INTRODUCTION TO STATISTICS AND PROBABILITY

Book · March 2019

CITATIONS
0

READS
6,002

11 authors, including:

Ngozi. Paulin Olewuezi
Federal University of Technology Owerri
28 PUBLICATIONS **76** CITATIONS
SEE PROFILE

Chukwudi Justin Ogbonna
Federal University of Technology Owerri
47 PUBLICATIONS **107** CITATIONS
SEE PROFILE

Ngozi Umelo-Ibemere
Federal University of Technology Owerri
8 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Hycinth C. Iwu
Federal University of Technology Owerri
15 PUBLICATIONS **31** CITATIONS
SEE PROFILE

FEDERAL UNIVERSITY OF TECHNOLOGY, OWERRI

STA 211 MANUAL

(INTRODUCTION TO STATISTICS AND PROBABILITY)

DEPARTMENT OF STATISTICS
SCHOOL OF PHYSICAL SCIENCES, FEDERAL UNIVERSITY OF TECHNOLOGY OWERRI.

ISBN: 978-978-56884-6-7

DEPARTMENT OF STATISTICS

First edition, 2019
Second edition, 2020 by Hephzibah Prints
Owerri, Imo State
Email: hephcom@gmail.com
Phone no: 08122053378; 08035809074

DEDICATION

This manual is dedicated to the families of the contributors.

EDITOR IN CHIEF

PROF. I.S IWUEZE

(Director: Academic Planning & Development, FUTO.
Lecturer, Department of Statistics, FUTO).

EDITORS

PROF. E.C. NWOGU - Department of Statistics, FUTO

PROF. (MRS) N.P. OLEWUEZI - Department of Statistics, FUTO

DR. G.U. UGWUANYIM - Department of Statistics, FUTO

CONTRIBUTORS

1. PROF. E.C. NWOGU
   Analyses of Covariance and Ratios, Rates and Index Numbers.

2. PROF. (MRS.) N.P. OLEWUEZI
   Correlation and Regression Analyses and Ratios, Rates and Index Numbers.

3. DR H.C. IWU
   Measures of Central Tendency and Partition

4. DR C.J. OGBONNA
   Analyses of Variance

5. DR G.U. UGWUANYIM
   Measures of Dispersion

6. DR C.C. NWAIGWE
   Test of Hypothesis

7. MRS. N.C. UMELO-IBEMERE
   Graphical Presentation of Data

8. MRS. B.N. OKECHUKWU
   Contigency Tables and Chi-square Analysis

9. MR. H.O. AMUJI
   Probability

10. MRS. C.H. IZUNOBI
    Measures of Central Tendency and Partition

11. MR. O. ELECHI
    Estimation & Frequency Distribution

12. MR. T.W. Owolabi
    Meaning and Scope of Statistics, Correlation & Regression Analysis

13. MR. I.C. OBINWANNE
    Meaning and Scope of Statistics

14. MR. C.P. OBITE
    Measures of Central Tendency/Partition & Frequency Distribution

15. MR. D.C. BARTHOLOMEW
    Estimation & Ratios, Rates and Index Numbers.

FORWARD TO THE FIRST EDITION

This STA 211 Manual is intended to be used in the University and other Institutions of Higher Learning for an Introductory Course in Statistics. It does not assume an elementary knowledge of Statistics and Probability. Sets of exercises are provided on each topic and it is hoped that these exercises will be useful to both teachers and students.

This First Edition covers almost all the topics in Introduction to Statistics and Probability. The manual has Eleven Chapters which consists of three (3) major sections, namely: Descriptive Statistics, Probability and Inferential Statistics. The Chapters are self-contained with sections and sub-sections. It is laid out for easy reference so that teachers and students can find quickly any topic they wish to study. The authors welcome criticisms and corrections from the users of the manual.

On behalf of Statistics Departmental Board of Studies (STADBOS) Federal University of Technology Owerri, I urge students and colleagues to appreciate and use the manual as appropriate.

Prof. I.S Iwueze

FORWARD TO THE SECOND EDITION

The first edition has been reviewed, corrections made and two new chapters (Eleven and Twelve) added. We thank the members of the public for the errors they pointed out. We hope as we earlier pointed out, that this book will not only benefit our students but the research public.

Prof. I.S Iwueze

PREFACE

A BURNING PURPOSE

The production of STA 211 Manual is as old as the Department. STA 211 (Introduction to Statistics and Probability) is a general course that cuts across every Department in Federal University of Technology, Owerri (FUTO). The idea of producing this manual was mooted in order to assist students have a mastery of the course by following up their lectures with self-study. Thus, the contributors, who are lecturers in the Department, have didactically and elementarily written out the various topics in the course outline with several examples. At the end of each chapter, questions with answers are provided. Students are required to work through these questions to enable them understand the course better.

This manual is ideal for any introductory course in statistics. It can also be very helpful to researchers.

Being the first edition of this work, it may not be error-free. Any corrections shall be made in subsequent editions.

I thank the Editor-in-Chief and his Editors for their enormous efforts in turning our dream of the production of this manual into a delightful reality. I thank also the contributors for their respective inputs. Lastly, I thank the University Administration for providing the Department with the enabling environment for its production.

Yes, 'a burning purpose' according to Edmund Burke, 'attracts others who are drawn along with it and help fulfill it'.

Dr. G. U. Ugwuanyim

Ag. Head, Statistics Department.

TABLE OF CONTENTS

CHAPTER ONE

MEANING AND SCOPE OF STATISTICS

1.0 DEFINITION

Statistics is the science of collecting, organizing, analyzing and interpreting data.
A robust definition of Statistics as a subject was given by Afonja (1975, p.11) as the study of the methods of collection and analysis of data in such a way as to minimize any uncertainty in the conclusions drawn from the data and be able to assess the degree of such uncertainty.
Statistics is divided into two categories: descriptive and inferential statistics. The methods for describing entire population through tables and diagrams are referred to as **descriptive statistics**. On the other hand, methods for studying only a part of the population through the use of mathematics and probability are generally referred to as **inferential statistics**. Inferential statistics is used to make an educated guess about a population parameter (see section 1.2) based on a statistic computed from a sample randomly drawn from that population.

Observation Unit: This is an identifiable physical entity on which measurements or observations are made e.g. students, patients, crops, machines etc.

Characteristics: These are the properties of the defined group of individuals. A characteristic of observations may be a variable or a constant. It is a constant if the characteristic remains the same for all observation units. It is a variable if it assumes different values for different observation units e.g. height is a variable. It varies from one observation to another just as colour of a person is a variable varying from one person to another. Characteristics can be measured quantitatively or differentiated qualitatively. Quantitative characteristics include heights, weights, test scores, market prices, etc. Qualitative characteristics include sex, colour, attitude, taste, etc.

Population: This is a set of all objects or (observation) units about which conclusions are to be drawn. **Sample** is a part of the population. **Target population** is the population about which information is wanted. **Sampled population** is the population from which sample is drawn. When a sample is obtained in such a way that every unit in the population has equal chance of being selected into the sample, it is referred to as a **random sample**. Population may be finite or infinite. A population is **finite** when it contains a definite number of observation units no matter how large the number. When no upper limit can be put on the number in the population it is said to be **infinite**.

Frame: This refers to the list of all the population units from which sample units are obtained.

Example1.1: All registered voters in Nigeria could be a population of voters while registered voters from Imo state are sample of voters.

1.1 STATISTICAL DATA: Data is simply a collection of facts or figures. These are observations obtained from groups of objects such as human beings, animals, crops, machines, schools etc. When these observations are obtained using statistical methods, they are called statistical data. Data may be numerical or non-numerical. When there exists a definite unit of measurement like weight, volume, length etc, the data is referred to as metric data. When each observation is classifiable into two or more categories (as in qualitative data) they are referred to as nominal data. Data that can be put in ranked form are called ordinal data. It is worth noting that though data is the plural of datum, the word is often used in the singular sense to denote a body of facts or figures.

1.1.1 TYPES OF DATA

There are two types of data obtainable in every field of endeavor; Qualitative and Quantitative.

Qualitative: Data may be a set of qualitative observations, that is, non-numerical characteristics or labels such as colour, complexion, sex, attitude of a person to some issue, marital status, etc.

Quantitative: Data may be a set of numerical measurements or quantities, such as heights, weights, examination scores, market prices, daily temperatures and many others that can be measured on an object. Quantitative data may be;

Discrete: when it assumes only zero or integer values (whole numbers) such as number of children in a family, number of students in a class, length of a queue etc.

Continuous: when it can assume any value (integer or non-integer) , such as height, weight, temperature, etc., within an interval.

1.1.2 CLASSES OF DATA

There are many ways of classifying data. A common classification is based upon who collected the data and/or the purpose for which the data was collected.

Primary data: Data collected by the investigator for a specific purpose.

Example 1.2: Going into workplaces and asking workers questions, Records taken at the production line of a manufacturing company, etc.

Secondary data: Data that has already been collected by someone else for some other purpose but being utilized by another investigator and/or for another purpose.

Example1.3: Central Bank of Nigeria (CBN) data being used to analyze the impact of inflation on the economy, etc.

1.1.3 DATA SOURCES

There are two major sources of existing data, unpublished and published.

Unpublished sources: Data in their original form exist in the files, log-books, and various registration forms of many government and non-government departments/agencies, vital registration offices of National Population Commission (NPC). Data obtained through these sources are referred to as unpublished data.

Example1.4: In Nigeria, one can obtain data on births, deaths and marriages from files of all such institutions as the law courts, churches, mosques and hospitals.

Published sources: Published data are naturally more readily accessible than unpublished ones. Main sources of published data include

1) Statistical bulletins, abstracts and reports issued by government departments.
2) Miscellaneous reports of government and non-government agencies.
3) Research reports/articles and learned journals.
4) Daily newspapers, magazines and periodicals.

Data obtained via these sources are called published data.

1.1.4 METHODS OF DATA COLLECTION

The common methods of generating statistical data include the following.

Census: A study that obtains data from every member of a well-defined population. In most studies, a census is not practical, because of the cost and/or time required. It is used when population is small and this is one of the limitations of a census study. The sense of census study in Statistics is slightly different from the sense of population census.

Example1.5: 1991, 2006 census in Nigeria.

Sample survey: A study that obtains data from a subset (part) of a population, in order to estimate population attributes (see section 1.4). This is employed when the population is large.

Example 1.6: In order to find out the opinion of FUTO students about the FUTO intra-campus shuttle service, data may be collected from census or from a randomly selected sample survey. However, there must be some rationale behind choosing the sample.

Experiment: A controlled study in which the researcher attempts to understand cause-and-effect relationships. The study is controlled in the sense that the researcher controls how subjects are assigned to groups and which treatments each group receives. Subjects can be human, animal or the environment.

Example 1.7: In mathematics, students might investigate sine waves using weights and springs. In physics, students might investigate properties of circuits, center of gravity of a meter rule using a beam. In marketing, students might examine how information about a food's health benefits affects consumer purchasing decisions. In political science, students might investigate voting behavior by participating in an election exercise etc.

Direct Observation: This attempt to understand cause-and-effect relationships like in experiments. However, the researcher is not able to control how subjects are assigned to groups and/or which treatments each group receives. In a nutshell, the researcher doesn't manipulate anything. This method draws a conclusion by comparing subjects against a control group.

Example1.8: A very simple example would be a survey of some sort. Consider someone on the busy street of an Owerri neighborhood asking random people that pass by how many

pets they have, then taking this data and using it to decide if there should be more pet food stores in that area. This is an observational study, because the researcher is simply observing the answers of the survey without influencing the outcome in any way.

Example1.9: Another example of an observational study would be if a researcher was trying to determine the effects that eating strictly organic foods has on overall health. The researcher finds 200 individuals, where 100 of them have eaten organically for the past three years, and the other 100 haven't eaten organically in the past three years. They then give each subject an overall health assessment. Lastly, they analyze the data and use it to draw conclusions on how eating organically can affect one's overall health. This is an observational study, because the researcher hasn't done anything other than observe the individuals in the study.

Questionnaire: This is a set of questions and answers. The answers to the questions constitute data. Questionnaires are used mainly in studies involving human beings and their social activities. Parties to a questionnaire includes

i. Enumerator: This name is sometimes used for the interviewer;
ii. Respondent: This is the person who answers the questions in the questionnaire;
iii. Enumeration (Observation) unit: This is an identifiable physical quantity on which observations are made. Sometimes the respondent is the enumeration unit.

Questionnaire may be administered orally or in written form. Written questionnaire may be sent by mail to the respondent (mail questionnaire) or administered directly by the enumerator (canvasser method). Both methods of administering written questionnaires have their advantages and disadvantages.

Simulation: This is done by using a statistical computer program (e.g. Minitab, SPSS, R, etc) to mimic or replicate a population data. It is often used when it is impossible to collect a real life sample data.

Example1.10: Temperature at the core of the Sun, Monte Carlo Simulations, etc.

1.2 PARAMETER AND STATISTIC

Measurements vary depending on the data set of interest, that is population or sample data sets.

Parameter: This is a numerical measurement made using the population data set describing a characteristic of a population. Thus, a parameter is a descriptive characteristic of a population. e.g. population mean μ, population variance σ^2, population correlation coefficient ρ, etc.

Statistic: This is a measurement made using a sample data set. Thus, a statistic is a descriptive characteristic of a sample (e.g. sample mean $\bar{x}$, sample variance s^2,etc) estimating a population parameter.

Parameters are sometimes difficult to measure. However, parameter measurement is possible when the population is very small otherwise it is probably impossible. This obstacle is overcome by drawing a representative sample, computing the sample statistic, and using

same to make an estimate of the population parameter. However, statistical conclusions can be up to 90% or 95% or 99% certain. They are never 100% certain.

Example1.11: FUTO has a record of staff salary. Using the staff salary data set, we could calculate the average salary for the lecturers. The average calculated from the population (all FUTO lecturers) data set would be the parameter. The average calculated from the sample of 100 lecturers would be a statistic.

1.3 LEVELS (SCALES) OF MEASUREMENT

There are four measurement scales commonly used in Statistics.

Nominal scale: Data in nominal scale is only classifiable into one of the mutually exclusive unordered classes. This can be qualitative only. Data values serve as labels, but the labels have no meaningful order.
Example1.12: blood group, breed of dog, tribe of a person, religion, sex, colour, University major, preferred brand of chocolate, etc.

Ordinal scale: Data in this scale is classifiable into one of the mutually exclusive ordered classes. This can be qualitative or quantitative. Data values serve as labels but the labels have a natural meaningful order. Differences between val ues, however, are meaningless.
Example1.13: Mathematics grade, rank, terror threat level, Satisfaction, Fanciness, etc.

Interval scale: These are always quantitative. Data are numerical, so they have a natural meaningful order, and differences between data values are meaningful. However, the ratio of two data values is meaningless. This occurs when zero is an arbitrary measurement rather than actually indicating 'nothing'.
Example1.14: temperature, year of birth, etc

Ratio scale: These are always quantitative. Data values are numerical, have order, and both differences and ratios between values are meaningful. Zero measurement indicates absence of the quantity being measured.
Example1.15: weight, height, volume, number of children in a family, etc.

1.4 SAMPLE SURVEY

Sampling is a scientific method of selecting and using a representative part (sample) of a population to make conclusions about the population. Consciously or unconsciously, we make use of sampling in everyday life to obtain the required information or carry out a course of action.

Example1.16: A Foodstuff dealer would normally use handfuls of rice, beans or *garri* from different parts of a sack of the item to check the quality of the item he is buying. A medical doctor, by using a few millilitres of a patient's blood sample obtains the quantity of malaria parasites in the blood system. Also a cook uses a small quantity of soup from a pot of soup to ascertain whether there is adequate salt in the soup.

In example1.16, an alternative method of obtaining the necessary information is to inspect the whole items. This may require the consumption of the whole pot of soup for instance, or draining out all the entire blood of the patient.

Sampling is therefore defined as the collection and examination of data from a sample in order to make inferences about the whole (Okafor 2002 p.1).

1.4.1 ADVANTAGES OF SAMPLING

In a destructive study, such as determination of the mean lifetime of electric bulbs, it is better to use sample survey rather than complete enumeration.
Sampling saves money, time and energy than in a complete enumeration where a lot of these factors are required.
Data can be collected and analyzed quickly from a sample than from the whole population.

1.4.2 METHODS OF SAMPLING

Random Sampling: The sample is chosen as a result of chance occurrences. Randomization can be achieved using a table of Random Digits or a computer program.
Example1.17: Raffle draw, telephone polling random telephone numbers, etc.

Systematic Sampling: The population is placed on a list, a random starting point is chosen and then every k-th member is selected.
Example1.18: Choosing a sample of registered voters in Imo State by choosing every 25^{th} voter from the State's voters register.
Testing every 50^{th} product from the packaging line of a manufacturing company.

Stratified Sampling: The population is divided into groups (strata) usually with meaningful differences (stratifying variable), and a sample is chosen from each group.
Example1.19: Choosing 500 men and 500women for a sample.
Stratify the population of Owerri by income level and then choose a sample of low, middle and high income individuals.

Cluster Sampling: The population is divided into groups by putting elements, which are physically close to each other together (i.e. in a more or less random way), and then a sample of these cluster units is then selected from the total number of clusters by an appropriate sampling scheme. Information is then obtained from all the units in sampled groups.

Example1.20: Randomly choose 29 polling stations in Owerri Municipal and then obtain information on all the voters at those stations.

1.5 APPLICATIONS OF STATISTICS

Statistics can be applied by Government, Companies and individuals. Government collects data on such items as population, housing, salaries and wages, agriculture, health and education. Such data reveal such information as how many people there are in each state, what services/facilities they require and how best these services/facilities can be provided.

Data are also collected on such areas as income to the Government through taxes, natural resources, import duties etc, rate of economic activities (employment and unemployment). Statistics can also be applied in research in business and industry, medicine, engineering, agriculture, sciences, law, humanities and social sciences.

EXERCISE ONE

1. In a survey conducted in Ihiagwa to find out how popular kidnapping is. You randomly choose people to call, and make 1,000 phone calls to people scattered across the community. In this study, what is the statistics term for THE PEOPLE OF IHIAGWA, and what is the statistics term for THE PEOPLE YOU CALLED?
2. When an investigator uses data, which have already been collected by others, such data is called?
3. What technique is used to ensure that a sample is representative of a population?
4. Statistical methods may be categorized into
5. A researcher is gathering data from three Senatorial zones designated
   Okigwe = 1; Orlu = 2; Owerri = 3.
   The designated Senatorial zones represent what type of data?
6. A small scale study in which all the operation intended to be used in the main study are used is called?
7. The characteristic feature possessed by the units of the population that change during the period under consideration is called
8. Memory lapse is a source of error in data collection. True/False?
9. Data on gender is categorized as
10. Indicate whether the following variables are qualitative or quantitative:
    i. Favorite food.
    ii. Favorite profession.
    iii. Number of goals scored by Man United last team season.
    iv. Number of students in FUTO.
    v. The eye color of your course mates.
    vi. IQ of Students in your hall/lodge.
11. Indicate whether the following variables are discrete or continuous:
    i. Number of yam tubers sold every day at the Eziobodo market.
    ii. Hourly temperatures recorded at FUTO observatory.
    iii. Lifetime of a car.
    iv. The diameter of the wheels of several cars.
    v. Number of children from 50 families.
    vi. Annual Census of Nigerians.
12. Classify the following variables as qualitative, quantitative discrete or continuous.
    i. The nationality of a person.
    ii. Number of liters of water contained in a tank.
    iii. Number of books on a library shelf.

iv. Sum of points tallied from a set of dice.
v. The profession of a person.
vi. The area of the different tiles on a building.

