{ where } Q = 1 - P$$

and n is large

Then to test the above null hypothesis, under H_0 ,

$$Z = \frac{p - p_o}{\sqrt{\frac{P_o Q_o}{n}}} \text{ is a Standard Normal Variate (SNV); } Q_o = 1 - P_o$$

if calculated value of $|Z|$ is greater than the table value of Z at 5% level of significance, our null hypothesis will be rejected, otherwise accepted. (the table value of $Z = 1.96$ at 5% level for two sided test and the table value of $Z = 1.645$ for one sided test)

Example- In a sample of 400 burners, there were 12 burners whose internal diameters were not within tolerance. Is this sufficient evidence for concluding that the manufacturing process is turning out more than 2% defective burners?

Here $P = .02$, $Q = 1 - P = 0.98$ and $p = 12/400 = .03$

The null hypothesis is:

$H_o: P = 0.02$

To test the above null hypothesis we use:

$$\begin{aligned} Z &= \frac{p - p_o}{\sqrt{\frac{P_o Q_o}{n}}} = \frac{(.03 - .02)}{\sqrt{\frac{.02 \times .98}{400}}} \\ &= \frac{0.01}{\sqrt{.000049}} = \frac{.01}{.007} \\ &= 1.429 \end{aligned}$$

The calculated value of Z is less than the table value of $Z = 1.645$ (for one tail test) at 5% probability level, so our null hypothesis is accepted. Therefore, it can be concluded that the process is under control.

Testing Equality of Proportions- Let n_1 and n_2 be two large samples taken from two different population and we wish to test $H_o: P_1 = P_2$, where P_1 and P_2 are population proportions of two quality characteristics. Suppose p_1 and p_2 be the corresponding sample proportions obtained from the two random samples drawn from these populations. If n_1 and n_2 are sufficiently large, then under H_o :

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

When $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$

However if n_1 and n_2 are moderately large with define

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \text{ and } q = 1 - p.$$

In this case under H_o ,

$$Z = \frac{p_1 - p_2}{\sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0,1)$$

Example- A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defectives in a batch of 300. Has the machine improved?

Ho: there is no difference in the improvement of the machine before and after overhauling.

Or Ho: $p_1 - p_2 = 0$

$P_1 = 20/400 = .05$ so $q_1 = 1 - p_1 = 1 - .05 = .95$

$P_2 =$ proportion of defective articles after overhauling $= 10/300 = .033$

So $q_2 = 1 - p_2 = 1 - .033 = 0.97$

$$\begin{aligned} SE(p_1 - p_2) &= \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \\ &= \sqrt{\frac{.05 \times .95}{400} + \frac{.033 \times .967}{300}} \\ &= \sqrt{0.00023} = 0.015 \end{aligned}$$

Then

$$Z = \frac{p_1 - p_2}{SE(p_1 - p_2)} = \frac{.05 - .033}{.015} = \frac{.017}{.015} = 1.134$$

The calculated value of $|Z|$ is smaller than the table value of Z at 5% level of significance, so our null hypothesis Ho is accepted. Therefore, it can be concluded that the machine has not improved after overhauling

Testing significance of standard derivation- The variance of a sample standard deviation say's is $\frac{\sigma^2}{2n}$ if a large sample of size n is drawn from a normal population of variance. It is to be noted that if parent population is not normal then this formula is not to be relied upon. Thus, for normal parent population to test Ho: $\sigma = \sigma_0$

Under Ho:

$$Z = \frac{s - \sigma_0}{\sqrt{\frac{\sigma_0^2}{2n}}} \text{ is a S.N.V.}$$

Testing Equality of Standard Deviations- Let s_1 and s_2 be the sample standard deviation of the two large samples of sizes n_1 and n_2 taken from two normal population with variance σ_1^2 and σ_2^2 respectively. For testing Ho: $\sigma_1 = \sigma_2$ under Ho, we have

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} \quad (\sigma_1, \sigma_2 \text{ known})$$

Or

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} \quad (\sigma_1, \sigma_2 \text{ not known})$$

Or

$$Z = \frac{s_1 - s_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where

$$S^2 = \frac{1}{n_1 + n_2} [n_1 s_1^2 + n_2 s_2^2]$$

as standard Normal Variate (S.N.V.).

Example- Random samples of sizes 300 and 270 taken from two normal population with variance σ_1^2 and σ_2^2 respectively gave sample standard deviations as 240 and 202.

Test the hypothesis $H_0: \sigma_1 = \sigma_2$ at 1% level of significance.

Solution- Here $H_0: \sigma_1 = \sigma_2$ Against $H_1: \sigma_1 \neq \sigma_2$

Under H_0 ,

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{240 - 202}{\sqrt{\frac{(240)^2}{2 \times 300} + \frac{(202)^2}{2 \times 270}}} = \frac{38}{13} \approx 3$$

For two sided test the tabulated value of Z at 1% level of significance is 2.58.

As calculated value falls in the critical region, H_0 may be rejected at 1% level of significance and we may conclude that the difference between the standard deviations is significant at 1% level of significance.

5.10 SUMMARY

In this unit an attempt is made to explain the basic concepts related to the testing of hypotheses.

5.11 CHECK YOUR PROGRESS

1. In a rat feeding experiment the following results were obtained gain in weight in gm.

High Protein	13	14	10	11	12	14	10	8	11	12	9	12
Low Protein	7	11	10	8	10	12	9					

Find if there is any evidence of superiority of one diet over the other.

Given t (5%) on 17 d.f. is 2.11

Ans. $t = 1.77$

2. A random sample of 900 members is found to have a mean of 3.4 cms. Could it be regarded as a sample from a large population with mean 3.25 cms? and s.d. 2.61 cm. at 5% level of significance?
3. The mean of sample of size 25 from a normal population with mean μ and s.d. 4 is found to be 15. Do you accept or reject $H_0: \mu = 20$ at the 10% level of significance?

5.12 FURTHER READINGS

1. A.M. Mood, F. Ar. Graybill & D.C. Boes. Introduction to the theory of Statistics, III. Editions Pub: Mac.Graw Hill.
2. Rohtagi V.K. (1984): An Introduction to Probability theory and Mathematical Statistics chapter VIII, IX & X Pub; John Wiley & Sons, New York.
3. Goon A.N., Gupta M.K. & Das Gupta B (1987) Fundamentals of Statistics Vol. I The World Press Pvt. Ltd., Kolkata.
4. Kapoor V.K. & S.C. Saxena: Fundamentals of Mathematical Statistics, Chapter Seventeen, Pub: S. Chand.

UNIT-6 ANALYSIS OF VARIANCE

Structure

- 6.1 Introduction
- 6.2 Objectives
- 6.3 Analysis of Variance
- 6.4 Analysis for one way Classified Data
- 6.5 Analysis for two way Classified Data
- 6.6 Analysis for three way Classified Data
- 6.7 Ancova
- 6.8 Summary
- 6.9 Check Your Progress
- 6.10 Further Readings

6.1 INTRODUCTION

Frequently biologists select their samples from numerous different populations. The differences among more than two population means cannot be tested by the methods which described earlier. Analysis of variance (ANOVA) is used for determine whether or not the means of more than two populations are equal. Analysis of variance is large area of application of statistics developed by Prof. R.A. Fisher. The term “analysis of variance” is used because the total variability in the set of data can be broken into the sum of variability among the sample means and the variability within the samples. This procedure is based on the question- is there significantly more variation among the group means than there is within the groups? The pooled variation within groups is used as a standard of comparison, because it measures the inherent observational variability in the data. The difference in means should be large relative to this inherent variability if it is to be meaningful.

6.2 OBJECTIVES

After going through this unit you should be able to

- Understand the role of analysis of variance.
- Know and apply the basic principles on experimental units.

6.3 ANALYSIS OF VARIANCE

Once the statistical problem is formulated, the next step is to perform experiment for collecting relevant information regarding the statistical hypothesis

by an appropriate method which forms the basis for drawing valid inferences in the best possible manner. After the observations are obtained they are statistically analyzed using proper techniques so as to get relevant information regarding the objective of the experiment. Usually the objective or hypotheses of the experiments are to make comparison among the effects of several treatments. If the number of treatments is more than two, then analysis of variance technique is mainly adopted for the analysis and introduction of the observations, collected from an experiment, to make comparisons among the effect of the treatments.