13. Analysis of labor turnover rates, performance appraisal, training programs and planning of incentives are examples of role of
A. statistics in personnel management
B. statistics in finance
C. statistics in marketing
D. statistics in production

14. Focus groups, individual respondents and panels of respondents are classified as
A. pointed data sources
B. itemized data sources
C. secondary data sources
D. primary data sources

15. Variables whose measurement is done in terms such as weight, height and length are classified as
A. continuous variables
B. measuring variables
C. flowchart variables
D. discrete variables

16. One of the following is not an example of Ordinal data
A. rank
B. volume
C. statistics grade
D. satisfaction level

17. Numerical methods and graphical methods are specialized procedures used in
A. inferential statistics
B. military statistics
C. descriptive statistics
D. education statistics

18. Scale used in statistics which provides difference of proportions as well as magnitude of differences is considered as
A. satisfactory scale
B. ratio scale
C. goodness scale
D. exponential scale

19. Sample statistics are denoted by the
A. upper case Greek letter
B. associated roman alphabets

C. roman letters
D. lower case Greek letter

20. Which of the following points do not reflect statistics?
A. It is a while subject of study
B. They can be inferential
C. It describes methods of collecting, quantitative data
D. It describes ways of analyzing qualitative data

21. Which of the following is a statistic?
A. Sample mean
B. Population variance
C. None of these
D. Population mean

22. What name is given to data which can be ranked?
A. Categorical data
B. Ordinal data
C. Interval data
D. Ratio data

23. What name is given to data which is on a continuous scale with a neutral zero?
A. Interval data
B. Ranked data
C. Ratio data
D. Ordinal data

24. What is the first stage in statistics?
A. Analyze data
B. Collect data
C. Organize data
D. Identify the group of people to be studied

25. Which of these is an example of a categorical variable?
A. flavor of soft drink ordered by each customer at a fast food restaurant
B. height, measured in inches, for each student in a class
C. points scored by each player on a team
D. color of a car

26. Numerical and pictorial information about variables are called
A. analytical statistics
B. descriptive statistics
C. inferential statistics
D. parametric statistics

27. The entire group of interest for a statistical conclusion is called the

A. data
B. population
C. sample
D. statistic

28. A subgroup that is representative of a population is called
A. a category
B. data
C. a sample
D. census

29. Statistical inference is
A. the process of estimates and conclusions carefully based on data from a sample
B. the process of estimates and conclusions carefully based on data from an entire population
C. pictorial displays that summarize data
D. tabulation of data.

30. Two types of statistical variables are
A. categorical and descriptive
B. categorical and numerical
C. descriptive and numerical
D. constant and numerical

CHAPTER TWO

FREQUENCY DISTRIBUTION

2.0 INTRODUCTION

The classification and summary of a data pertaining to any statistical enquiry is an important step in statistical computing. A statistical data consists of a list of numeric or non-numeric information related to any statistical enquiry. This information may occur a number of times. The occurrence of numeric or non-numeric information in a data set is referred to its frequency. The frequencies of variables in a statistical data are to be listed in a table in order to enhance proper classification and summary leading to easy analysis and interpretation of statistical data. This table is known as frequency distribution.

Hence, in Statistics efforts are made to categorize statistical data into distinct classes with associated frequencies which is known as the frequency distribution. At a glance, the frequency distribution provides the overview and/or summaries of the distinct values or classes of the numerical data with their corresponding frequencies. There are many types of frequency distribution but this Chapter discusses the ungrouped frequency distribution, grouped frequency distribution, cumulative frequency distribution and percentage cumulative frequency distribution, relative frequency distribution and cumulative relative frequency distribution.

2.1 UNGROUPED FREQUENCY DISTRIBUTION

When a set of data covers a few finite countable range of values or consist only discrete set of values and it is reasonable to list each value with the associated frequency, an ungrouped frequency distribution is used to provide the summary result for such set of data. Therefore, the frequency distribution for ungrouped data, referred to as regular frequency distribution, is a statistical table which provides the overview and/or summary of the distinct values of the data with their corresponding frequencies. The procedural steps for the development of an ungrouped frequency distribution are:

1. Determine the range of values of the ungrouped distribution. The least value is the first value in the first category of the frequency distribution while, the highest value is the optimum value in the last category of the frequency distribution. The incremental order of the categories does not follow any mathematical sequence or ordering but, depends on the set of the distinct values which lies between the ranges of values of the ungrouped data.
2. **Tally the** values of the ungrouped data **against** each determined category or value in 1 above in bundles of five.
3. Count the tallied strokes for each category or value and **write it as the frequency (the number of times each value occurs) which corresponds to such category.**

Example 2.1

The data on a number of pet dogs owned by households in Owerri Metropolitan is given below: 3, 0, 1, 4, 4, 1, 2, 0, 2, 2, 0, 2, 0, 1, 3, 1, 2, 1, 1and 3

Construct a frequency distribution for the ungrouped data and determine the number of households in Owerri Metropolitan that own two pet dogs.

Solution

Step 1: The range of the data is 0, 1, 2, 3, and 4. These values form the five distinct categories suitable for classification of the data.

Steps 2 and 3: Execute steps 2 and 3 respectively to obtain the frequency distribution Table 2.1.

Table 2.1: Frequency Distribution of the Number of Pet Dogs owned by Households in Owerri Metropolitan

Number of Pets (X)	Tally	Frequency (f)
0	||||	4
1	𝍸 |	6
2	𝍸	5
3	||	2
4	||	2
Total		19

The number of households in Owerri Metropolitan that own two pet dogs are five households.

2.2 GROUPED FREQUENCY DISTRIBUTION

When a set of data consists of a large number of observations or values and it is reasonable to summarize the data, a grouped frequency distribution is used to provide the summary for such set of observations. The grouped frequency distribution, referred to as continuous frequency distribution, is a statistical table which provides the overview and/or summary of the distinct classes of the numerical data with their corresponding frequencies. The distinct classes are specified with a class interval. Each class interval is specified and bounded by two values. The questions now imposed are; what is the suitable class interval and the number of classes required for proper grouping of a given set of observations? The procedural steps for formation of the class interval and the construction of the grouped frequency distribution are discussed in section 2.2.1.

2.2.1 Procedural Steps for the Determination of the Number of Classes, Class Size and the Construction of the Grouped Frequency Distribution.

Step 1: Determine the range of the data. The range is the difference between the largest and the smallest values.

Step 2: Determine the Number of Classes. A rough guide to aid in the selection of the number of classes required for data categorization was given by Sturges (1926) as

$$k = 1 + 3.322 \log_{10} n \quad (2.1)$$

where n is the number of observations, k is the number of classes required and $5 \leq k \leq 20$.

Note: when fractional result is obtained as the value of k using Equation 2.1, the nearest whole number is taken as the value of k.

Step 3: Determine the approximate class size (c). The class size (c) is obtained by dividing the range of data by the number of classes

$$\text{Class size } (c) = \frac{\text{Range}}{\text{Number of classes}} = \frac{\text{Range}}{k} \quad (2.2)$$

Note: If the set of observations are integer values, c must be an integer. If the set of observations are rounded off in one place of decimal, c must be in one place of decimal etc.

Step 4: Decide the starting point**.** Pick a starting value that is equal to the smallest value among the set of observation as the lowest class limit.

Step 5: Determine the remaining class limits by continually adding the class size to the lower class limit to obtain the rest of the lower class limits. To determine the upper class limit of the first class, subtract 1 from the lower class limit of the second class if the observations are integer valued (if the observations are rounded off in one decimal place, subtract 0.1 from lower class limit of the second class, when the observations are rounded off in two decimal places, subtract 0.01 from the lower class limit of the second class, etc.). Then add the class size to the successive upper class limits sequentially to obtain the rest of the upper class limits.

Step 6: Distribute the data into the respective classes determined in Step 5. Count the number of items in each class to obtain the frequency of each class. The total frequency of the classes must be equal to the number of observations.

Example 2.2

Construct a grouped frequency distribution for the data on IQ scores for gifted pupils in a classroom of a particular elementary school. The IQ scores are: 118, 123, 124, 125, 127, 128, 129, 130, 130, 133, 136, 138, 141, 142, 149, 150 and 154.

Solution

n = 17

Range = 154 – 118 = 36

Number of classes (k) required $= 1 + 3.322 \log_{10} 17 = 5.08755 \approx 5$ classes

Class size (c) = $\frac{36}{5}$ = 7.2 or 8 (approximate to the next whole number)

Starting point (least IQ score) = 118

Table 2.2: Grouped Frequency Distribution of IQ scores for 17 gifted pupils in a classroom of a particular elementary school

Classes	Tally	Frequency (f)
118-125	||||	4
126-133	𝍸 |	6
134-141	|||	3
142-149	||	2
150-157	||	2
Total		17

Example 2.3

Construct a grouped frequency distribution for the set of 50 measurements on a particular item.

2.5, 5.9, 3.2, 1.4, 7.0, 4.3, 8.9, 0.7, 4.2, 9.7, 3.4, 4.6, 5.0, 6.4, 1.1, 9.2, 7.7, 0.9, 4.0, 2.3, 5.6, 2.2, 3.1, 4.7, 5.5, 6.6, 1.9, 3.9, 6.1, 5.2, 8.2, 3.3, 2.2, 5.8, 4.1, 3.8, 1.2, 6.8, 9.5, 0.8, 0.1, 4.8, 0.4, 0.3, 6.1, 6.6, 1.5, 1.3, 0.8, 0.9

Solution:

n = 50

Range = 9.7 – 0.1 = 9.6

Number of classes (k) required $=1+3.322\log_{10}50 = 6.64397 \approx 7$ classes

Class size = $\frac{9.6}{7} = 1.37 \approx 1.4$ (since the observations are in one place of decimal)

Starting point = 0.1

Table 2.3: Grouped Frequency Distribution for the set of fiftymeasurements on a particular item

Classes	Tally	Frequency (f)
0.1-1.4	𝍸 ||	12
1.5-2.8	𝍸 |	6
2.9-4.2	𝍸 ||||	9
4.3 – 5.6	𝍸 |||	8
5.7 – 7.0	𝍸 |||	8
7.1 – 8.4	|||	3
8.5 – 9.8	||||	4
Total		50

2.2.2 Relationship between Class Interval, Class Limit, Class Boundaries and Class Mark

Class interval, class limit, class boundaries and class mark are four important but related concepts in statistics. As discussed in Section 2.2.1, class interval refers to the length of class specified by the range between two values or numbers. The two values or numbers are referred to as the class limits. The smallest value is the lower class limit (denoted LCL)

while the upper value is the upper class limit (denoted UCL). Given the frequency distribution of the IQ scores data of Example 2.2 shown in Table 2.2, the first class interval is 118-225. The value 118 is the lower class limit (least value belonging to the first class) while 125 is the upper class limit (highest value belonging to the first class).

Class boundaries (lower class boundary (denoted as LCB) and upper class boundary (denoted as UCB)) are achieved by adding a shift index value to the upper class limits and subtracting same shift index value from the lower class limits of the classes. Specifically the class boundaries are determined as follows: adding and subtracting a shift index value of 0.5 from the upper and lower class limits respectively for class limits with whole or integer values, adding and subtracting a shift index value of 0.05 from the upper and lower class limits respectively for class limits with a decimal point rounded off values, adding and subtracting a shift index value of 0.005 from the upper and lower class limits respectively for class limits with two decimal points rounded off values etc.

Class mark is the mid-point of the class interval. The class mark is given as

$$\text{Class mark} = \frac{\text{LCL} + \text{UCL}}{2} \qquad 2.3$$

Example 2.4

Determine the class boundaries and class mark of the grouped frequency distribution of IQ scores for 17 gifted pupils in a classroom of a particular elementary school shown in Table 2.2.

Solution

Table 2.4: Class Boundaries and Class Mark of the Grouped Frequency Distribution of IQ scores for 17 gifted pupils in a classroom of a particular elementary school

Classes LCL - UCL	Class Boundaries LCB - UCB	Class mark X	Frequency (f)
118-125	117.5 - 125.5	121.5	4
126-133	125.5 - 133.5	129.5	6
134-141	133.5 - 141.5	137.5	3
142-149	141.5 - 149.5	145.5	2
150-157	149.5 - 157.5	153.5	2
Total			17

2.3 CUMULATIVE FREQUENCY AND PERCENTAGE CUMULATIVE FREQUENCY DISTRIBUTIONS

Cumulative frequency distribution is one of the important types of frequency distribution. Cumulative frequency can be defined as the sum of all previous frequencies up to the current point. The cumulative frequency distribution is formed in two ways by either cumulating the frequencies from the first to the last class or from the last class to the first class. The former

is called the "Less Than" Cumulative Frequency Distribution" while the latter is called the "More Than" Cumulative Frequency Distribution"

The cumulative percentages are formed from the frequency distribution by converting the cumulative frequencies into percentages. The resulting values which indicate the percentage of values that are accumulated as you move down or up the distribution table is called the percentage cumulative frequency distribution. It gives the percentage of values at or below each i^{th} class value, i = 1, 2, … , k.

Cumulative frequency is used to determine the number of observations that lie above (or below) a particular value in a data set. It is also used to find the median and some other measures of fractiles.

Example 2.5

Calculate the cumulative frequency and the percentage cumulative frequency distribution of the grouped frequency distribution of the IQ scores for 17 gifted pupils in a classroom of a particular elementary school shown in Table 2.2.

Solution

Table 2.5: The "Less Than" Cumulative Frequency and Percentage Cumulative Frequency of the IQ scores for 17 gifted pupils in a classroom of a particular elementary school

Classes	Frequency (f)	"Less Than" Cumulative Frequency	"Less Than" Percentage Cumulative Frequency
118-125	4	4	(4/17) x 100 = 23.53
126-133	6	4+6 = 10	(10/17) x 100 = 58.82
134-141	3	10+3 = 13	(13/17) x 100 = 76.47
142-149	2	13+2 = 15	(15/17) x 100 = 88.24
150-157	2	15+2 = 17	(17/17) x 100 = 100

Table 2.6: The "More Than" Cumulative Frequency and Percentage Cumulative Frequency of the IQ scores for 17 gifted pupils in a classroom of a particular elementary school

Classes	Frequency (f)	"More Than" Cumulative Frequency	"More Than" Percentage Cumulative Frequency
118-125	4	13+4 = 17	(17/17) x 100 = 100
126-133	6	7+6 = 13	(13/17) x 100 = 76.47
134-141	3	4+3 = 7	(7/17) x 100 = 41.18
142-149	2	2+2 = 4	(4/17) x 100 = 23.53
150-157	2	2	(2/17) x 100 = 11.77

2.4 Relative Frequency and Cumulative Relative Frequency Distribution

A relative frequency is the ratio (fraction or proportion) of the frequencies of the distribution. The relative frequencies are obtained by dividing each frequency by the total frequency and they can be written as fractions, percents, or decimals. Relative frequency is used to define probability. Especially, the value of the relative frequency for each class may be used to determine the chance or proportion or percentage of the total observations falling into a distinct class.

The cumulative relative frequency is the accumulation of the previous relative frequencies. It is obtained by adding all the previous relative frequencies to the relative frequency of a particular class.

Example 2.6

Calculate the relative frequencies and cumulative relative frequencies of the grouped frequency distribution of IQ scores for 17 gifted pupils in a classroom of a particular elementary school shown in Table 2.2.

Solution

Table 2.7: Relative Frequencies and Cumulative Relative Frequencies of the IQ scores for 17 gifted pupils in a classroom of a particular elementary school

Classes	Frequency (f)	Relative Frequency	Cumulative Relative Frequency
118-125	4	4/17 = 0.235	0.235
126-133	6	6/17 = 0.353	0.235 + 0.353 = 0.588
134-141	3	3/17 = 0.176	0.588 + 0.176 = 0.764
142-149	2	2/17 = 0.118	0.764 + 0.118 = 0.882
150-157	2	2/17 = 0.118	0.882 + 0.118 = 1
Total	17		

Note: The value of the relative frequency for the first class equals 0.235 implies that about 24% of the IQ scores for all gifted pupils in a classroom of a particular elementary school examined lie between 118 – 125.

EXERCISE TWO

1. Twenty students were asked how many hours they studied per day. Their responses, in hours, are as follows: 5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3. Construct an ungrouped frequency distribution table showing the number of hours studied per day and the cumulative frequency.

Use the information below to answer question 2, 3 and 4

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown in Table 2.8 below:

Table 2.8: Part-time students' Course Table.

No. of courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

Part-time Student Course Loads

2. Fill in the blanks in Table 2.8 above.
3. What percent of students take exactly two courses?
4. What percent of students take one or two courses?

Use the information below to answer question 5, 6 and 7

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in Table 2.9 below.

Table 2.9: Gum Disease Report

No. of Flossing per week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.45	
1	18		
3			0.9333
6	3	0.05	
7	1	0.0167	

Flossing Frequency for Adults with Gum Disease

5. Fill in the blanks in Table 2.9.
6. What percent of adults flossed six times per week?
7. What percent flossed at most three times per week?

8. Thirty immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2, 5, 7, 2, 2, 10, 20, 15, 3, 7, 20, 5, 12, 15, 12, 4, 5, 10, 4, 11, 18, 24, 25, 21, 4, 6, 21, 18, 17and 15. Construct a grouped frequency distribution table showing the class mark and class boundary.

Using the information contained below, answer questions 9, 10 and 11.
18 friends wrote an online exam and had the following scores14, 14, 13, 15, 11, 15, 13, 10, 12, 13, 14, 13, 14, 15, 17, 14, 14 and 15

9. Find the relative frequency of those that score 12.
10. Calculate the cumulative frequency for those that score of 14?
11 What is the most common score in the data?
12. A political pollster asked voters to indicate the degree of their agreement with the following question: "Global warming is real and is at least partially caused by human activity." The voters response were scaled using the following psychometric scales: 1 = strongly disagree, 2 = disagree, 3 = Neither disagree nor agree, 4 = agree and 5 = strongly agree. The Frequency distribution, Percentages and Cumulative Percentages of Response of Voters indicating their degree of agreement to the question on Global Warming are presented in Table 2.10 below.