The total variation in any set of numerical observations may be due to a number of causes. These variations may be classified in two groups

- (i) Due to assignable causes
- (ii) Due to chance causes.

The variations due to assignable causes can be located and measured whereas the variations due to chance causes is beyond the control and therefore cannot be traced separately.

The ANOVA technique is not special to test the significance of two sample variances. Its purpose is to test the equality of several means (mean effect of several treatments) in ANOVA technique the F-test for testing the significance of two measures of variance is applied. These measures of variances differ only if means under consideration are not homogeneous.

Assumption- The analysis of variance (ANOVA) produce is based upon the following two assumptions:

- (i) The observation are independent
- (ii) Parent population from which the observations has a normal probability distribution
- (iii) Various treatment and environment effects are additive in nature.

The components variances are always shown and tested in an analysis of variance table (or ANOVA Table). The ANOVA table consist of five columns as given below, using A,B,C..... as sources of variation. It may be noted that the experimental error is also a source of variation.

Table- The Skelton ANOVA Table

Sources of Variation	Degrees of Freedom (df)	Sum of squares (SS)	Mean Sum squares (SS)	F-values
A				
B				
C				
.				
.				
Error				
Total				

The ANOVA enables us to test whether the variance due to a component factor (say A) is significantly more than the variance due to experimental error or not. Here, the mean sum of square due to component is nothing but the variance due to that component. Further, the ratio of component means square error with error mean sum of square (MSE) is distributed as F distribution with corresponding df's.

Basic Principles- We have seen the randomization, replication and local/error control are the three main principles.

The one way ANOVA is used when the field or the experimental material is homogeneous, so there is no need of local control used if there is one source of variation i.e. the treatment. In two way ANOVA, when there are two sources of variations and the three way ANOVA when there are three sources of variation are discussed. These designs take into account the variation in the experimental units. The local control principle plays an important role in these designs.

6.4 ONE WAY ANALYSIS OF VARIANCE

It is very simplest design using the two essential principles as replication and randomization and also known as completely randomized design (CRD). In CRD the units are taken in a single group, as far as possible the units forming the group should be homogeneous.

Let there be k treatment $T_1, T_2, \dots, T_k$ (or level of a factor) in an experiment. Let the i^{th} treatment be replicate r_i times for $i=1,2,3,\dots,k$. Thus the total number of experimental units required for the design is $n = \sum_{i=1}^k r_i$. In the CRD, we allocate the k treatments completely at random to the n units subject to the condition that the i^{th} treatment appears in r_i units for $i=1,2,\dots,k$. A particular case of this equal replication for different treatments, where $r_1=r_2=r_3,\dots,r_k=r$ so that $n=rk$.

Layout- The terms layout refers to the placement of treatments to the experimental units according to the condition of the design. There are several methods for random allocation of treatment to the experimental units; (i) use of random number table, (ii) Lottery system, (iii) Tossing of coin. Suppose we have four treatments A,B,C,D and we have to replicated each treatment four times then the arrangement may be

A	B	C	D
A	C	B	D
C	A	D	B
D	A	B	C

Analysis- This design provides a one way classified data according to level of a single factors treatment. For its analysis following modes is taken

$$y_{ij} = \mu + t_i + e_{ij}; \quad \begin{pmatrix} i = 1, 2, \dots, k \\ j = 1, 2, \dots, r \end{pmatrix}$$

Where y_{ij} is the random variable corresponding to the observations obtained from the j^{th} replicate of the i^{th} treatment and e_{ij} is the error component. The error components are independently normally distributed with mean zero and constant variance σ^2 , μ is the general effect, t_i is the fixed effect due to the i^{th} treatment. We are interested in hypothesis $H_0 = T_1 = T_2 = \dots = T_k$ against the alternative hypothesis that t_i are not all equal.