Table 2.10: Frequency, Percentages and Cumulative Percentages of Response of Voters indicating their degree of agreement with Global Warming

Psychometric Scale	Frequency f	Percentages %	Cumulative Frequency cf
1.00	5	21.7	21.7
2.00	7	30.4	52.2
3.00	1	4.3	56.5
4.00	4	17.4	73.9
5.00	6	26.1	100.0
Total	23	100	-

a. How many people responded to the survey?
b. How many people strongly agreed with the statement that Global warming is real and is at least partially caused by human activity ?
c. What percentage of people strongly disagreed with the statement that Global warming is real and is at least partially caused by human activity?
d. What percentage of the sample disagreed with the statement that Global warming is real and is at least partially caused by human activity (i.e., gave an answer of 2 or lower)?

13. The number of classes required to partition a given distribution on students score in a test is nine (9). Determine the total number of students that participated in the test using sturge's rule?

14. The range of a class is twice the class size. If there are X^2 number of classes required for proper grouping of the class, determine the value of X^2. (A) 8
(B) 6 (C) 4 (D) 2

Using the information contained below, answer Questions 15, 16 and 17.

The upper class limit of a class is five times its lower limit X. If the class mark of the class is 30.

15. Determine the upper class limit of the class?

(A) 80.0 (B) 94.0 (C) 50.0 (D) 32.0

16. Determine the lower class limit of the class?

(A) 16.0 (B) 18.8 (C) 10.0 (D) 16.4

17. What is the class size?

Using the information contained below, answer questions 18 - 23.

The distribution shown in Table 2.11 represents the distance (Km) covered by some selected students from their Home to School.

Table 2.11: Distance covered by some students

Serial Number	Distance (Km)	Frequency
1	10-14	3
2	15-19	12
3	20-24	17
4	25-29	9
5	30-34	4
6	35-39	1

18. Determine the upper class boundary of the third class?
(A) 20.5 (B) 24.5 (C) 22.0 (D) 22.0

19. Determine the lower class boundary of the fifth class?
(A) 29.5 (B) 34.5 (C) 34.5 (D) 32.0

20. Determine the class mark of the second class?
(A) 10 (B) 19 (C) 15 (D) 17

21. Determine the cumulative frequency of the third class

22. What is the relative frequency of the sixth class?

23. Determine the percentage cumulative frequency of the fourth class

Using the information contained below, answer questions 24 and 25

The frequency distribution of weights (in kg) of 40 persons is given in Table 2.12 below

Table 2.12: Frequency Distribution of the Weight of 40 persons

Weights (in kg)	30 – 34	35 –39	40 – 44	45 - 49	50 – 54
Frequency	6	13	14	4	3

24 (a) what is the lower class limit of fourth class?
(b) What is the class size of each class?

25 (a)which class interval has the highest frequency?
(b) Find the class marks of all the class intervals?

26. Construct the frequency distribution table for the data on heights (cm) of 20 boys using the class intervals 130 - 134, 135 - 139 and so on. The heights of the boys in cm are: 140, 138, 133, 148, 160, 153, 131, 146, 134, 136, 149, 141, 155, 149, 165, 142, 144, 147, 138, 139. Also, find the range of heights of the boys.

Using the information contained below, answer questions 27, 28, 29, 30, 31, 32, 33 and 34.

The scores of 30 students in FUTO PUTME is given below
31, 41, 46, 33, 44, 51, 56, 63, 71, 71, 62, 63, 54, 53, 51, 43, 36, 38, 54, 56, 66, 71, 74, 75, 46, 47, 59, 60, 61, 63.

27. Construct a frequency distribution table for the following scores of 30 students using 30-34, 35 – 39, …
28. What is the range of the above scores?
29. How many classes are there?
30. What is the class mark of the class intervals 50-54?
31. Which class has the highest frequency?
32. Determine the cumulative frequency of the third class
33. What is the relative frequency of the sixth class?
34. Determine the percentage cumulative frequency of the fourth class

CHAPTER THREE

GRAPHICAL PRESENTATION OF DATA

3.0 INTRODUCTION

Graphs are used to (i) summarize and bring outstanding features of a set of data readily to the eye. (ii) suggest possible methods of further analysis (iii) explain conclusions based on the data. Some of the graphs most commonly used to illustrate a set of data are: (a) line graph (b) bar chart (simple and multiple) (c) component graph (pie chart and block diagram) (d) curves (e) histogram (f) frequency polygon (g) cumulative frequency graph (ogive). Graphs (a) – (c) are used for both qualitative and quantitative data while (d) – (g) are used only for quantitative data.

3.1 BAR CHART

This consists of separated vertical bars (rectangles) of equal width in which the height of each bar is proportional to the frequency of the class it represents. We have basically three types of bar charts, namely: Simple Bar Chart, Multiple Bar Chart and Component Bar Chart

3.1.1. Simple Bar Chart: Rectangular bars are used to represent items or class intervals. The bars have equal width and the height of each bar is proportional to the frequency of the item. There are also equal gaps between the bars.

Example 3.1: Draw a bar chart given the number of births in Hospital A.

Table 3.1 Number of births by year in Hospital A

Year	2000	2001	2002	2003	2004
Number of birth	15	23	9	13	32

Solution: The simple bar chart for the number of births by year only is shown in Figure 3.1.

3.1.2 Multiple Bar Chart: This is a chart in which two or more bars lie side by side to each other. Two or more bars are drawn for each item. The height of each bar is also proportional to the frequency and equal gaps must be between bars.

Example 3.2: The table that follows shows the number of births in Hospitals A and B. Draw a multiple bar chart to compare the number of births by year and hospitals.

Table 3.2 Number of births by year in two Hospitals A and B

year		2000	2001	2002	2003	2004
Number of birth	A	15	25	9	13	32
	B	12	30	15	7	45

Solution: The multiple bar chart for comparing the number of births by year in hospitals A and B is shown in Figure 3.2.

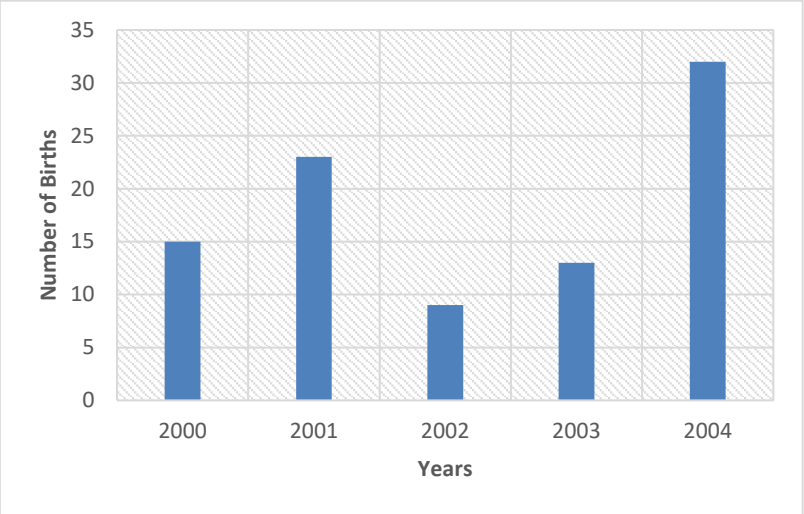


Figure 3.1: Simple bar chart of the number of births in hospital A

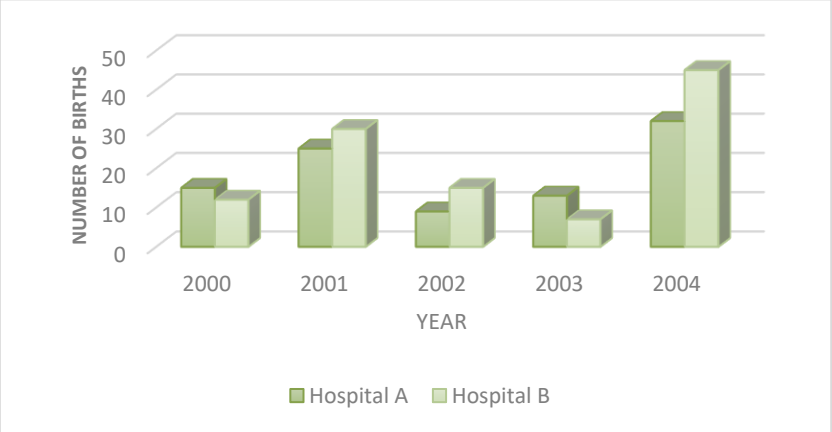


Figure 3.2: Multiple bar chart for the number of births in hospitals A and B

3.1.3. Component Bar Graph (Block Diagram): This is a special bar chart in which each bar is sub-divided into components. It can also be referred to as Block Diagram. To draw a component bar chart, find the total of the frequencies for each item. The height of each bar is proportional to the totals and the component bars are shaded differently.

Example 3.3. The data below shows the monetary assets to the nearest billion. Construct a component bar chart to graphically represent the data.

Table 3.3 Monetary Assets in nearest billion

	Year			
Monetary Assets	1999	2000	2001	2002
Domestic Credit	34	50	65	81
Private Sector	60	70	83	90
Commercial Banks	20	25	52	70
Total	114	145	200	241

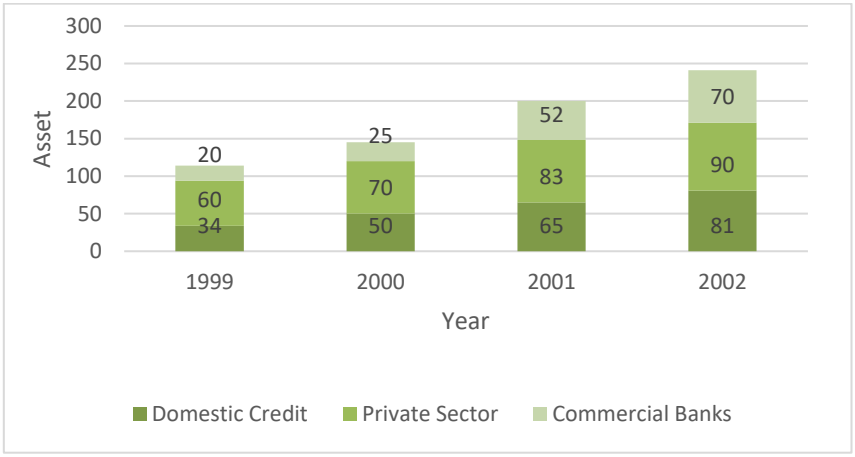


Figure 3.3: Component Bar Chart of Monetary Assets by year.

3.2 PIE CHART:

In a pie chart, each item is represented as a sector of a circle. The frequency of each item is converted to relative frequency. Then draw a circle and measure each sector angle

$$\text{Relative frequency} = \frac{frequency\ of\ item}{sum\ of\ all\ frequencies} \tag{3.1}$$

$$\text{sector angle} = relative\ frequency\ \times 360°. \tag{3.2}$$

Example 3.4: Draw a pie chart for the number of birth in Hospital A using Table 3.4

Table 3.4: Number of births in Hospital A

year	Number of birth	Relative frequency	Sector Angle (°)
2000	15	15/92	58.70
2001	23	23/92	90.00
2002	9	9/92	35.22
2003	13	13/92	50.87
2004	32	32/92	125.21
Total	92	1.0	360

Solution

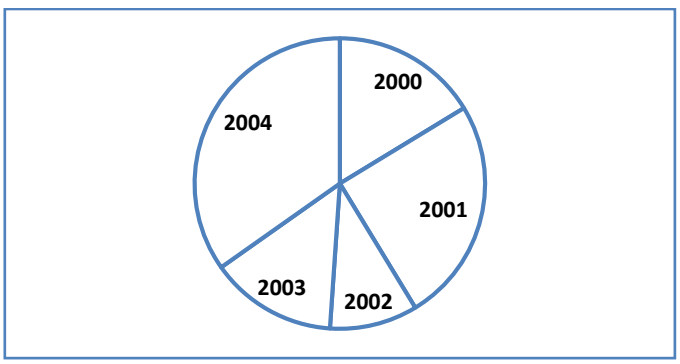


Figure 3.4: Pie chart for the number of births in Hospital A

3.3. HISTOGRAM

To draw a histogram, rectangular bars are mounted on class boundaries. The area of each bar is proportional to the frequency of the class interval and there is no gap between bars.

Area= class width x height (h)

$$\text{Height} = \frac{\text{Relative frequency}}{\text{Class width}} \tag{3.3}$$

For example: In the first class of Example 3.5,

$$\text{Area}, A = 5 = 10 \times h$$

$$\Rightarrow h = \frac{5}{10} = 0.5$$

The total area under a histogram is equal to the total frequency. To plot a histogram, the class height is plotted against class boundaries. When the class size is equal, the frequency is plotted against class boundaries. The histogram may be plotted based on either: (i) the absolute values or (ii) the relative values of the frequency. When it is based on the relative values, the total area under a histogram is equal to one. To construct a histogram when the class width is not equal: (i) find the class width of each class (ii) obtain the class height (iii) plot height on class boundaries.

Example 3.5: Present the data that follows using a histogram.

Class	Frequency	Class boundary
0 - 9	5	-0.5 – 9.5
10 - 19	10	9.5 – 19.5
20 - 29	15	19.5 – 29.5
30 - 39	10	29.5 – 39.5
40 - 49	5	39.5 – 49.5

Solution

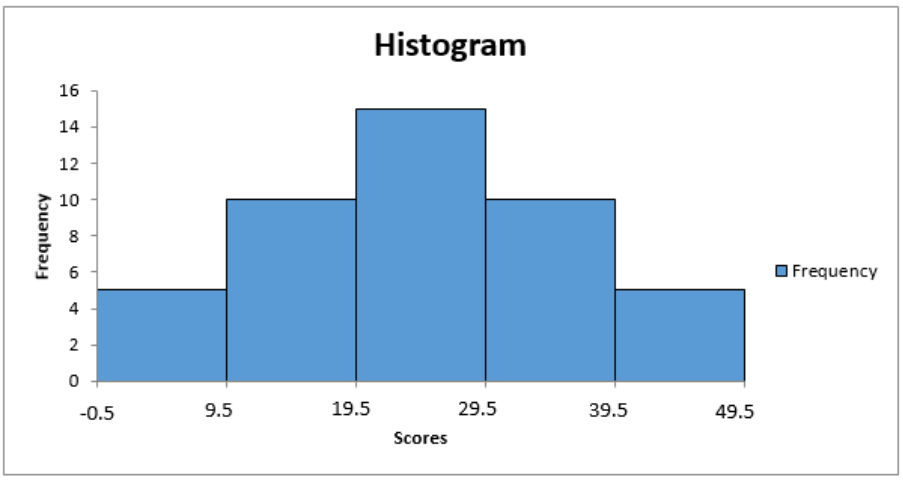


Figure 3.5: Histogram for Example 3.5

3.4 FREQUENCY POLYGON

This graph is obtained by plotting height (h) against the class mark, or simply join the top mid points of the bars in the histogram.

Example 3.6 Represent the data in Table 3.6 using a frequency polygon

Table 3.6: Scores of 24 students in PHY 101 examination

Class	*Frequency*	*Class mark*
61-63	*3*	*62*
64-66	*5*	*65*
67-69	*6*	*68*
70-72	*7*	*71*
73-75	*3*	*74*

Solution

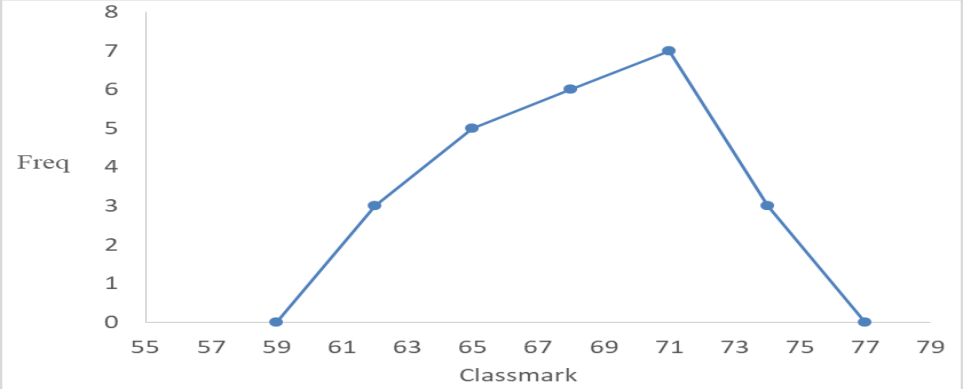


Figure 3.6: Frequency Polygon Curve for Example 3.6

3.5 CUMULATIVE FREQUENCY GRAPH (OGIVE)

An ogive is produced by plotting cumulative frequency against (a) the upper class boundaries if cumulative is from the lowest to the highest values (b) lowest class boundaries if cumulative is from the highest to the lowest values. However, the most commonly used is the first (a) and has been illustrated below. Here, the cumulative frequencies give the number of observations not exceeding (at most) the upper class boundaries. For example, the number of observations not exceeding 29.45 is 24. To ensure that the graph is stepped down to the horizontal axis. We introduce an upper class boundary of the lowest side with zero cumulative frequency. An ogive is used to estimate the number of observations that are less than or equal to a particular value. It is used to estimate the quartiles and percentiles.

Example 3.7: Draw a cumulative frequency polygon for the frequency distribution that follows.

Class	Frequency	Cumulative frequency	Class boundaries
9.5 – 19.4	2	2	9.45 – 19.45
19.5 – 29.4	22	24	19.45 – 29.45
29.5 – 39.4	31	55	29.45 – 39.45
39.5 – 49.4	2	57	39.45 – 49.45
49.5 – 59.4	3	60	49.45 – 59.45

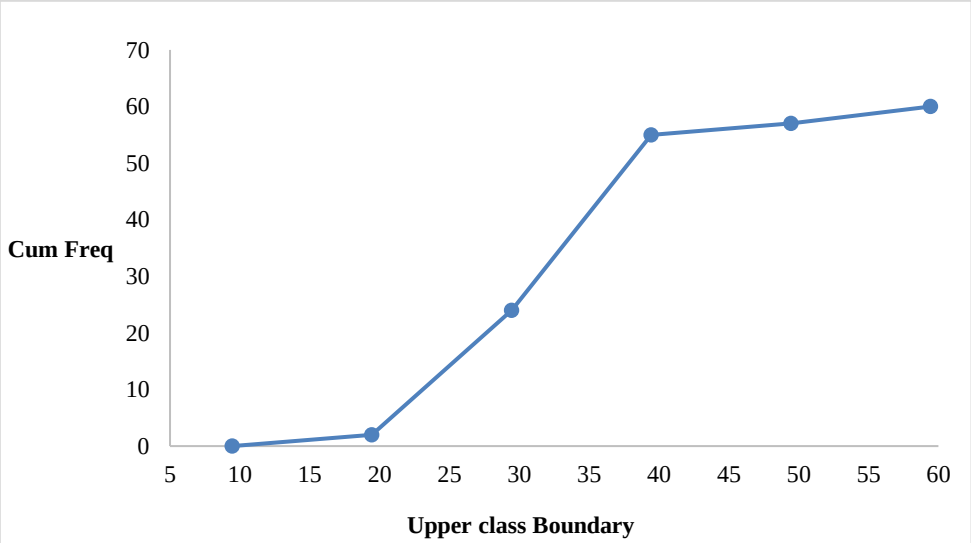


Figure 3.7: Cumulative Frequency Polygon curve of Example 3.7

3.6 STEM-AND-LEAF PLOT (STEM PLOT)

For the stem-and-leaf plot, each observation is regarded as consisting two pieces: stem and leaf. To construct the stem plot, (i) list all the leading digits (stem) in the first column preferably in ascending order and the remaining digit (s) become the leaf (ii) list all possible stem values in a vertical column and write the leaf for every observation beside the corresponding stem value (iii) order the leaves (iv) provide a key that explains in context what the stems and leaves represent.