Let $\sum_j y_{ij} = T_i$ be the observation total of the i^{th} treatment.

$$\text{We have treatment sum of the equal} = \sum_{i=1}^k \frac{T_i^2}{r_i} - \frac{G^2}{n}$$

Where

$$G = \sum_{i=1}^k T_i, n = \sum_{i=1}^k r_i$$

Total sum of square

$$= \sum_{i=1}^k \sum_{j=1}^{r_i} y_{ij}^2 - \frac{G^2}{n}$$

Error sum of square = Total sum of square – treatment sum of squares

The analysis of variance is given below:

ANOVA for one way classified data or completely randomized design

ANOVA: Table of Completely Randomized Designs:

Sources of Variation	Degrees of Freedom (df)	Sum of squares (SS)	Mean squares (s.s./d.f)	F-values
Treatments	k-1	$\sum_{i=1}^k \frac{T_i^2}{r_i} - \frac{G^2}{n}$	S_t^2	$\frac{S_t^2}{S^2}$
Error (within treatment)	n-k	By subtraction	S^2	
Total	n-1	$\sum \sum y_{ij}^2 - \frac{G^2}{n}$		

The hypothesis that treatment have equal are tested by the F-test. If F is not significance, at desired level of significance, the treatments can be considered to have equal effects. If F is significant the treatment effects are not equal. In such

cases it becomes necessary to estimate and test individuals treatment combinations in which experimenter may be interested. Estimate of any treatment contrasts

$\sum_{i=1}^k l_i \bar{y}_i$, where t_i denotes the effect of i^{th} treatment and

Where $\bar{y}_i = \frac{T_i}{r_i}$

and

$$V \left(\sum_i l_i \bar{y}_i \right) = \sigma^2 \sum_i \frac{l_i^2}{r_i}$$

Where σ^2 is the error variance which is estimate by the error mean squares, S^2 in the above table. Significance of the contrast can be tested by t-test, where

$$t = \frac{|\sum_i l_i \bar{y}_i|}{S^2 \sum_i \frac{l_i^2}{r_i}} \text{ with } (n - k) d.f.$$

Contrasts of the type $(t_i - t_m)$ in which experiment are often interested are obtained from $\sum_i l_i t_i$ by putting $l_i = 1$, $l_m = -1$ and zero for the other l 's.

Merits and Demerits- The CRD is useful in small preliminary experiment and also in certain types of animal or laboratory experiments where the experimental units are homogeneous. There is complete flexibility in the number of treatment and the number of their replications, which may vary from treatment to treatment units or on entire treatment are missing. The CRD provides maximum degree of freedom for the estimation of the experimental error.

The main objection against the CRD is that the principle of local control had not been used in the design. It is seldom used in field experiments because the plots are not homogeneous. It is generally useful in laboratory experiments.

6.5 TWO WAY ANOVA

We have seen that in a completely randomized design no local control measure was adopted as the experimental units are supposed to be homogeneous. Usually, when experiments require a large number of experimental units, one way design cannot ensure precision of the estimates of treatment effects. An improvement of the one way design can be obtained by providing error control measures as described below. The resulting design is called two way/randomized block design.

Let there be k treatments. Each of the treatments is replicated the same number of times in this design. Let r denote the number of replications of each treatment. The total number of experimental units is, therefore, kr . These units are arranged into r blocks each of size k . The error control measures in this design consists of making the units in each of blocks homogeneous.

In clinical or similar trials where animals like rats, guinea pigs, etc. are the experimental units, animals coming from the same litter may form blocks. In

general blocks are formed with unites having common characteristics, which are known to have influence on the variate under study.

The number of blocks in the design is the same as the number of replications. The k treatments are allocated at random to the k plots in each block.

Number of treatments = k = Number of plots in a block

Number of blocks = r = Number of replication of a treatment.

This type of homogeneous grouping of the experimental units and the random allocation of the treatments separately in each block are the two main characteristic features of randomized block design. Actual number of replication in the design is determined by the availability of resources and considerations of cost and precision. A method of finding the number of replications from considerations of sensitivity of comparison of the treatments has been discussed earlier.

Analysis- The data collected from experiments with randomized block design form a two way classification, that is, classified according to the levels of two factors, viz, block and treatments. There are kr cells in the two way table with one observation in each cell.

We take the model

$$y_{ij} = \mu + t_i + e_{ij}; \quad \begin{pmatrix} i = 1, 2, \dots \dots k \\ j = 1, 2, \dots \dots r \end{pmatrix}$$

Where y_{ij} denotes the observation on the variable from ith treatment in jth block, $\mu + t_i + b_j$ are respectively the general mean, effect of the ith treatment and affect of the jth block as explained earlier. These effects are fixed and e_{ij} is the error component which is random variable. e_{ij} 's are assumed to be normally and independently distributed with zero mean and a constant variance σ^2 .

Following the method of analysis of variance for finding sums of squares due to block treatments and error we proceed as follows:

Let

$$\sum_i y_{ij} = T_i (i = 1, 2, \dots \dots k)$$

= observation total of ith treatment and

$$\sum_i y_{ij} = B_j (j = 1, 2, \dots \dots r)$$

= observation total of jth block.

These are the marginal total of the two-way data table.

Further,

$$\sum_i T_i = \sum_i B_j = G$$

We shall call G^2/rk as correction factor denoted by C.F.,

$$\text{So that } CF = \frac{G^2}{rk}$$

$$\text{Sum of squares due to treatment} = \sum_i \frac{T_i^2}{r} - C.F.$$

$$\text{Sum of squares due to blocks} = \sum_i \frac{B_i^2}{k} - C.F.$$

$$\text{Total sum of squares} = \sum_{ij} y_{ij}^2 - C.F.$$

Table: Analysis of Variance of Two way/Randomized Block Design

Sources of Variation	Degrees of Freedom (d.f.)	Sum of squares (s.s.)	Mean squares M.S.=(s.s./d.f)	F-ratio
Block	r-1	$\sum_i \frac{B_i^2}{k} - C.F.$	$S_b^2 = \frac{SSB}{r-1}$	$F = \frac{S_i^2}{S^2}$ $= \frac{MST}{MSE}$
Treatments	k-1	$\sum_i \frac{T_j^2}{k} - C.F.$	$S^2 = \frac{SST}{k-1}$ $= MST$	
Error	(r-1)(k-1)	By subtraction= SSE		
Total	Rk-1	$\sum_{ij} y_{ij}^2 - C.F.$		

The null hypothesis H_0 that the treatments have equal effects is tested by F-test.

Under H_0 , we have

$$F = \frac{MST}{MSE} \sim F_{(k-1), (r-1)(k-1)}$$

Test at level of significance α as follows:

Rejected H_0 if $F \geq F_0$

Accept H_0 if $F < F_0$

6.6 THREE WAY ANOVA

Two way designs is an improvement over one way designs in the sense that it provides error control measures for the elimination of block variations. This principle can be extended further to improve two way designs by eliminating more sources of variations. Three way/Latin square designs is one such improved

design with provision for the elimination of two sources of variation.

Let there be k treatments, each replicated k times so that the total number of experimental units required is k^2 . Let P and Q denote two factors whose variability are to be eliminated from the experimental error by having a suitable design. Evidently both these factors should be related to the variate under study so that their variability may influence the variability of the variate under study. These two are actually the controlled factors. Each of the factors P and Q is taken at k levels. The total number of level combinations of the two factors is k^2 . The k^2 experimental units are now so chosen that each unit possesses a different level combination of the two factors.

Originally Latin square designs were defined for eliminating the variation of two factors which are generally called row and column. Though it is necessary that the two factors should always be called row and column it has become customary to define latin square by calling the factors as row and column. The experimental units are obtained as specified above the treatments are allotted to these units in the following way; the k treatments are allotted to the k^2 units in such a manner that each treatment occurs only once in each level of the factor P and once in each level of the factor Q . This requires that each treatment should be replicated k times.

If a two-ways table is formed with the levels of the factors P and Q such that the levels of P denote the rows and the level of Q denote the column, then the Latin square designs requires that the treatments should be so allotted to the k^2 cells of this table that each treatment occurs once in each row and once in each column. Such an arrangement is called a latin square of order k .