Example 3.8: Suppose we want to present the exam scores of 20 students in STA 211 using the stem plot.

Solution:

78 47 90 75 60 100 81 69 54 41 73
59 78 51 80 60 42 73 65 64

4	1 2 7
5	1 4 9
6	0 0 4 5 9
7	3 3 5 8 8
8	0 1
9	0
10	0

Key: 5|4 represent score 54

EXERCISE THREE

1. 20 women were asked if their marital status is Single (S), Married (M), Widowed (W) or Divorced (D). The responses are as follows:
   M, S, S, M, D, W, S, M, D, M, S, D, S, M, M, S, S,D, S, M.
   Represent the data using a bar chart.

2 State the difference between the block diagram and the multiple bar chart

3 List three features of a bar chart.

4 List two ways used to display qualitative data graphically.

5 In a bar chart, the height of each bar is proportional to the ---------- of the item.

6 The data represent an online survey that asked adults the type of phone they like to use;

Nokia	800
Tecno	255
Samsung	586
Motorolla	99
LG	120

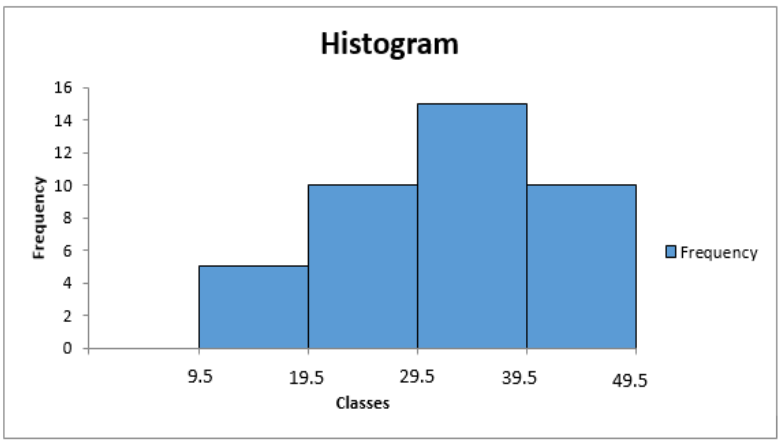
(a) Use a pie chart to display the data.
(b) What percentage of adults use Samsung phones?

7. In a histogram, the area of each bar is proportional to the --------- of a class.

8. ---------- are used to ensure that consecutive bars of a histogram touch

9. Use the histogram in the figure that follows to:

(a) determine the number of classes
(b) determine the class width
(c) estimate the frequency of the modal class



10. Construct a histogram using the frequency table that follows;

class	frequency
9.5 – 19.4	2
19.5 – 29.4	12
29.5 – 39.4	31
39.5 – 49.4	2
49.5 – 59.4	3

11 20 undergraduates were asked the number of times they call their parents in a week. The following data show their responses.

10 6 7 9 3 2 1 6 12 11
4 8 5 7 7 5 10 5 15 8

Draw a histogram for the data. (Hint : Use 1 as the lower class limit of the first class and 3 as the class width).

12 To draw a frequency polygon, frequency is plotted on ---------

13 List three ways used to display quantitative data graphically.

14 In an opinion poll, 118 adults were asked questions on their highest educational qualification.

Others	6
HND	36
B.Sc.	48
M.Sc.	20
Ph.D	8

Use a pie chart to display the data.

15. What advantage does the stem-and-leaf plot have over grouped frequency distribution?

16. The stem-and-leaf plot has _____ and _____ major parts.

17. (a) List the actual values of the data displayed in the stem-and-leaf plot below.

1 | 1 3 6
2 | 7
3 | 5
4 | 1 1 4 5 7 7 8

key 1|1 = 1.1

(b). What is the maximum value?

18. The data that follow give the Grade Point Averages of a sample of Statistics students

1.60 1.72 2.51 4.74 3.25 2.35 3.56 3.70 0.81 0.91

Draw a stem-and-leaf plot for the data.

19. These data give the time (in minutes) taken by 12 students to move from their hostels to the lecture hall;

30 15 25 28 32 17 10 15 9 12 30 28

Display the data in a stem-and-leaf plot

20. The following data show the incomes (in thousands of naira) of a sample of 10 households.

25 40 22 100 80 17 35 20 62 50

Construct a Stem Plot for the data.

21. State one advantage of an Ogive.

22. Draw a stem-and-leaf plot for the following data

60 52 78 60 63 60 47 81 42 50 73
90 45 57 39 70

23. ABC Technology manufactures computer keyboard. The following are the numbers of keyboards produced at the company for a sample of 15 days

42 34 38 35 38 36 31 38 33 39 45
41 33
45 37

Prepare a stem plot.

A random sample of 15 students were selected and their scores in Biology test are given below.

0	3 4 6 9
1	0 1 2 5 5 8
2	1 5 6 7
3	0

key 1|1 = 11

Use the data to answer questions 24 and 25

24. Write the data that was used to construct the display
25. Obtain the highest score
26. State the differences between a histogram and a simple bar chart.
27. Which graph would you choose to display the percentage of 6 nutrients in a food item?

A. Stem plot B. Pie Chart C. Multiple bar chart D. Histogram

28. State one difference between a frequency histogram and a relative frequency histogram.

CHAPTER FOUR

MEASURES OF CENTRAL TENDENCY AND PARTITION

4.0 MEASURES OF CENTRAL TENDENCY

A measure of central tendency is a summary statistic that represents the center point of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. In statistics, the three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.

Choosing the best measure of central tendency depends on the type of data you have. In this chapter, the following measures of location will be considered:

a) The Arithmetic mean
b) The geometric mean
c) The harmonic mean
d) The median
e) The mode

Quantiles/Fractiles are partition values of the variate which divides the total frequency into a number of equal parts. Mostly used quantiles in the analysis of data are (i) Quartiles (ii) Deciles and (iii) percentiles.

4.1 Arithmetic Mean

This is the most common measure of central location. It is defined as the sum of n numbers divided by n. The mean of a set of n observations, $x_1, x_2, x_3, x_4, ..., x_n$ is denoted as $\bar{x}$ and it is given by

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (4.1)$$

If the numbers occur with frequencies, the mean is calculated by multiplying each number by corresponding frequency. This is given by

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{n} \qquad (4.2)$$

where f_i is the respective frequency, k is the number of groups and $n = \sum_{i=1}^{k} f_i$

Example 4.1

The number of books ordered by 11 salesmen are 10,9,6,5,7,12,6,8,8,4 and 3. Find the mean order.

Solution

Let x_i be the number of books ordered by the ith salesman

The mean is calculated as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{11}(10+9+6+5+7+12+6+8+8+4+3) = \frac{78}{11} = 7.09 \approx 7.$$

The mean order is 7.

There are different types of mean, namely: arithmetic mean, weighted mean (WM), geometric mean (GM) and harmonic mean (HM). If mentioned without an adjective (as mean), it generally refers to the arithmetic mean.

4.2 Geometric Mean (GM)

GM is the nth root of the product of the data values $x_1, x_2, ..., x_n$. This measure is valid only for data that are measured absolutely on a strictly positive scale. The geometric mean can be good for combining numbers that are expressed in different units

$$GM = \sqrt[n]{x_1 * x_2 * x_3 * ... * x_n} \tag{4.3}$$

4.3 Harmonic Mean (HM)

HM is the reciprocal of the arithmetic mean of the reciprocals of the data values. This measure too is valid only for data that are measured absolutely on a strictly positive scale. It is appropriate for situations when the average of rates is desired.

$$\text{HM} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{4.4}$$

4.4 Weighted Arithmetic Mean (WM)

WM is an arithmetic mean that incorporates weights to certain data elements. Weighted average is used whenever some data points are worth more than others and when each data point shows a different probability of occurring.

$$\text{WM} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{4.5}$$

where w_i is the weight

Example 4.2

The score of 5students in STA 211 test are 4, 6, 8, 5 and 7. Find

a. The arithmetic mean

b. The geometric mean
c. The harmonic mean

Solution

Let x_i be the score of the students

a. $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{4+6+8+5+7}{5} = \frac{30}{5} = 6$

The arithmetic mean is 6

b. $GM = \sqrt[n]{x_1 * x_2 * x_3 * ... * x_n} = \sqrt[5]{4 \text{ x } 6 \text{ x } 8 \text{ x } 5 \text{ x } 7} = \sqrt[5]{6720} \approx 5.83$

The geometric mean is 5.83

c. $\text{HM} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{5}{\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\frac{1}{5}+\frac{1}{7}} = \frac{5}{0.8845} \approx 5.65$

The harmonic mean is 5.65

Example 4.3

Here is how a teacher may decide to grade his class. 10% for homework assignments, 15% for quizzes, 25% for midterm test, and 50% for the final exam. A student scores 45 in homework assignment, 50 in quiz, 65 in midterm test and 95 in exam. What is the weighted mean?

Solution

Let x_i be the student's score and w_i be the weight

$$\text{WM} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{(0.1*45)+(0.15*50)+(0.25*65)+(0.5*95)}{0.1+0.15+0.25+0.5} = \frac{75.75}{1.00} = 75.75$$

The weighted mean is 75.75

4.5 MEAN FOR GROUPED DATA

When data is grouped, x_i becomes the midpoint of each class and the midpoint is the average of the lower class limit and the upper class limit. The mean for grouped data can be obtained using the Long method, coding method I and coding method II. The formulae for the three methods are given as follows:

4.5.1 The Long Method

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{n} \tag{4.6}$$

where x_i is the respective midpoint (class mark), f_i is the respective frequency and k is the number of groups

Example 4.4

The daily temperature observed in a given community is given below. Obtain the mean using the Long method.

Class	Frequency
20-29	5
30-39	7
40-49	8
50-59	6
60-69	7
70-79	9
80-89	8

Solution

The summary of the table is given below

Class	Midpoint (x)	f	fx
20-29	24.5	5	122.5
30-39	34.5	7	241.5
40-49	44.5	8	356
50-59	54.5	6	327
60-69	64.5	7	451.5
70-79	74.5	9	670.5
80-89	84.5	8	676
Total		**50**	**2,845**

The Long Method:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{n} = \frac{2845}{50} = 56.9$$

When the numbers are very large, one's calculators may not be able to carry them. Coding methods I and II are then used to reduce the numbers in order to make the calculations less cumbersome.

4.5.2 The Coding Method I

$$\bar{x} = A + \frac{\sum_{i=1}^{k} f_i d_i}{n} \tag{4.7}$$

where A is a constant (assumed mean) and d_i is the difference between each observation and the assumed mean.

$d_i = x_i - A;\ x_i = A + d_i$

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i (A + d_i)}{\sum_{i=1}^{k} f_i} = \frac{A\sum_{i=1}^{k} f_i + \sum_{i=1}^{k} f_i d_i}{\sum_{i=1}^{k} f_i} = A + \frac{\sum_{i=1}^{k} f_i d_i}{\sum_{i=1}^{k} f_i}$$

where n = $\sum_{i=1}^{k} f_i$

Example 4.5

Obtain the mean of Example 4.4 using the coding method I (Use A = 54.5)

Solution

The summary of the table is given below

Class	Midpoint (x)	f	d = x –A	*fd*
20-29	24.5	5	-30	-150
30-39	34.5	7	-20	-140
40-49	44.5	8	-10	-80
50-59	54.5	6	0	0
60-69	64.5	7	10	70
70-79	74.5	9	20	180
80-89	84.5	8	30	240
Total		50		120

The Coding Method I:

$$\bar{x} = A + \frac{\sum_{i=1}^{k} f_i d_i}{n} = 54.5 + \left(\frac{120}{50}\right) = 54.5 + 2.5 = 56.9$$

4.5.3 The Coding Method II

$$\bar{x} = A + \left(\frac{\sum_{i=1}^{k} f_i u_i}{n}\right) C \tag{4.8}$$

where f_i = frequency of each class

A= assumed mean; $u_i = \frac{d_i}{c}$; $d_i = x_i - A$; $x_i = cu_i + A$

x_i = Midpoint of each class; c = class size

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i (A + cu_i)}{\sum_{i=1}^{k} f_i}$$

$$= \frac{A\sum_{i=1}^{k} f_i + c\sum_{i=1}^{k} f_i u_i}{\sum_{i=1}^{k} f_i}$$

$$= A + \frac{c\sum_{i=1}^{k} f_i u_i}{\sum_{i=1}^{k} f_i}$$

$$= A + \frac{c\sum_{i=1}^{k} f_i u_i}{n};$$

where $n = \sum_{i=1}^{k} f_i$

Example 4.6

Obtain the mean of Example 4.4 using the coding method II (Let A = 54.5)

Solution

The summary of the table is given below:

Class	Midpoint (x)	f	d = x – A	$u = \frac{d}{c}$	fu
20-29	24.5	5	-30	-3	-15
30-39	34.5	7	-20	-2	-14
40-49	44.5	8	-10	-1	--8
50-59	54.5	6	0	0	0
60-69	64.5	7	10	1	7
70-79	74.5	9	20	2	18
80-89	84.5	8	30	3	24
Total		50			12

A = 54.5; c = 10

Using the Coding Method II

$$\bar{x} = A + \frac{c\sum_{i=1}^{k} f_i u_i}{n} = 54.5 + \left(\frac{12}{50}\right)10 = 54.5 + 2.4 = 56.9$$

The advantages of mean are as follows:

a. It is easy to calculate and understand.
b. The calculation utilizes the whole data
c. It is a more representative of the distribution than any other measure of location.

Limitations of the mean:

a. The mean cannot be calculated for categorical data, as the values cannot be summed.
b. As the mean includes every value in the distribution the mean is influenced by outliers – an outlier is an observation point that is distant from other observations.

4.6 THE MEDIAN

The median is the middle score for a set of data that has been arranged in ascending or descending order. It is easy to calculate the median. If the number of observations is odd, then $(\frac{n+1}{2})^{th}$ observation (in the ordered set) is the median. When the total number of observations is even, it is given by the mean of $\frac{n}{2}$th and $(\frac{n}{2}+1)^{th}$ observation.

4.6.1 The Median for Grouped Data

Estimated Median $(\tilde{x}) = L_m + \left[\frac{\left(\frac{N}{2}\right) - Cf}{f_m}\right] * w$ (4.9)

where:

- **L_m** is the lower class boundary of the group containing the median
- **N** is the total number of values
- **Cf** is the cumulative frequency of the groups before the median group
- **f_m** is the frequency of the median group
- **w** is the group width of the median class

Example 4.7

Find the median of the grouped data given below

IQ	Frequency (f)
118-125	4
126-133	6
134-141	3
142-149	2
150-157	2

Solution

IQ	Frequency (f)	Class Boundaries	Cumulative Frequency
118-125	4	117.5-125.5	4
126-133	6	125.5-133.5	10
134-141	3	133.5-141.5	13
142-149	2	141.5-149.5	15
150-157	2	149.5-157.5	17
Total	17		

Estimated Median ($\tilde{x}$) = $L_m + \left[\frac{\left(\frac{N}{2}\right) - Cf}{f_m}\right] * w$; $\frac{\mathbf{N}}{\mathbf{2}} = \frac{\mathbf{17}}{\mathbf{2}} = 8.5$,

where the 8.5th IQ falls into the 126-133 class

Cf = 4, f_m = 6, w = 133.5 – 125.5 = 8 (width of the median class), L_m = 125.5

Median = $125.5 + \left[\frac{8.5-4}{6}\right] * 8 = 131.5$

Advantages of the median

1. The median is less affected by outliers and skewed data.
2. The median can be used for ordinal and numeric data

Limitation of the median:

1. The median cannot be identified for nominal data, as it cannot be logically ordered.

4.7 THE MODE

The mode is defined as the value of a variable or attribute which occurs with the most frequency. The mode can be applied to both quantitative and qualitative data. A distribution may contain more than one mode and such a distribution is said to be multi-modal. If it contains only two modes, it is called bimodal and one modal distribution is called uni-modal.

Example 4.8

Find the mode of the following set of data.

a) 15,19,16,21,18,24,14,19
b) 2, 4, 2, 7, 4, 5, 1, 3, 6, 8.
c) 24,14,16,21,19,18

Solution

a) 14, 15, 16, 18, 19, 19, 21, 24.
   The mode is 19. This is called uni-modal
b) 1, 2, 2, 3, 4, 4, 5, 6, 7, 8.
   The mode is 2 and 4. This is bimodal.
c) 14, 16, 18, 19, 21, 24. No mode.

4.7.1 The Mode for Grouped Data

$$\text{Mode} = L_m + \left(\frac{f_m - f_{m-1}}{2f_m - f_{m+1} - f_{m\_1}}\right) * w \qquad (4.10)$$

where:

- L_m is the lower class boundary of the modal group

- f_{m-1} is the frequency of the group before the modal group
- f_m is the frequency of the modal group
- f_{m+1} is the frequency of the group after the modal group
- w is the modal group width

Example 4.9

Given the table below, find the mode of the IQ score

IQ	Frequency (f)
118-125	4
126-133	6
134-141	3
142-149	2
150-157	2

Solution

IQ	Frequency (f)	Class Boundaries
118-125	4	117.5-125.5
126-133	6	125.5-133.5
134-141	3	133.5-141.5
142-149	2	141.5-149.5
150-157	2	149.5-157.5
Total	17	

Mode = $L_m + \left(\frac{f_m - f_{m-1}}{2f_m - f_{m+1} - f_{m\_1}}\right)$ * w; Modal class = 126-133 because it has the highest frequency (6); L_m = 125.5, f_m = 6, f_{m-1} = 4, f_{m+1} = 3, w = 8

Mode = $125.5 + \left(\frac{6-4}{(2*6)-3-4}\right) * 8 = 128.7$

Advantages of the mode

1. The mode has an advantage over the median and the mean as it can be used for both nominal, ordinal and numeric data

Limitations of the mode:

1. In some distributions, the mode may not reflect the center of the distribution very well.
2. It is also possible for there to be more than one mode for the same distribution of data (bi-modal, or multi-modal). The presence of more than one mode can limit the ability of the mode in describing the center or typical value of the distribution because a single value to describe the center cannot be identified.
3. In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).

Example 4.10

Given the following set of numbers 2, 4, 2, 4, 4, 5, 1, 3, 6, 6.

a. Find the mean
b. Find the median
c. Find the mode

Solution

a. Let x be the set of numbers

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+4+2+4+4+5+1+3+6+6}{10} = 3.7$$

Or

Using $\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$

Numbers (x)	Frequency (f)	fx
1	1	1
2	2	4
3	1	3
4	3	12
5	1	5
6	2	12
Total	10	37

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{37}{10} = 3.7$$

b. Arrange the numbers in either ascending or descending order

1, 2, 2, 3, 4, 4, 4, 5, 6, 6; $\frac{n}{2} = \frac{10}{2} = 5th$ number and $\frac{n}{2}+1 = \frac{10}{2}+1 = 6$th number

Median $= \frac{4+4}{2} = 4$

b. Mode = 4

4.8 FRACTILES OR MEASURES OF PARTITION

Fractiles are measures of location or position which include not only central location but also any position based on the number of equal divisions in a given distribution. The quartiles, deciles and percentiles are measures of partition or fractiles, which partitions the data into four, ten and hundred equal parts respectively.

4.8.1 The Quartiles

Q**uartiles** are the values that divide a data set into four equal parts when the data are arranged in order of magnitude. To find the quartiles, we arrange the data set in order of magnitude. We have the first, second and third quartiles which are denoted by Q_1, Q_2, Q_3 respectively. Q_1 is referred to as the lower quartile while Q_2 is the median and Q_3 is the upper quartile.

When data are ungrouped, the position of the quartiles are

$$Q_1 \ \ is\ the \left(\frac{N+1}{4}\right)th\ \ item \tag{4.11}$$

$$Q_2 \ \ is\ the\ 2\left(\frac{N+1}{4}\right)th\ \ item \tag{4.12}$$

$$Q_3 \ \ is\ the \ \ 3\left(\frac{N+1}{4}\right)th\ \ item \tag{4.13}$$

For grouped data, **the quartile group** is found using

$$\left(\frac{N}{4}\right)th\ item\ \ for\ the\ Q_1\ \ group \tag{4.14}$$

$$\left(\frac{2N}{4}\right)th\ item\ \ for\ the\ Q_2\ \ group \tag{4.15}$$

and

$$\left(\frac{3N}{4}\right)th\ item\ \ for\ the\ Q_3\ \ group \tag{4.16}$$

The **actual value of the quartiles** is computed from grouped data using the formula

$$Q_k = L_k + \left[\frac{\left(\frac{kN}{4}\right) - Cf}{f_k}\right] * w \qquad , \ k = 1, 2, 3 \tag{4.17}$$

where:
L_k is the lower class boundary of the quartile group
N is the total number of values
Cf is the cumulative frequency of the groups before the quartile group
f_k is the frequency of the quartile group
w is the class width or class size of the quartile group

Example 4.11
The scores of student in Sta211 assignment are 5, 7, 4, 4, 6, 2 and 8. Find the quartiles

Solution
Arrange the numbers in ascending order

2, 4, 4, 5, 6, 7, 8

N = 7

The first quartile

$$Q_1 \ is\ the \left(\frac{N+1}{4}\right)th\ item;\ Q_1 = \left(\frac{7+1}{4}\right)th\ item.$$

$Q_1 = 2^{nd}$ item

$Q_1 = 4$

The second quartile

$$Q_2 \ is\ the\ 2\left(\frac{N+1}{4}\right)th\ item;\ Q_2 = 2\left(\frac{7+1}{4}\right)th\ item$$

$$Q_2 = 4th\ item$$

$Q_2 = 5$

The third quartile

$$Q_3 \ is\ the\ 3\left(\frac{N+1}{4}\right)th\ item;\ Q_3 = 3\left(\frac{7+1}{4}\right)th\ item$$

$$Q_3 = 6^{th}\ item$$

$Q_3 = 7$

4.8.2 The Deciles

Deciles divide the distribution into 10 equal parts. There are 9 points which can be computed. These nine points are denoted as $D_1, D_2, \ldots, D_9$

When data are ungrouped, the position of the deciles are

$$D_1 \ is\ the \left(\frac{N+1}{10}\right)th\ item \quad (4.18)$$

$$D_2 \ is\ the\ 2\left(\frac{N+1}{10}\right)th\ item \quad (4.19)$$

and

$$D_9 \ is\ the\ 9\left(\frac{N+1}{10}\right)th\ item \quad (4.20)$$

For grouped data, **the decile group** is found using

$$\left(\frac{N}{10}\right)^{th} \text{ for the } D_1 \text{ group} \quad (4.21)$$

$$\left(\frac{2N}{10}\right)^{th} \text{ for the } D_2 \text{ group} \quad (4.22)$$

and

$$\left(\frac{9N}{10}\right)^{th} \text{ for the } D_9 \text{ group} \quad (4.23)$$

The **actual value of the deciles** is computed from grouped data using the formula

$$D_k = L_k + \left[\frac{\left(\frac{kN}{10}\right) - Cf}{f_k}\right] * w \qquad , \; k = 1, 2, \ldots\ldots, 9 \tag{4.24}$$

L_k is the lower class boundary of the decile group
N is the total number of values
Cf is the cumulative frequency of the groups before the decile group
f_k is the frequency of the decile group
w is the class width or class size of the decile group

4.8.3 The Percentiles

Percentiles divide the data set into 100 equal parts, denoted as P_1, P_2, …, P_{99}, so that 1% of the data lies below P_1, 2% of the data below P_2,…, and 99% of the data below P_{99}.

When data are ungrouped, the position of the percentiles are

$$P_1 \;\; is\; the \left(\frac{N+1}{100}\right) th \;\; item \tag{4.25}$$

$$P_2 \;\; is\; the\; 2\left(\frac{N+1}{100}\right) th \;\; item \tag{4.26}$$

and

$$P_{99} \;\; is\; the \;\; 99\left(\frac{N+1}{100}\right) th \;\; item \tag{4.27}$$

For grouped data, **the percentile group** is found using

$$\left(\frac{N}{4}\right) th\; item \;\; for\; the\; P_1 \;\; group \tag{4.28}$$

$$\left(\frac{2N}{4}\right) th\; item \;\; for\; the\; P_2 \;\; group \tag{4.29}$$

$$\left(\frac{99N}{4}\right) th\; item \;\; for\; the\; P_{99} \;\; group \tag{4.30}$$

The **actual value of the percentiles** is computed from grouped data using the formula

$$P_k = L_k + \left[\frac{\left(\frac{kN}{100}\right) - Cf}{f_k}\right] * w \qquad k = 1, 2, \ldots.., 99 \tag{4.31}$$

where:
L_k is the lower class boundary of the percentile group
N is the total number of values

Cf is the cumulative frequency of the groups before the percentile group

$\mathbf{f_k}$ is the frequency of the percentile group

w is the class width or class size of the percentile group

The 25th percentile is also called the first quartile $(P_{25} = Q_1)$.

The 50th percentile is generally the median $(P_{50} = Q_2)$.

The 75th percentile is also called the third quartile $(P_{75} = Q_3)$.

Example 4.12

The following are the numbers of defective items produced in a month by a machine for the last 24 months. 45, 30, 36, 26,16,21,33,40,32,14,10,29,23,39,17,11,18,35,19,24,21,35,42,37. Find:

a. D_1 and D_5
b. P_{10}, and P_{75}

Solution

a. Arrange the numbers in ascending order

10,11,14,16,17,18,19,21,21,23,24,26,29,30,32,33,34,35,36,37,39,40,42,45

$D_1 = \left(\frac{N+1}{10}\right)$th item; $D_1 = \left(\frac{24+1}{10}\right)$th item

$D_1 = 2.5^{th}$ item

2.5^{th} item lies between the 2^{nd} and the 3^{rd} observation in the ordered value.

2^{nd} item = 11

3^{rd} item = 14

D_1 is the mean of 11 and 14

$$D_1 = \frac{11+14}{2} = \frac{25}{2} = 12.5$$

$D_5 = \left(\frac{5(N+1)}{10}\right)$th item; $D_5 = \left(\frac{5(24+1)}{10}\right)$th item

$D_5 = 12.5^{th}$ item

12.5^{th} item lies between 12^{th} and 13^{th} item in the ordered group

12^{th} item = 26; 13^{th}item = 29

D_5 is the mean of 26 and 29

$$D_5 = \frac{26+29}{2} = \frac{55}{2} = 27.5$$

b. $P_{10} = \left(\frac{10(24+1)}{100}\right)$th item

$P_{10} = 2.5^{th}$ item

2.5^{th} item lies between the 2^{nd} and the 3^{rd} observation in the ordered value.

2^{nd} item = 11

3^{rd} item = 14

P_{10} is the mean of 11 and 14

$$P_{10} = \frac{11+14}{2} = \frac{25}{2} = 12.5$$

$$P_{75} = \left(\frac{75(24+1)}{100}\right)\text{th item} = 18.5^{th}\text{ item}$$

This lies between the 18^{th} and 19^{th} observation

18^{th} item =35; 19^{th} item = 36

P_{75} = is the mean of 35 and 36

$$P_{75} = \frac{35+36}{2} = 35.5$$

Example 4.13

Given the table below

IQ	Frequency (f)
118-125	4
126-133	6
134-141	3
142-149	2
150-157	2

a. Find Q_1 and Q_3

b. Find D_4

c. Find P_{75}

Solution

IQ	Frequency (f)	Class Boundaries	Cumulative Frequency
118-125	4	117.5-125.5	4
126-133	6	125.5-133.5	10
134-141	3	133.5-141.5	13
142-149	2	141.5-149.5	15
150-157	2	149.5-157.5	17
Total	17		

a. $Q_1 = L_1 + \left[\frac{\left(\frac{1N}{4}\right) - Cf}{f_1}\right] * w$; $\frac{1N}{4} = \frac{1*17}{4} = 4.25$, $L_1 = 125.5$, Cf = 4, w =8, $f_1 = 6$

$$Q_1 = 125.5 + \left(\frac{4.25-4}{6}\right) * 8 = 125.83$$

$Q_3 = L_3 + \left[\frac{\left(\frac{3N}{4}\right) - Cf}{f_3}\right] * w$; $\frac{3N}{4} = \frac{3*17}{4} = 12.75$, $L_3 = 133.5$, Cf = 10, w =8, $f_3 = 3$

$Q_3 = 133.5 + \left(\frac{12.75 - 10}{3}\right) * 8 = 140.83$

b. $D_4 = L_4 + \left[\frac{\left(\frac{4N}{10}\right) - Cf}{f_4}\right] * w$; $\frac{4N}{10} = \frac{4*17}{10} = 6.8$, $L_4 = 125.5$, Cf = 4, w =8, $f_4 = 6$

$D_4 = 125.5 + \left(\frac{6.8 - 4}{6}\right) * 8 = 129.23.$

c. $P_{75} = L_{75} + \left[\frac{\left(\frac{75N}{100}\right) - Cf}{f_{75}}\right] * w$; $\frac{75N}{100} = \frac{75*17}{100} = 12.75$, $L_{75} = 133.5$, Cf = 10, w =8, $f_{75} = 3$

$P_{75} = 133.5 + \left(\frac{12.75 - 10}{3}\right) * 8 = 140.83$

4.9 BOX AND WHISKER PLOT (BOXPLOT)

Graphing a boxplot requires the knowledge of (i) the minimum value (ii) first quartile (iii) second quartile (iv) third quartile (v) maximum value.

The steps required to construct a boxplot are;

a. Construct a horizontal scale for the range of the data
b. Plot the minimum value, first quartile, second quartile, third quartile and the maximum value above the horizontal scale
c. Draw a box above the horizontal scale from the first quartile to the third quartile and draw a vertical line in the box at the second quartile.
d. Draw whiskers from the box to the minimum and maximum values.

Example 4.14: Draw a box-and-whisker plot for the following data:

5 13 6 4 15 17 18 3 32 20 31 8

Solution

Min. value = 3, Q1 = 5, Q2 = 14, Q3 = 20 and Max. Value = 32

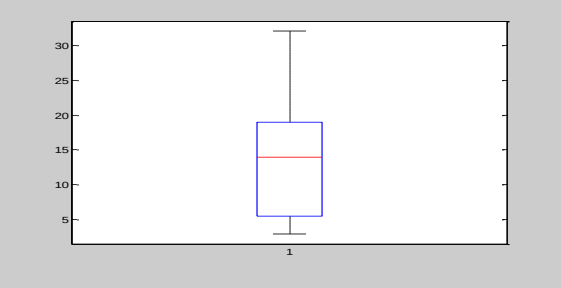


Figure 4.1: Box and Whisker plot

EXERCISE FOUR

1. Consider the following population test scores for a class:
   99, 99, 62, 75, 81, 68, 74, 86, 79, 91, 77, 82, 96, 84, 71. Find the mean, median and mode
2. The weights of 10 boys are 65, 55, 68, 71, 55, 64, 75, 67, 69 and 58
   What is P_{80} and P_{90}?
3. In order to increase the number of customers, a fast-food restaurant regularly monitors its times of services to ensure that its speed of service is improving in time. A sample of 20 times of service has been taken on a randomly selected day and is shown here (time in seconds): 45, 48, 49, 56, 61, 66, 66, 66, 70, 72, 72, 75, 78, 79, 81, 81, 83, 95, 102, 135;
   a) Determine the average, the most frequent and the median time of service.
   b) Calculate the quartiles.
4. Which of the following is not a measure of location?
   (a) Mean (b) Median (c) Variance (d) Mode (e) Quartile
5. Which of the following statement is false?
   (a) The median is a measure of central tendency
   (b) The 50th percentile is the median
   (c) An extreme value is likely to have a greater effect on the median than the mean
   (d) The 25th percentile is the first quartile (Q1)
6. For each of the following, indicate the appropriate statistical measures that may be used for analysis (e.g., median, mode and mean). List as many as are appropriate.
   a. For data measured on a nominal scale, you may use: _____
   b. For data measured on an ordinal scale, you may use: _____
   c. For data measured on an interval scale, you may use: _____
7. Find out the weighted mean of the following data:

Group	Index Number	Weights
Food	352	48
Fuel	220	10
Cloth	230	8
House Rent	160	12
Misc.	190	15

8. Calculate the geometric mean of the following data: 5, 15, 25, 35 and 45.
9. Below are the courses, unit and grade of a year 2 student in Statistics Department, FUTO. Calculate the weighted mean of the student, where A= 5, B = 4, C = 3, D = 2 and F = 0 and using the unit as weight.

Courses	Grade	Unit
Sta211	A	3
Sta221	C	3
Sta212	D	2
Sta223	A	1
Sta224	B	1

10. The following are the number of rooms in the houses of a particular locality. Find the median of the data:

No. of rooms	3	4	5	6	7	8
No of houses	38	654	311	42	12	2

Use the information below to answer question 11-13

The table below is the distribution of marks secured by some students in an examination:

Marks	No. of students
1-20	42
21-30	38
31-40	120
41-50	84
51-60	48
61-70	36
71-80	31

11. Find the median mark.
12. Calculate Q_1 and Q_3
13. Calculate D_2 and P_{50}
14. In the frequency distribution of 100 families given below: the number of families corresponding to expenditure groups 20 – 39 and 60 – 79 are missing from the table. However the median is known to be 49.87. Find out the missing frequencies.

Expenditure	0-19	20-39	40-59	60-79	80-99
No. of families	14	X	27	y	15

15. Calculate the mode of the following data:

Number	1	2	3	4	5	6	7	8	9	10
Marks obtained	10	27	24	12	27	27	20	18	15	30

16. The table below is the marks of students in an examination:

Marks	Frequency
0 – 9	5
10 – 19	10
20 – 29	15
30 – 39	14
40 – 49	10
50 – 59	5
60 – 69	3

Calculate the mean of the mark using the assumed mean method. (Use A = 29.5)

17. The mean temperature for the past ten days was 22° Celsius. If the sum of the temperatures for the first nine days was 200, what was the temperature on day 10?

18. Imagine that you received the following data on a test: 20, 22, 23, 23, 23, 23, 23, 23, 24, 25, 28, 29, 30, 30, 30, 30, 31, 32, 32, 33, 33, 34, 35, 35, 36, 36, 37 and 37. Compute the mean and mode of the data

19. The table below gives the frequency distribution of marks obtained by a group of students in a Statistics test.

Marks	3	4	5	6	7	8	9
Frequency	6	x-2	x	8	6	4	2

If the mean mark is 5.5, calculate (i) the value of x (ii) the Mode (iii) the Median

20. Prove the following:

(i) If n numbers $X_1, X_2, ..., X_n$ have deviation from any number A given respectively by $d_1 = x_1 - A$, $d_2 = x_2 - A$, ……, $d_n = x_n - A$ then

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = A + \frac{\sum d}{n}$$

(ii) Use the above method to find the arithmetic mean of the numbers 5,8,11,9,12,6,14 and 10.Take (i) 9 and (ii) 20 as the value of A.

21. (a) Four groups of students consisting of 15, 20, 10 and 18 individuals reported mean heights of 1.62, 1.48, 1.53 and 1.40 meters respectively. Find the mean height of all the students.
(b) Find the harmonic mean H of the numbers 3, 5, 6,6,7,10,12.

22. Find (a) the geometric mean and (b) the arithmetic mean of the numbers 3, 5, 6, 6, 7, 10 and 12.

23. (a) A man travels from A to B at an average speed 30km/h and returns from B to A along the same route at an average speed of 60km/h. find the average speed for the entire trip.

24. The following table shows the distribution of the maximum loads in Kilonewtons supported by certain cables produced by a company. Determine the mean maximum loading using the long method.

Maximum Load(KN)	93 – 97	98 – 102	103 – 107	108 –112	113 –117	118 -122	123 -127	128 -132
Number of Cables	2	5	12	17	14	6	3	1

25. The following table gives the frequency distribution of the monthly allowance of 65 employees in Naira in a company. Compute the Mean using the coding method I. (Let A = 75)

Allowance (A)	50.00 – 59.99	60.00 – 69.99	70.00 – 79.99	80.00 – 89.99	90.00 – 99.99	100.00 – 109.99
No. of Employee	8	10	16	14	10	5

26. Using the frequency distribution table of problem 25. Compute the mode monthly allowance of the employees.
27. Using the frequency distribution table of problem 25. Compute the median monthly allowance of the employees
28. Two machines "A" and "B" produce 50kg bags of cement. The quality control department of the cement company makes regular random checks to insure that the bags are of correct weights. Such random checks on 7 bags of each type gave the following:

Machine A	50.21	49.88	50.60	49.63	50.40	49.62	49.90
Machine B	50.04	49.60	50.25	50.12	49.82	50.36	50.10

(a) Calculate the harmonic means for each type, giving your answer correct to two decimal places.
(b) Compute the arithmetic means for each type correct to two decimal places.
(c) Determine the median in each type.