For example suppose there are five treatments denoted by A, B, C, D and E . Then the following arrangements in a 5×5 square is a latin square design.

	Level of P				
Level of Q	P_1	P_2	P_3	P_4	P_5
q_1	A	B	C	D	E
q_2	D	E	A	B	C
q_3	B	C	D	E	A
q_4	E	A	B	C	D
q_5	C	D	E	A	B

It has been pointed out earlier that this design is effective if the two factors P and Q can cause variability in the variate under study. If one of the factors does not have substantial influence on the variate under study, elimination of its variance may not reduce the error variance. In such a situation a latin square design is no improvement over the randomized block designs. So, unless it is known that both the factors cause sufficient variation in the variate under study, it is better to adopt a randomized block design. In agriculture experiment if there is soil fertility

in two mutually perpendicular directors, then the adoption of a latin square design with rows and column along the director of fertility gradients proves useful.

The LSD is an incomplete three layouts, where three factors namely, row, column and treatment, are at the same number of levels (k). From the layout, it is obvious that in LSD there is need of k^2 experimental units (plots). But for a complete three ways layout with each factor at k level one needs k^3 units. Hence there is saving of $(k^3 - k^2)$ experimental units in using LSD.

Analysis of L.S.D.- In latin square designs there are three factors. These are the factors P,Q and treatments. The data collected from this design are, therefore analyzed as a three way classified data. Actually, there should have been k^3 observations as there are three factors each at k level. But because of the particular allocation of treatments to the cell, there is only one observation per cell instead of k in the usual three way classified orthogonal data. As a result we can obtain only the sums of squares due to each of the three factors and error sum of squares. None of the interaction sums of squares of the factors can be obtained accordingly, we take the model.

$$y_{ijs} = \mu + r_i + c_j + t_s + e_{ijs}$$

Where y_{ijs} denotes the observation on the variable of interest in j^{th} row j^{th} column and under s^{th} treatment $\mu, r_i, c_j, t_s (i, j, s = 1, 2, \dots, k)$ are fixed effects denoting in order the general mean, the row, the column and the treatment effects. The e_{ijs} is the error component assumed to be independently and normally distributed with zero mean and a constant variance.

The analysis is conducted by following a similar procedure as described for the analysis of two way classified data. The different sums of squares are obtained as below:

Let the data be arranged first in a row x column table such that denotes the observation of $(i,j)^{\text{th}}$ cell of the table.

Let $R_i = \sum_j y_{ij} = i^{\text{th}}$ row total ($i=1,2,\dots,k$).

$C_j = \sum_i y_{ij} = j^{\text{th}}$ column total ($j=1,2,\dots,k$).

$T_s =$ sum of those observations which come from s^{th} treatment

$=$ s^{th} treatment observation total ($s=1,2,\dots,k$).

$G = \sum_i R_i =$ grand total

Correction Factors $CF = \frac{G^2}{k^2}$

Treatment sum of squares $= \sum_i \frac{T_s^2}{k} - C.F.$

Row sum of squares $= \sum_i \frac{R_i^2}{k} - C.F.$

$$\text{Column sum of squares} = \sum_i \frac{C_j^2}{k} - C.F.$$

Table: Analysis of Variance of Completely Randomized Designs

Sources of Variation	d.f.	S.S.	M.S.= (s.s./d.f)	F
Row	k-1	$\sum_i \frac{R_i^2}{k} - C.F. = S.S.R$	S_R^2	
Column	k-1	$\sum_j \frac{C_j^2}{k} - C.F. = S.S.C$	S_C^2	
Treatments	K-1	$\sum_i \frac{T_s^2}{k^2} - C.F. = SST$	S_t^2	$\frac{S_t^2}{S_E^2}$
Error	(k-1)(k-2)	By subtraction= SSE		
Total	K ² -1	$\sum_{ij} y_{ij}^2 - C.F. = TSS$		

The hypothesis of equal treatment is tested by F-test where F is the ratio of treatment mean squares to error mean squares. If F is not significant, treatment effects do not differ significantly among themselves. If F is significant further studies to test the significance of any treatment contrast can be made in exactly the same way as discussed for randomized block designs.