29. The monthly expenditure on food by sample of households in a city is distributed as follows:

Expenditure (N'000)	**Number of households**
6.01 – 6.50	25
6.51 – 7.00	40
7.01 – 7.50	50
7.51 – 8.00	70
8.01 – 9.00	45
9.01 – 10.00	30
10.01 – 10.50	20

(a) Compute the arithmetic Mean.
(b) Compute the Median and the other quartiles of the distribution.

30. The following table gives the frequency distribution of scores by students in a quiz.

Scores	0	1	2	3	4	5	6
No. of Students	14	10	8	7	5	4	2

Find the mean, median and mode

31. The following data show the incomes (in thousands of naira) of a sample of 10 households.

    25 40 22 100 80 17 35 20 62 50

    Construct a box-and-whisker plot for these data.

32. Draw a Box-and-Whisker plot for the following data

    60 52 78 60 63 60 47 81 42 50 73
    90 45 57 39 70

33. ABC Technology manufactures computer keyboard. The following are the numbers of keyboards produced at the company for a sample of 15 days

    42 34 38 35 38 36 31 38
    33 39 45 41 33 45 37

    Prepare a Box-and-Whisker plot.

CHAPTER FIVE

MEASURES OF DISPERSION

5.0 INTRODUCTION

In the previous chapter, Measures of Central Tendency were treated. Concepts like "Coding Method 1" and "Coding Method 2" were introduced and explained. These concepts shall be explored further in the treatment of Measures of Dispersion. The mean of a set of measurements only locates the centre of the distribution of data and by itself, it does not provide an adequate description of a set of measurements. This is because two sets of measurements could have widely different frequency distributions but equal means. The difference lies in the variation or dispersion of measurements on either side of the mean. Therefore, to describe data adequately, we must also define measures of data variability. These measures of variation are also called "Measures of Dispersion". They are measures of spread of observations from certain values. They play an important role in the study of Statistics as some examples in this chapter will readily show. There are many measures of variation but the widely used ones are: the variance, the standard deviation, the mean deviation, the range, the interquartile range and the semi-interquartile range. We shall study these measures in the sections that follow.

5.1. THE VARIANCE

The variance is a measure of spread of the observations from the central value, the mean. Its positive square root is called the **STANDARD DEVIATION.** The standard deviation is often used more than the variance because it has the same unit of measure as the variables used in computing it. In other to convert the observations back to the original values, then, the standard deviation is taken. In this section, the variance, the standard deviation and the applications of the standard deviation will be treated. Treated also will be the calculation of these statistics using coding methods 1 and 2. We shall also calculate these measures for the population and sample as well as for ungrouped and grouped cases.

5.1.1 THE VARIANCE AND THE STANDARD DEVIATION FOR UNGROUPED DATA

Let $X_1, X_2, \ldots, X_N$ be the population values, then, the variance, written as 'σ^2' for the population is given for ungrouped data by:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \qquad (5.1)$$

where 'N' signifies population size and 'μ'the population mean as given in the previous chapter. Hence, the standard deviation is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}} \tag{5.2}$$

And if $X_1, X_2, \ldots, X_n$ is a random sample from the population, then, the sample variance, denoted by '**S^2**', is given by:

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} \tag{5.3}$$

where 'n' signifies sample size and '$\overline{X}$' the sample mean. As in (5.2), the sample standard deviation is given as:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} \tag{5.4}$$

We note that the divisor of S^2 is "n-1" and not "n" because the population mean, μ has been estimated by the sample mean, $\overline{X}$. So S^2 defined in this way, provides a "better" estimator for σ^2. However, when the sample size is at least 30, S^2 can be divided by "n" only.

5.1.1(a) Steps in the Calculation of the Variance and Standard deviation for the Ungrouped data

1. Find the population mean, μ, or the sample mean, $\overline{X}$, from the given data.
2. Subtract 'μ' or '$\overline{X}$' from each observation.
3. Square each of the values in (2).
4. Sum (3) and divide by N (or 'n-1' or 'n' as the case may be).

The value in (4) gives the variance while the positive square root gives the standard deviation.

Example 5.1: Given the following data: 6, 13, 25, 7, 19, 22, 4, find the variance and the standard deviation treating the data (i) as a population and (ii) as a sample.

Solution:

	X	**(X - 13.71)**	**$(X - 13.71)^2$**
	6	-7.71	59.51
	13	-0.71	0.51
	25	11.29	127.37
	7	-6.71	45.08
	19	5.29	27.94
	22	8.29	68.65
	4	-9.71	94.37
Total	**96**		**423.43**

where X represents the values. Note from the four steps above the following:

Treating the data as population,

1. $\mu = \frac{1}{7}\{6+13+25+7+19+22+4\} = \frac{96}{7} = 13.71.$

And treating the data as a sample,

$\overline{X} = \frac{1}{7}\{6+13+25+7+19+22+4\} = \frac{96}{7}$

$= 13.71.$

2. Step 2 is shown in column 2.
3. Step 3 is shown in column 3.
4. Step 4 – sum of the squared deviations – is shown at the foot of column 3 (423.43).

The students are expected to confirm these values with their calculators.

$\therefore \sigma^2 = \frac{423.43}{7}$

$= 60.49$

$\therefore S^2 = \frac{423.43}{7-1}$

$= 70.57$

while,

$\sigma = \sqrt{60.49}$

$= 7.78$

$S = \sqrt{70.57} = 8.40.$

Therefore, (i) variance for the population = 60.49 and for the sample, 70.57.
(ii) standard deviation for the population =7.78 and for the sample, 8.40

5.1.1.(b). Alternative Formula for the Calculation of Equations (5.1) and (5.3):

Equations (5.1) and (5.3) can alternatively be written as:

$$\sigma^2 = \frac{N\left(\sum_{i=1}^{N} X_i^2\right) - \left(\sum_{i=1}^{N} X_i\right)^2}{N^2} \quad (5.5)$$

$$S^2 = \frac{n\left(\sum_{i=1}^{n} X_i^2\right) - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)} \quad (5.6)$$

These formulas are often referred to as **Computational Formulas.**

Example 5.2: Repeat Example 5.1 using Equations (5.5) and (5.6).
Solution:

	X	**X^2**
	6	36
	13	169
	25	625
	7	49
	19	361
	22	484
	4	16
Total	**96**	**1740**

$$\sigma^2 = \frac{7(1740)-(96)^2}{7X7} = 2964/49$$

$$= 60.49$$

Hence, the standard deviation is:

$$\sigma = \sqrt{60.49}$$

$$= 7.78.$$

For the sample,

$$S^2 = \frac{7(1740)-(96)^2}{7(7-1)} = 2964/42 = 70.57$$

and the sample standard deviation is:

$$S = \sqrt{70.57} = 8.40.$$

Therefore, (i) variance for the population = 60.49 and for the sample, 70.57.

(ii) standard deviation for the population =7.78 and for the sample, 8.40

which correspond with the answers given in Example 5.1.

5.1.1.(c). Coding Methods for the Calculation of Equations (5.1) and (5.3):

(1) Coding Method 1:

Recall that in coding method 1, a central value (or any convenient value), X_o, is subtracted from X and the result called "D". Using D therefore, Equations (5.1) and (5.3) can be written respectively as:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(D_i - \mu_D)^2}{N} \tag{5.7}$$

where

$$\mu_D = \frac{\sum_{i=1}^{N} D_i}{N}$$

and

$$S^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1} \tag{5.8}$$

where

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

You may see Note A at the end of this chapter.

Example 5.3: Repeat Example 5.1 using Equations (5.7) and (5.8). Take $X_o = 13$.

Solution:

	X	D=X-13	D - 0.71	$(D - 0.71)^2$
	6	-7	-7.71	59.4441
	13	0	-0.71	0.5041
	25	12	11.29	127.4641
	7	-6	-6.71	45.0241
	19	6	5.29	27.9841
	22	9	8.29	68.7241
	4	-9	-9.71	94.2841
Total		**5**		423.4287

where mean = 0.71.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(D_i - 0.71)^2}{7}$$

$$= \frac{423.4287}{7}$$

= 60.49 with a standard deviation $\sqrt{60.49}$ = 7.78 (for the population)

Again,

$$S^2 = \frac{\sum_{i=1}^{N}(D_i - 0.71)^2}{7-1}$$

$$= \frac{423.4287}{6}$$

= 70.57 with a standard deviation $\sqrt{70.57}$ = 8.40 (for the sample).

which correspond with the answers given in Example 5.1.

In accordance with equations (5.5) and (5.6), using coding method 1, the variance can be written as:

$$\sigma^2 = \frac{N\left(\sum_{i=1}^{N} D_i^2\right) - \left(\sum_{i=1}^{N} D_i\right)^2}{N^2} \qquad (5.9)$$

$$S^2 = \frac{n\left(\sum_{i=1}^{n} D_i^2\right) - \left(\sum_{i=1}^{n} D_i\right)^2}{n(n-1)} \qquad (5.10)$$

You may see Note B at the end of this chapter.

Example 5.4: Repeat Example 5.1 using Equations (5.9) and (5.10).

Solution:

	X	D=X-13	D^2
	6	-7	49
	13	0	0
	25	12	144
	7	-6	36
	19	6	36
	22	9	81
	4	-9	81
Total		**5**	**427**

$$\sigma^2 = \frac{7(427)-(5)^2}{7X7}$$

$$= 60.49$$

$$\sigma = \sqrt{60.49}$$

$$= 7.78.$$

Also,

$$S^2 = \frac{7(427)-(5)^2}{7(7-1)}$$

$$= 70.57$$

$$S = \sqrt{70.57}$$

$$= 8.40.$$

Therefore, (i) variance for the population = 60.49 and for the sample, 70.57.
(ii) standard deviation for the population =7.78 and for the sample, 8.40
which correspond with the answers given in Example 5.1.

Observe that by subtracting X_o from each X, the variance is unchanged. Therefore, variance is invariant (not affected) by subtraction of the same quantity from every observation.

(2) Coding Method 2:

In order to scale down the data further for easy computation, "D" is divided by a scale factor, "C" and the result called "U". Using U therefore, Equations (5.7) and (5.8) can be written as:

$$\sigma^2 = \frac{C^2\sum_{i=1}^{N}(U_i-\mu_U)^2}{N} \qquad (5.11)$$

where

$$\mu_U = \frac{\sum_{i=1}^{N}U_i}{N}$$

$$S^2 = \frac{C^2 \sum_{i=1}^{n} \left(U_i - \overline{U}\right)^2}{n-1} \qquad (5.12)$$

where

$$\overline{U} = \frac{\sum_{i=1}^{n} U_i}{n}$$

You may see Note C at the end of this chapter.

Example 5.5: Repeat Example 5.1 using Equations (5.11) and (5.12) taking C = 3.

Solution:

	X	D=X-13	U=D/3	U-0.24	$(U-0.24)^2$
	6	-7	-2.33	-2.57	6.62
	13	0	0	-0.24	0.06
	25	12	4	3.76	14.14
	7	-6	-2	-2.24	5.02
	19	6	2	1.76	3.10
	22	9	3	2.76	7.62
	4	-9	-3	-3.24	10.50
Total		**5**	**1.67**		**47.05**

where mean = $1.67/7$ = **0.24.**

Observe that by dividing through by 3, the sum of squares of deviation from the mean has reduced from 423.43 to 47.05.

Hence,

$$\sigma^2 = \frac{3^2 \sum_{i=1}^{N} \left(U_i - 0.24\right)^2}{7}$$

$$= \frac{9(47.05)}{7}$$

= 60.49 with a standard deviation $\sqrt{60.49}$ = 7.78 (for the population)

And,

$$S^2 = \frac{3^2 \sum_{i=1}^{N} \left(U_i - 0.24\right)^2}{7-1}$$

$$= \frac{9(47.05)}{6}$$

= 70.57 with a standard deviation $\sqrt{70.57}$ = 8.40 (for the sample).

which correspond with the answers given in Example 5.1.

In accordance with equations (5.5) and (5.6), using coding method 2, the variance can be written as:

$$\sigma^2 = C^2\left\{\frac{N\left(\sum_{i=1}^{N}U_i^2\right)-\left(\sum_{i=1}^{N}U_i\right)^2}{N^2}\right\} \quad (5.13)$$

$$S^2 = C^2\left\{\frac{n\left(\sum_{i=1}^{n}U_i^2\right)-\left(\sum_{i=1}^{n}U_i\right)^2}{n(n-1)}\right\} \quad (5.14)$$

You may see Note C at the end of this chapter.

Example 5.6: Repeat Example 5.1 using Equations (5.13) and (5.14) taking C = 3.

Solution:

	X	**D=X-13**	**U=D/3**	**U^2**
	6	-7	-2.33	5.44
	13	0	0	0
	25	12	4	16
	7	-6	-2	4
	19	6	2	4
	22	9	3	9
	4	-9	-3	9
Total		**5**	**1.67**	**47.44**

$$\sigma^2 = 3^2\left\{\frac{7(47.44)-(1.67)^2}{7X7}\right\}$$

$$= 60.49$$

$$\sigma = \sqrt{60.49}$$

$$= 7.78.$$

Also,

$$S^2 = 3^2\left\{\frac{7(47.44)-(1.67)^2}{7(7-1)}\right\}$$

$$= 70.57$$

$$S = \sqrt{70.57}$$

$$= 8.40.$$

Therefore, (i) variance for the population = 60.49 and for the sample, 70.57.
(ii) standard deviation for the population =7.78 and for the sample = 8.40 which correspond with the answers given in Example 5.1.

5.1.2. THE VARIANCE AND THE STANDARD DEVIATION FOR GROUPED DATA

For grouped data with classes: a_1-b_1, a_2-b_2, …, a_k-b_k, let X_1, X_2, …, X_k denote the classmarks and f_1, f_2, …, f_k denote the corresponding frequencies, then, the variance, written as 'σ^2' for the population for grouped data is given by:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (X_i - \mu)^2}{\sum_{i=1}^{k} f_i} \qquad (5.15)$$

where 'k' is the number of classes, a_i = the lower limit of class i, b_i = the upper limit of class i and

$$X_i = \frac{1}{2}(a_i + b_i) \forall i;\ i = 1, 2, 3, ..., k. \qquad (5.16a)$$

And for the sample, the variance, denoted by '**S^2**', is given by:

$$S^2 = \frac{\sum_{i=1}^{k} f_i (X_i - \overline{X})^2}{\sum_{i=1}^{k} f_i - 1} \qquad (5.16b)$$

where 'k' is the number of classes.

5.1.2(a) Steps in the Calculation of the Variance and Standard Deviation for the Grouped Data:

1. Find the population mean, μ, or the sample mean, $\overline{X}$, from the given data.
2. Subtract 'μ' or '$\overline{X}$' from each observation (here, the classmark or the class-midpoint).
3. Square each of the values in (2) and multiply each by f_i.
4. Sum (3) and divide by $\sum f_i$ for the population and by $\sum f_i$ -1 for the sample.

The value in (4) gives one the variance and the positive square root, the standard deviation.

Example 5.7: The following data is a grouped data for the scores of 79 students in a Statistics test. Find the variance and the standard deviation treating the data (i) as a population and (ii) as a sample.

Exam Score	**Freq.**
10 - 19	13
20 - 29	44
30 - 39	19
40 - 49	3
Total	**79**

Solution:

Exam Score	F	X	Fx	X-26.02	$(X-26.02)^2$	$f(X-26.02)^2$
10 - 19	13	14.5	188.5	-11.52	132.71	1725.24
20 - 29	44	24.5	1078	-1.52	2.31	101.66
30 - 39	19	34.5	655.5	8.48	71.91	1366.30
40 - 49	3	44.5	133.5	18.48	341.51	1024.53
Total	**79**		**2055.5**			**4217.72**

where X = classmark and mean = 26.02

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i \left(X_i - 20.02\right)^2}{79}$$

$$= \frac{4217.72}{79}$$

= 53.39 with standard deviation, σ =7.31 for the population.

Also,

$$S^2 = \frac{\sum_{i=1}^{k} f_i \left(X_i - 20.02\right)^2}{79-1}$$

$$= \frac{4217.72}{78}$$

= 54.07 with standard deviation, S = 7.35 for the sample.

5.1.2.(b). Alternative Formula for the Calculation of Equations (5.15) and (5.16):

Equations (5.15) and (5.16) can alternatively be written as:

$$\sigma^2 = \frac{N\left(\sum_{i=1}^{k} f_i X_i^2\right) - \left(\sum_{i=1}^{k} f_i X_i\right)^2}{N^2} \qquad (5.17)$$

where $N = \sum_i f_i$

$$S^2 = \frac{n\left(\sum_{i=1}^{k} f_i X_i^2\right) - \left(\sum_{i=1}^{k} f_i X_i\right)^2}{n\left(\sum_{i=1}^{k} f_i - 1\right)} \qquad (5.18)$$

where $n = \sum_i f_i$

Example 5.8: Repeat Example 5.7 using Equations (5.17) and (5.18).

Solution:

Exam Score	f	X	X^2	fX	fX^2
10 - 19	13	14.5	210.25	188.5	2733.25
20 - 29	44	24.5	600.25	1078	26411
30 - 39	19	34.5	1190.25	655.5	22614.75
40 - 49	3	44.5	1980.25	133.5	5940.75
Total	**79**			**2055.5**	**57699.75**

$$\sigma^2 = \frac{\mathbf{79}(\mathbf{57699.75}) - (\mathbf{2055.5})^2}{79X\,79}$$

$$= 53.39$$

$$\sigma = \sqrt{53.39}$$

$$= 7.31.$$

Also,

$$S^2 = \frac{\mathbf{79}(\mathbf{57699.75}) - (\mathbf{2055.5})^2}{79(79-1)}$$

$$= 54.07$$

$$S = \sqrt{54.07}$$

$$= 7.35.$$

Therefore, (i) variance for the population = 53.39 and for the sample, 54.07.
(ii) standard deviation for the population =7.31 and for the sample, 7.35.

5.1.2.(c). Coding Methods for the Calculation of Equations (5.15) and (5.16):

(1) Coding Method 1:

Recall that coding method 1 was used to calculate the variance for the ungrouped data in section (5.1.1(c)). Using D therefore, equations (5.15) and (5.16) can be written as:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (D_i - \mu_D)^2}{\sum_{i=1}^{k} f_i} \tag{5.19}$$

$$S^2 = \frac{\sum_{i=1}^{k} f_i \left(D_i - \overline{D}\right)^2}{\sum_{i=1}^{k} f_i - 1} \tag{5.20}$$

for the population and the sample respectively.

Example 5.9: Repeat Example 5.7 using Equations (5.19) and (5.20).

Solution:

Exam Score	f	X	D = X-24.5	fD	D-1.52	$(D-1.52)^2$	$f(D-1.52)^2$
10 - 19	13	14.5	-10	-130	-11.52	132.71	1725.24
20 - 29	44	24.5	0	0	-1.52	2.31	101.66
30 - 39	19	34.5	10	190	8.48	71.91	1366.30
40 - 49	3	44.5	20	60	18.48	341.51	1024.53
Total	**79**		**20**	**120**			**4217.72**

where $X_o = 24.5$, $\overline{D} = 1.52$.

$$\therefore \quad \sigma^2 = \frac{\sum_{i=1}^{4} f_i \left(D_i - 1.52\right)^2}{79}$$

$$= \frac{4217.72}{79}$$

= 53.39 with standard deviation, σ = 7.31 for the population.

Also,

$$S^2 = \frac{\sum_{i=1}^{4} f_i \left(X_i - 1.52\right)^2}{79-1}$$

$$= \frac{4217.72}{78}$$

= 54.07 with standard deviation, s =7.35 for the sample.

Like Equations (5.9) and (5.10), using coding method 1, the variance can be written as:

$$\sigma^2 = \frac{N\left(\sum_{i=1}^{k} f_i D_i^2\right) - \left(\sum_{i=1}^{k} f_i D_i\right)^2}{N^2} \tag{5.21}$$

where $N = \sum_i f_i$

$$S^2 = \frac{n\left(\sum_{i=1}^{k} f_i D_i^2\right) - \left(\sum_{i=1}^{k} f_i D_i\right)^2}{n(n-1)} \tag{5.22}$$

where $n = \sum_i f_i$

Example 5.10: Repeat example 5.7 using Equations (5.21) and (5.22).

Solution:

Exam Score	f	X	D = X-24.5	D^2	fD	fD^2
10 - 19	13	14.5	-10	100	-130	1300
20 - 29	44	24.5	0	0	0	0
30 - 39	19	34.5	10	100	190	1900
40 - 49	3	44.5	20	400	60	1200
Total	**79**		**20**		**120**	**4400**

$$\sigma^2 = \frac{\mathbf{79}(4400) - (120)^2}{79X\,79}$$

$$= 53.39$$

$$\sigma = \sqrt{53.39}$$

$$= 7.31.$$

Also,

$$S^2 = \frac{\mathbf{79}(4400) - (120)^2}{79(79-1)}$$

$$= 54.07$$

$$S = \sqrt{54.07}$$

$$= 7.35.$$

Therefore, (i) variance for the population = 53.39 and for the sample, 54.07.
(ii) standard deviation for the population=7.31 and for the sample, 7.35.

(2) Coding Method 2:

As above, the population and sample variances are computed using Equations (5.23) and (5.24) below respectively but note that the class size, C, can be used as the scale factor for "U".

$$\sigma^2 = \frac{C^2\sum_{i=1}^{k} f_i(U_i - \mu_U)^2}{\sum_{i=1}^{k} f_i} \qquad (5.23)$$

$$S^2 = \frac{C^2\sum_{i=1}^{k} f_i(U_i - \bar{U})^2}{\sum_{i=1}^{k} f_i - 1} \qquad (5.24)$$

Example 5.11: Repeat Example 5.7 using Equations (5.23) and (5.24).

Solution:

Exam Score	**f**	**X**	**D = X-24.5**	**U=D/10**	**fU**	**U-0.15**	**$(U-0.15)^2$**	**$f(U-0.15)^2$**
10 - 19	13	14.5	-10	-1	-13	-1.15	1.32	17.19
20 - 29	44	24.5	0	0	0	-0.15	0.02	0.99
30 - 39	19	34.5	10	1	19	0.85	0.72	13.73
40 - 49	3	44.5	20	2	6	1.85	3.42	10.27
Total	**79**		**20**	**2**	**12**			**42.18**

where C = 10 and $\bar{U}$ = 0.15.

$$\therefore \quad \sigma^2 = \frac{10^2\sum_{i=1}^{k} f_i(U_i - 0.15)^2}{79}$$

$$= \frac{100(42.18)}{79}$$

$$= 53.39 \text{ with standard deviation, } \sigma = 7.31 \text{ for the population.}$$

Also,

$$S^2 = \frac{10^2 \sum_{i=1}^{k} f_i \left(X_i - 0.15\right)^2}{79-1}$$

$$= \frac{100(42.18)}{78}$$

= 54.07 with standard deviation, S =7.35 for the sample.

Also, in accordance with equations (5.13) and (5.14), using coding method 2, the variance can be written as:

$$\sigma^2 = C^2 \left\{ \frac{N\left(\sum_{i=1}^{N} f_i U_i^2\right) - \left(\sum_{i=1}^{N} f_i U_i\right)^2}{N^2} \right\} \qquad (5.25)$$

where N = $\sum_i f_i$

$$S^2 = C^2 \left\{ \frac{n\left(\sum_{i=1}^{n} f_i U_i^2\right) - \left(\sum_{i=1}^{n} f_i U_i\right)^2}{n(n-1)} \right\} \qquad (5.26)$$

where n = $\sum_i f_i$

Example 5.12: Repeat Example 5.7 using Equations (5.25) and (5.26).

Solution:

Exam Score	f	X	D = X-24.5	U=D/10	fU	U^2	fU^2
10 - 19	13	14.5	-10	-1	-13	1	13
20 - 29	44	24.5	0	0	0	0	0
30 - 39	19	34.5	10	1	19	1	19
40 - 49	3	44.5	20	2	6	4	12
Total	**79**		**20**		**12**		**44**

$$\therefore \sigma^2 = 100^2 \left\{ \frac{79(44) - (12)^2}{79^2} \right\}$$

= 53.39 with standard deviation, σ =7.31 for the population.

$$S^2 = 100^2 \left\{ \frac{79(44) - (12)^2}{79(79-1)} \right\}$$

= 54.07 with standard deviation, S =7.35 for the sample.

5.1.3. APPLICATIONS OF THE STANDARD DEVIATION

(a). THE Z SCORE: The Z-Score measures how many standard deviations are above or below the mean. It is unitless and is used to compare two or more sets of data measured on the same or different units relative to their groups. The formulas are given as follows:

$$Z = \frac{X - \mu}{\sigma} \quad \text{(for the population)} \qquad (5.27)$$

$$Z = \frac{X - \overline{X}}{S} \quad \text{(for the sample)} \qquad (5.28)$$

where Z = Z score

X = the observation from the population with mean, μ, and standard deviation, σ or the observation from the sample with mean, $\overline{X}$, and standard deviation, S

Example 5.13: Supposing that in a Harmattan Semester, a student scored 85% in a Physics course with a mean grade of 55% and a standard deviation of 9%; he further scored 60% in a Statistics course with a mean grade of 45% and a standard deviation of 3%. Relatively speaking, in which course did the student perform better?

Solution:

Let X represent the scores in Physics, and let Y represent the grades in Statistics. Then,

$$Z_x = \frac{85-55}{9} = 3.33$$

$$Z_y = \frac{60-45}{3} = 5$$

Since the student's Z score is higher in Statistics course, then we say that the student performed better in Statistics than in Physics.

(b). THE COEFFICIENT OF VARIATION:

The standard deviation does not tell us much about the variability of a single set of data. A more appropriate measure is the coefficient of variation which is defined, for the random variable X, as:

$$CV_x = \frac{\sigma_x}{\mu}.100\% \quad \text{(for the population)} \qquad (5.29)$$

$$CV_x = \frac{S_x}{\overline{X}}.100\% \quad \text{(for the sample)} \qquad (5.30)$$

It can also be used to compare the variability of two or more sets of data measured on the same or different units since it is measured as a percentage.

Example 5.14: Compare the variability of the data in Examples 5.1 and 5.7, taking the data as a population.

Solution:

For the data in Example 5.1, the coefficient of variation, CV = $\frac{7.78}{13.71} x100\%$ = 57%

For the data in Example 5.7, the coefficient of variation, CV = $\frac{7.31}{26.02} x100\%$ = 28%

Since the CV for Example 5.7 is smaller than that of Example 5.1, we conclude that the data in Example 5.7 is less variable than that of Example 5.1.

Example 5.15: Consider the following samples of length measurements, in metres, of ready to wear trousers made by companies X and Y. In buying ready-made trousers, which company would one feel confident to buy from and why?

X	9.7	9	8.7	10.1	8.5	14.2
Y	10.1	7.7	8.2	13.4	11.2	9.6

Solution:

The mean and the standard deviation of the data are presented in the table below:

Company	n	Mean	Std. Dev.	Min	Max
X	6	10.03	2.13	8.5	14.2
Y	6	10.03	2.08	7.7	13.4

We observe that the mean measurements for the two companies are the same but with different standard deviations. Actually, the standard deviation of the measurements from Company Y is smaller than that of Company X. Therefore, one would feel more confident to buy from Company Y since her products seem to be closer to the advertised average measurement.

5.2. THE MEAN DEVIATION:

Another measure of variation is the mean deviation.

5.2.1. THE MEAN DEVIATION FOR THE UNGROUPED DATA:

For ungrouped data, the mean deviation is given by:

$$\frac{\sum_{i=1}^{N}|X_i-\mu|}{N} \qquad \text{(for the population)} \qquad (5.31)$$

and

$$\frac{\sum_{i=1}^{n}|X_i-\overline{X}|}{n} \qquad \text{(for the sample)} \qquad (5.32)$$

Steps in the Calculation of the Mean Deviation for Ungrouped Data:

1. For the random variable, X, find the population mean, μ, or the sample mean, $\overline{X}$, from the given data.
2. Subtract 'μ' or '$\overline{X}$' from each observation.
3. Take the absolute values of each of the values in (2).
4. Sum (3) and divide by "N" for the population, and by "n" for the sample.

Example 5.16: Find the Mean Deviation for the data in Example 5.1, treating it as (i) a population and (ii) as a sample.

Solution:

	X	(X - 13.71)	|X - 13.71|
	6	-7.71	7.71
	13	-0.71	0.71
	25	11.29	11.29
	7	-6.71	6.71
	19	5.29	5.29
	22	8.29	8.29
	4	-9.71	9.71
Total	96		49.71

∴ The mean deviation = $\frac{49.71}{7} = 7.1$ for both the population data and the sample data.

5.2.2. THE MEAN DEVIATION FOR THE GROUPED DATA

For grouped data, the mean deviation is given by:

$$\frac{\sum_{i=1}^{k} f_i |X_i - \mu|}{\sum_{i=1}^{k} f_i} \qquad \text{(for the population)} \qquad (5.33)$$

and

$$\frac{\sum_{i=1}^{k} f_i |X_i - \overline{X}|}{\sum_{i=1}^{k} f_i} \qquad \text{(for the sample)} \qquad (5.34)$$

Steps in the Calculation of Mean Deviation for Grouped Data:

1. For the random variable, X, find the population mean, μ, or the sample mean, $\overline{X}$, from the given data.
2. Subtract 'μ' or '$\overline{X}$' from each observation.
3. Take the absolute values of each of the values in (2).
4. Multiply (3) by f_i.
5. Sum (4) and divide by "$\sum_{i=1}^{k} f_i$" for both the population, and the sample.

Example 5.17: Find the Mean Deviation for the data in Example 5.7, treating it as (i) a population and (ii) as a sample.

Solution:

Exam Score	**f**	**X**	**fX**	**X-26.02**	**|X-26.02|**	**f|X-26.02|**
10 – 19	13	14.5	188.5	-11.52	11.52	149.76
20 – 29	44	24.5	1078	-1.52	1.52	66.88
30 – 39	19	34.5	655.5	8.48	8.48	161.12
40 – 49	3	44.5	133.5	18.48	18.48	55.44
Total	**79**		**2055.5**			**433.2**

∴ Mean Deviation = $\frac{433.2}{79} = 5.48$ for both the population data and sample data.

5.3. THE RANGE

The range is the difference between the highest and the least observation in a set of data. We met the range when we formed a group frequency distribution of a set of data. Given that the range can easily be computed with information on the maximum and minimum value of the data set, users requiring only a rough indication of the data spread may prefer to use this indicator over more sophisticated measures of spread like the standard deviation. The range is often used in Statistical Quality Control where there is need to keep a continuous check on manufactured products in view of meeting the required specifications.

Example 5.18: Find the range of the data in Example 5.15.

Solution:

For X, the range is 14.2 – 8.5 = 5.7

For Y, the range is 13.4 – 7.7 = 5.7.

Both X and Y have the same range.

5.4. THE INTERQUARTILE RANGE:

The interquartile range measures the variability in the middle 50% of a set of data. It is given by:

$$Q_3 - Q_1 \qquad (5.35)$$

where Q_3 = Upper Quartile and Q_1 is the Lower Quartile.

Example 5.19: Find the interquartile range for the data in Example 5.1.

Solution:

X	
4	
	Q1 = 6
6	
7	
13	
	Q3= 22
19	
22	
25	

∴ the interquartile range is 22 - 6 = 16.

Example 5.20: Find the interquartile range for the data in Example 5. 7.

Solution:

Exam Score	f	cf	
10 – 19	13	13	
20 – 29	44	57----------	Q1 class
30 – 39	19	76----------	Q3 class
40 – 49	3	79	
Total	**79**		

$$Q_1 = 19.5 + 10\left(\frac{19.75 - 13}{44}\right)$$

= 21.03

$$Q_3 = 29.5 + 10\left(\frac{59.25 - 57}{19}\right)$$

= 30.68

∴ Interquartile Range is: 30.68 – 21.03 = 9.65

5.5. THE SEMI-INTERQUARTILE RANGE

The Semi-Interquartile Range is given as:

$$\frac{Q_3 - Q_1}{2} \tag{5.36}$$

Example 5.21: Find the semi-interquartile range for the data in Example 5.1.

Solution:

The Semi-interquartile range is given as: $\frac{22 - 6}{2} = 8$

Example 5.22: Find the semi-interquartile range for the data in Example 5.7.

Solution:

The Semi-interquartile range is given as: $\frac{30.68 - 21.03}{2} = \frac{9.65}{2} = 4.83.$

EXERCISE FIVE

1. The following data represents the number of female births (per 1,000 women) in a certain locality in Nigeria.

Age of Mothers (Years)	Number of female births per 1,000 women
15 - 19	30
20 - 24	153
25 - 29	170
30 - 34	84
35 - 39	36
40 - 44	15
45 - 49	3
TOTAL	491

   i. Calculate the standard deviation

   ii. Calculate the coefficient of variation

   iii. Interpret your result in (ii) above.

2. The following data represents the charges(in naira) billed by an electrician for his 25 days home service calls:

   59.12, 81.74, 57.29, 64.35, 77.29, 84.10, 68.25, 87.25, 78.95, 59.95, 58.42, 75.80, 64.35, 58.75, 83.91, 71.50, 60.15, 80.29, 61.26, 69.32, 56.07,66.01, 61.13, 87.29, 65.68.

3. In question 1, using an assumed mean of 37, find the standard deviation using coding method 1.

4. In question 1 above, calculate the standard deviation, using coding method 2.

5. Find the variance and standard deviation of the ungrouped data in Question 2 above.

6. Solve Question 5 above using an assumed mean of 65.68 (i.e. by using coding method 1)

7. Solve Question 6 above, using coding method 2.
   Use c = 6.

8. In Question 1, assume the data to be population values, find the variance and standard deviation.

9. In Question 5, assume the data to be population values, find the variance and standard deviation.

10. In Question 1, find the mean deviation.

11. For the ungrouped data of Question 2, find the mean deviation.

12. Using the data in Question 1, find the interquartile range and interpret your result.

13. Find the interquartile range for the ungrouped data in Question 2 and interpret your result.

14. Compare the variability in Questions 1 and 2.

15. The minimum temperatures for the widely separated locations in Imo State on Jan 1, 2007 was recorded as 45^oF at location A, 62^oF at location B and 9^oC at location C. A check with the weather stations at these three locations produced the following data for these locations over the last 15 years:

	LOCATION		
	A	**B**	**C**
Mean Temperature	40°F	71°F	13°C
Standard Deviation	9°F	6.9°F	3.6°C

Relatively speaking, which location on January 1, 2007 experienced the coolest day?

16. If the standard deviation of the sample 76, 75, 90, 2, 76, 86, 74, 79 and 71 is 26.16 Find the standard deviation of: 38, 37.5, 45, 1, 38, 43, 37, 39.5 and 35.5.

17. Using Question 16, find the standard deviation of: 79, 78, 93, 5, 79, 89, 77, 82 and 74.

18. Compare the variability of these two sets of data which are samples of silt top soil parameter under oil palm and rubber land uses:

Oil Palm: 10, 9, 5, 51, 7, 3, 11, 7 and 5.

Rubber: 9, 4, 14, 9, 27 and 10.

19. If the average silt value of the Rubber land use is 12.17 with a standard deviation of 7.94, what is the value of silt parameter with a z-score of 0.45?

20. Find the interquartile range and the semi-interquartile range of the data in Question17.

21. Find the mean deviation of the data in Question 17.

22. The following is the volume of cargo via rail between 1998 and 2006 of a particular country.

YEAR	1998	1999	2000	2001	2002	2003	2004	2005	2006
RAIL	481090	582490	100480	390542	450498	210132	147919	193438	56778

Find the standard deviation.

23. Find the coefficient of variation of the data in Question 22 above.

24. If a salesman's average profit at the end of a particular year is N250,000.00 with a standard deviation of N5,600.00 when he advertised his product with an advertising medium A, and N377,000.00 with a standard deviation of N7,400.00 when he advertised his product in another year with medium B, relatively speaking in which medium is the salesman's profit relatively higher in the years under comparison if he made a profit of N350,000.00 in December under medium A and N230,000.00 under medium B ?

25. The random variable X, has the following frequency distribution:

Classes	9.2-14.2	14.3-19.3	19.4-24.4	24.5-29.5	29.6-34.6	34.7-39.7	**TOTAL**
F	5	2	14	23	6	11	**61**

Find the population variance and standard deviation.

26. Find the sample variance and sample standard deviation of the data in Question 25.

27. Find the Mean Deviation of the Data in Question 25.

28. Find the sample variance and standard deviation for the data in Question 25 using coding method 1. You may use $X_o = 27$.

29. Find the sample variance and standard deviation for the data in Question 25 using coding method 2. You may use $X_o = 27$ and $C = 5.1$.

QUESTION 30. Show that: $S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} = \frac{C^2\sum_{i=1}^{n}\left(U_i - \overline{U}\right)^2}{n-1}$.