6.7 SUMMERY

In this unit we have studied about ANOVA and the analysis of one way, two way and three way classified data.

6.8 CHECK YOUR PROGRESS

- Four groups of patients were subjected to the different treatments for a particular disease. At the end of a specified time period each was given a test to measure treatment effectiveness. the scores were obtained. Conduct appropriate test to find, if there is any significant difference in the treatments. Also give the ANOVA table.

A	B	C	D
60	76	58	95

88	70	74	90
70	80	66	80
80	90	60	87
76	75	82	88
70	79	75	85

Conduct appropriate test to find, if there is any significant difference in the treatment. ($F = -2.6$, $F_a = 4.26$)

- The following table gives the gains in weights of four different types of pigs fed on three different ratios. Test whether the ratios or the pig types differ in their on mean weight.

Types of pigs

Types of Rations		I	II	III	IV
	I	7	16	10	11
	II	15	14	15	14
	III	8	16	7	11

6.9 FURTHER READINGS

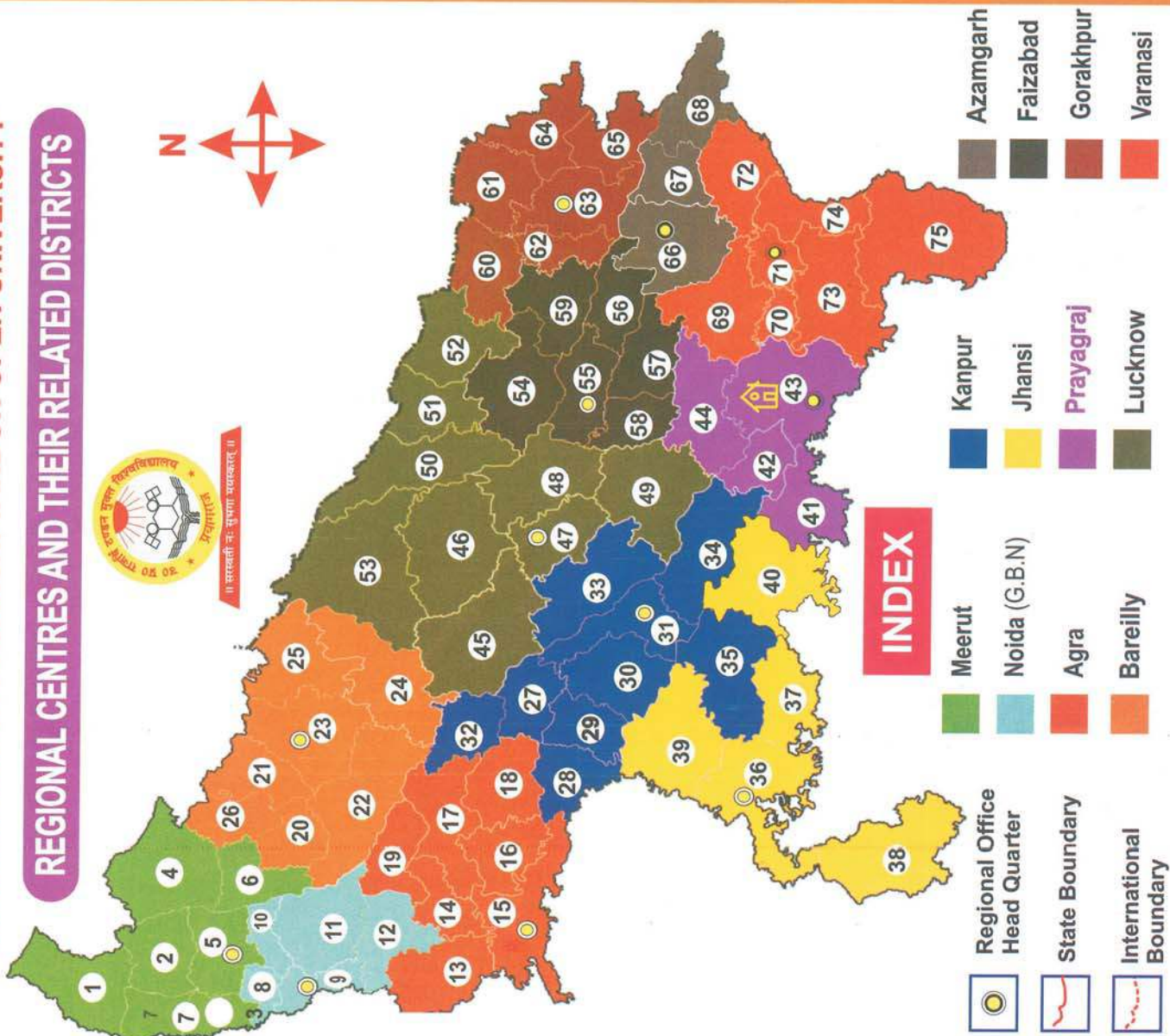
- Alok Dey (1986): Theory of Block Designs Wiley Eastern.
- Das M.N. and Giri N. (1979): Design and analysis of Experiments.
- Joshi D.D. (1987): Linear Estimation and Design of Experiments Wiley Eastern.
- Goon, Gupta & Dasgupta : Fundamentals of Statistics Vol. I and Vol. II The World Press Pvt. Ltd., Kolkata.

NOTES

DISTRICTS

1. Saharanpur	39. Jalaun
2. Muzaffarnagar	40. Banda
3. Baghpat	41. Chitrakoot
4. Bijnoor	42. Kaushambi
5. Meerut	43. Prayagraj
6. Amroha (Jyotiba Fule Nagar)	44. Pratapgarh
7. Shamli	45. Hardoi
8. Gaziabad	46. Sitapur
9. Noida (Gautam Buddha Nagar)	47. Lucknow
10. Hapur (Panchsheel Nagar)	48. Barabanki
11. Bulandshahr	49. Raebareli
12. Aligarh	50. Bahrach
13. Mathura	51. Shrivasti
14. Hathras	52. Balrampur
15. Agra	53. Lakhimpur Kheri
16. Firozabad	54. Gonda
17. Etah	55. Ayodhya
18. Mainpuri	56. Abmedkar Nagar
19. Kasganj	57. Sultanpur
20. Sambhal (Bhim Nagar)	58. Amethi (CSJ Nagar)
21. Rampur	59. Basti
22. Badaun	60. Siddharth Nagar
23. Bareilly	61. Mahrajanj
24. Shahjahanpur	62. Sant Kabir Nagar
25. Pilibhit	63. Gorakhpur
26. Moradabad	64. Kushinagar
27. Kannauj	65. Deoria
28. Etawah	66. Azamgarh
29. Auraiya	67. Mau
30. Kanpur Dehat	68. Balia
31. Kanpur Nagar	69. Jaunpur
32. Hamirpur	70. Sant Rabidas Nagar (Bhadobin)
33. Unnao	71. Varanasi
34. Fatehpur	72. Ghazipur
35. Farrukhabad	73. Mirzapur
36. Jhansi	74. Chndaluli
37. Mahoba	75. Sonbhadra
38. Lalitpur	

REGIONAL CENTRES AND THEIR RELATED DISTRICTS



सेक्टर-एफ, शान्तिपुरम, फाफामऊ, प्रयागराज-211021

“अपने भाइयों को मैं सचेत करना चाहता हूँ कि मोम न बनें और आसानी से पिघल न जायें। छोटी-छोटी सी बातों के लिए ही हम अपनी भाषा को या संस्कृति को न बदलें।”

-राजर्षि पुरुषोत्तमदास टंडन

उत्तर प्रदेश राजर्षि टण्डन मुक्त विश्वविद्यालय

प्रयागराज



॥ सरस्वती नः सुभगा मयस्करत् ॥



सेक्टर-एफ, शान्तिपुरम, फाफामऊ, प्रयागराज-211021

www.uprtou.ac.in

टोल फ्री नम्बर - 1800-120-111-333