Where:

S^2 = Sample Variance

X_i = A random Variable, $i = 1,2,.....,n$

C = Constant(A scale parameter).

X_O = Constant (usually the mean value)

$D_i = X_i - X_o$, as defined in (A1), $i = 1,2,.....,n$

$U_i = \frac{D_i}{C} = \frac{X_i - X_o}{C}$, as defined in (C1), $i = 1,2,.....,n$

n = Sample Size.

NOTES A, B AND C

PROOF OF VARIANCES USING CODING METHODS

A. CODING METHOD 1

To show that:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} = \frac{\sum_{i=1}^{N}(D_i-\mu_D)^2}{N}$$

SHOWING:

Define:

$$D_i = X_i - X_o \tag{A1}$$

Where X_o is a constant.

$$\therefore D_i + X_o = X_i \tag{A2}$$

Thus,

$$\mu_D + X_o = \mu \tag{A3}$$

$$\therefore \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} = \frac{\sum_{i=1}^{N}\left[D_i + X_o - (\mu_D + X_o)\right]^2}{N} \tag{A4}$$

$$= \frac{\sum_{i=1}^{N}\left[D_i + X_o - \mu_D - X_o\right]^2}{N}$$

$$= \frac{\sum_{i=1}^{N}(D_i-\mu_D)^2}{N} \quad . \tag{A5}$$

Again, if one defines:

$\overline{D} + X_o = \overline{X}$ *from A2*

It can be shown as in $A4$ that:

$$S^2 = \frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{n-1} = \frac{\sum_{i=1}^{N}(D_i-\overline{D})^2}{n-1}.$$

B. TO SHOW THAT:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(D_i-\mu_D)^2}{N} = \frac{N\sum_{i=1}^{N}D_i^2 - \left(\sum_{i=1}^{N}D_i\right)^2}{N^2}$$

SHOWING:

$$\frac{\sum_{i=1}^{N}(D_i-\mu_D)^2}{N} = \frac{\sum_{i=1}^{N}\left[D_i^2 - 2\mu_D D_i + \mu_D^2\right]}{N} \tag{B1}$$

$$= \frac{\sum_{i=1}^{N}D_i^2 - 2\mu_D\sum_{i=1}^{N}D_i + N\mu_D^2}{N} \tag{B2}$$

$$= \frac{\sum_{i=1}^{N} D_i^2 - 2N\mu_D^2 + N\mu_D^2}{N} \quad \text{(B3)}$$

$$= \frac{\sum_{i=1}^{N} D_i^2 - N\mu_D^2}{N} \quad \text{(B4)}$$

$$= \frac{\sum_{i=1}^{N} D_i^2 - N\left(\frac{\sum_{i=1}^{N} Di}{N}\right)^2}{N} \quad \text{(B5)}$$

$$= \frac{1}{N}\left\{\frac{N\sum_{i=1}^{N} D_i^2 - \left(\sum_{i=1}^{N} D_i\right)^2}{N}\right\} \quad \text{(B6)}$$

$$= \frac{N\sum_{i=1}^{N} D_i^2 - \left(\sum_{i=1}^{N} D_i\right)^2}{N^2}. \quad \text{(B7)}$$

In the same way, it can be shown that:

$$S^2 = \frac{\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2}{n-1} = \frac{n\sum_{i=1}^{n} D_i^2 - \left(\sum_{i=1}^{n} D_i\right)^2}{n(n-1)}.$$

The student is required to show this.

C. CODING METHOD 2

To show that:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} = \frac{C^2\sum_{i=1}^{N}(U_i - \mu_U)^2}{N}$$

SHOWING:

Define $U_i = \frac{D_i}{C}$ (C1)

i.e. $CU_i = D_i$ (C2)

i.e. $CU_i = X_i - X_o$ (C3)

i.e. $CU_i + X_o = X_i$ (C4)

i.e. $C\mu_U + X_o = \mu$ (C5)

Substituting (C4) and (C5) into $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$, one obtains:

$$\sigma^2 = \frac{\sum_{i=1}^{N}\left[CU_i + X_o - (C\mu_U + X_o)\right]^2}{N} \quad \text{(C6)}$$

$$= \frac{\sum_{i=1}^{N}\left[CU_i + X_o - C\mu_U - X_o\right]^2}{N} \quad \text{(C7)}$$

$$= \frac{\sum_{i=1}^{N}\left[C\left(U_i - \mu_U\right)\right]^2}{N} \qquad \text{(C8)}$$

$$= \frac{C^2\sum_{i=1}^{N}\left[\left(U_i - \mu_U\right)\right]^2}{N}. \qquad \text{(C9)}$$

In the same way, it can be shown that:

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} = \frac{C^2\sum_{i=1}^{n}\left(U_i - \overline{U}\right)^2}{n-1}.$$

The student should show that:

$$\sigma^2 = C^2\left\{\frac{N\sum_{i=1}^{N}U_i^2 - \left(\sum_{i=1}^{N}U_i\right)^2}{N^2}\right\}.$$

And

$$S^2 = C^2\left\{\frac{n\sum_{i=1}^{n}U_i^2 - \left(\sum_{i=1}^{n}U_i\right)^2}{n(n-1)}\right\}.$$

following the steps in B above.

CHAPTER SIX

PROBABILITY

6.0 INTRODUCTION

Probability is a measure of chance of a possible outcome of an experiment when the experiment is repeated a number of times. This implies that, in probability theory, our interest is on experiment with statistically regular (uniform) outcomes.

6.1. Random Experiment

A random experiment is an act that results to one or several possible outcomes that can repeat themselves over time under identical conditions. Example of a random experiment is the tossing of coin a number of times. For instance, if the two sides of a coin were labelled a "Head" and a "Tail" we know that either a "Head" or a "Tail" will eventually appear but we may be uncertain which of them will appear first. If the coin is tossed over a number of times, a statistical regularity (uniformity) in the outcomes of the toss is most likely to be observed. These outcomes are then collected as a set.

6.2. Sample Space

A sample space is the collection of all possible outcomes of an experiment. For example, in an experiment involving the toss of a coin, assuming the coin is tossed three times; then the sample space Omega (Ω) can be generated using a tree diagram as shown below.

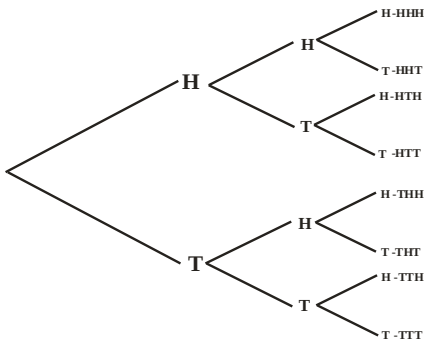


Figure 6.1

Hence, the sample space is

Ω = {(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)}

The outcomes are the individual results of a trial in the experiment, that is, (HHH), (HHT), . . .(TTT).

6.3. Event

An event is a subset of a sample space. An event may be an outcome or a group of outcomes of an experiment. An event can either be simple or compound. For example, an experiment involving either head or tail in the toss of a coin is an example of a simple event. A

compound event results from multiple experiments such as tossing the coin three or more times, or tossing a die at least two times etc.

6.3.1. Independent Events

Two events A and B are said to be independent if the occurrence of A does not prevent the occurrence of B. For example; the outcome from a toss of die cannot prevent the outcome of the occurrence of either head or tail if both die and coin are tossed together.

6.3.2 Mutually Exclusive Events

Two events A and B are said to be mutually exclusive if the occurrence of A prevents the occurrence of B at each trial. This implies that if A and B were distinct outcomes of an experiment, at each trial of the experiment, if Outcome A occurs, Outcome B will not occur and vice versa. For example, (i) the occurrence of life prevents the occurrence of death and the occurrence of death prevents the occurrence of life; (ii) the occurrence of a pass in an examination prevents the occurrence of a failure; (iii) the occurrence of a male birth prevents the occurrence of a female birth; etc. If the two events A and B are mutually exclusive, then, they have no intersection. In this case, their intersection could be described as a null set because they have nothing in common.

6.4. Set Theory

Definition: Set is defined as a collection of well-defined objects. The objects of interest must have a defined boundary. For example, a set of writing materials include: biro, pencil, and exercise books. Mathematical set include: pen, pencil, compass, protractor, sharpener, eraser and calculator. The objects contained in a set are alike. A set is denoted by capital letters, example, A, B, C, etc. and the elements of a set is denoted by small letters or figures; for example, a, b, c, etc. or 1, 2, 3, etc. The elements of a set are the points in the set. These elements are enclosed in either curly bracket called braces, that is,{ } or by a square bracket, that is, [ ].

Example 6.1:

A = {a, b, i, j, k, l}
B = {z, y, x}

The above are examples of sets. It is read as: “A” is a set that contains six elements and B is a set that contain three elements. The numbers of elements contained in a set is called the cardinality of that set. For example, the cardinality of set B above is 3 because it contains three elements, while the cardinality of set A is 6.

6.4.1. Subset

To understand this concept, let us define a new set K = {1, 2, …, 6 }. This is read as; K is a set containing, natural numbers, 1 to 6 elements; N = {3, 5}, N is a set containing two elements. Now set N is a subset of K because K contains all the elements in N. Hence, we can write this as follows: $N \subset K$ to mean N is a subset of K.

6.4.2. Superset

Since K contains all the elements in N, then, K is a superset of N written as $K \supset N$ to mean that K is a superset of N.

6.4.3. Null or Empty Set

A null or empty set does not contain any element. Null set is represented by { }. It should be noted that an empty set is a subset of every set.

6.4.4. Universal Set

The universal set is a set that contains all the elements under consideration. It is denoted by Ω.

Example 6.2:

Ω = {1, 2, …, 10}; A = {3, 5, 8}; B = {1, 5, 8, 10}. Hence, $A \subset \Omega$ and $B \subset \Omega$

6.4.5. Venn Diagram

A Venn diagram is the representation of a set in a diagram. For example, represent the set in Example 6.4 in a Venn diagram.

Solution

Ω

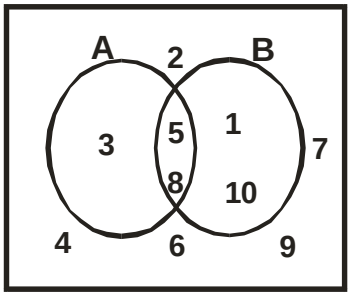


Figure 6.2

From the Venn diagram (Figure 6.2), the universal set contains all the elements under consideration. Set A and B contains 5 and 8 in common. Set A contains 3 separately and set B contains 1 and 10 separately. The rest of elements that are not contained in these two sets A and B are scattered all over the body of the Venn diagram. But none of them falls outside the Venn diagram. This is because a set has a defined boundary which any of its elements cannot exceed.

6.4.6 Complement of a Set

Using Example 6.4, the universal set Ω contains ten elements, that is, from 1 to 10 while Set A contains three elements. Then, all those elements in the universal set that are not in set A are called the complement of set A and it is denoted by A^c or A^1.

6.4.7 Equality of Two Sets

Two sets A and B are said to be equal if A is a subset of B and B is a subset of A, ie. $A = B \Rightarrow A \subset B \text{ and } B \subset A$.

Example 6.3:

Let A = {1, 3, 2} and B = {3, 2, 1}, Set A and Set B are said to be equal.

Note that the arrangement of elements in a set does not matter.

6.4.8 Set Builder and Roster Method

Set can be described using either set builder method or roster method. In Roster method, the elements of the sets are listed as we had done before this section. But in set builder method, we try to write the element of the set in a more compact form.

Example 6.4:

A = {x: x < 6}; x is a positive integer. List the elements of the set.

Solution

Since x is a positive integer (whole number), then we can list as follows;

A = {x: x = 1, 2,…, 5}.

The above is read as: "A" is a set containing x such that x is less than six.

Example 6.5:

Let B = {y: $-2 \leq y \angle 4$}. List the elements of set B, such that y is an integer.

Solution

Let B = {-2, -1, 0, 1, 2, 3}.

Example 6.6:

Let D = {x: $-1 \leq x \leq 3$} such that x is an integer . List the elements of the set D.

Solution

D = {-1, 0, 1, 2, 3}

It is important to always observe the inequality sign and its boundary so as not to exceed it or go below the range of number it specifies.

6.4.9 Basic Set Operation

6.4.9.1. Union of Two or More Sets

The union of two sets A and B denoted by $A \cup B$ is a set containing all the elements in A or B or both provided none of the elements is repeated.

Example 6.7:

A = {a, c, d} and B = {a, b, c, e, d}.

$A \cup B$ = {a, b, c, d, e}.

6.4.9.2. Intersection of Two or More Sets

The intersection of two sets A and B denoted by $A \cap B$ is a set containing all elements that are common to A and B.

Example 6.8:

Let A = {1, 2, 3, 5, 6} and B = {3, 6, 8, 10}. Find $A \cap B$.

So

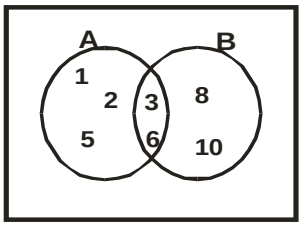


Figure 6.3

From the above Venn diagram, the intersection of the two sets A and B are {3, 6}, that is, the elements common to them. Hence, we can write $A \cap B$ = {3, 6}. This can be extended to three sets; A = {1, 2, 3, 5, 6}; B = {3, 6, 8, 10}; C = {1, 4, 6, 8, 9, 10}

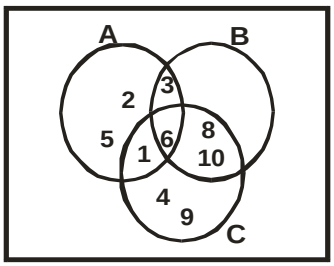


Figure 6.4

From the Venn diagram (Figure 6.4), we can see that $A \cap B = \{3,6\}$, $B \cap C = \{6,8,10\}$, $A \cap C = \{1,6\}$ and $A \cap B \cap C = \{6\}$

6.4.9.3 Application of Venn diagram in Solving Set Related Problems

Example 6.9:

In a class of 78 students, 39 students offer Mathematics and 48 students offer English Language. If 3 students offer neither of the courses, find:

(i) the number that offer both courses.
(ii) the number that offer Mathematics only.
(iii) the number that offer English only.

Solution

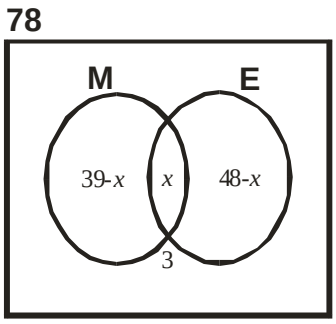


Figure 6.5

(i) In order to find out the number that offers both Mathematics and English, we solve for x.

$78 - 3 = 39 - x + x + 48 - x$

$75 = 87 - x \Rightarrow x = 12$

That is, 12 students offer both Mathematics and English Language.

(ii) The number that offer Mathematics only = $39 - x$

$= 39 - 12 = 27$ students.

(iii) The number that offer English only = $48 - x$

$= 48 - 12 = 36$ students.

6.4.9.4 Difference of Two Sets

Difference of two sets denoted by A – B (read as A difference B) is defined as a set containing all the elements in A but not in B.

Example 6.10:

Let A = {1, 2, 3, 4} and B = {1, 2, 3}.

Then A - B = {4}.

Example 6.11:

Given A = {2, 3, 5, 6} and B = {5}. Find A - B.

Solution

A - B = {2, 3, 6}.

6.4.9.5. Power Set

The power set of set A (denoted P (A)) is a set that contains all the subset of A. The number of subsets in any given set is given by 2^n, where n is the number (or cardinality) of the set under consideration.

Example 6.12:

Given A = {1, 2, 3}. List the power set of A.

Solution

P(A) = {(1), (2), (3), (1, 2), (1, 3), (2, 3), A, ϕ } $=2^3 = 8$ elements.

Example 6.13:
Given B = {5, 2}. Find the power set of B and list its elements.

Solution
The total number of subsets = $2^2 = 4$ elements.
The power set of Set B = {(5), (2), B, ϕ }.
It is important to mention here that, any set is a subset of itself and that the null set ϕ is a member of every set, that is $\{A \subset A, \quad \phi \in A\}$.

6.4.9.6. Cartesian product

The Cartesian product A x B (read as A cross B) of A and B is the set of possible pairs (x, y) where x is an element in A and y is an element in B. Suppose A = {0, 1, 2}; B = {3, 5}. We can now define A x B as a set containing the following elements.

		B	
		3	5
	0	(0,3)	(0,5)
	1	(1,3)	(1,5)
A	2	(2,3)	(2,5)

$\therefore$ A x B = {(0,3), (0, 5), (1, 3), (1, 5), (2, 3), (2, 5)}

6.5. Laws of Algebra of Set

Suppose we define three sets; A, B, and C as the subsets of a sample space denoted by Ω, then the following laws are valid.

1 Commutative Law

(a). $A \cup B = B \cup A$
(b). $A \cap B = B \cap A$

Using the same sets A and B for Associative Law, we illustrate the Commutative Law as follows;

(a). $A \cup B = \{1, 2, 3\}; \quad B \cup A = \{1, 2, 3\}$. Hence, $A \cup B = B \cup A$
(b). $A \cap B = \{2\}; \quad B \cap A = \{2\}$. Hence, $A \cap B = B \cap A$

In commutative law, the order of arrangement does not matter (i.e. the order of arrangement is immaterial).

2 Associative Law

(a). $(A \cup B) \cup C = A \cup (B \cup C)$
(b). $(A \cap B) \cap C = A \cap (B \cap C)$

Illustration of the above law: A = {1, 2}; B = {2, 3} and C = {3, 6}. Then
(a). $(A \cup B) \cup C = \{1, 2, 3\} \cup \{3, 6\} = \{1, 2, 3, 6\}$; and
$A \cup (B \cup C) = \{1, 2\} \cup \{2, 3, 6\} = \{1, 2, 3, 6\}$
Hence, $(A \cup B) \cup C = A \cup (B \cup C)$

(b). $(A \cap B) \cap C = \{2\} \cap \{3, 6\} = \phi$ and $A \cap (B \cap C) = \{1, 2\} \cap \{3\} = \phi$.
Hence $(A \cap B) \cap C = A \cap (B \cap C)$.

3 Identity Law

Let K to be a set defined in a sample space Ω, then we have the following

(a). $K \cup \phi = K$
(b). $K \cup \Omega = \Omega$
(c). $K \cap \phi = \phi$
(d). $K \cap \Omega = K$

Illustration:

Let K = {1, 2, 3, 6}; Ω = {1, 2, …, 8} and ϕ = {}. Then

(a). $K \cup \phi$ = {1, 2, 3, 6} = K
(b). $K \cup \Omega$ = {1, 2, 3, 4, 5, 6, 7, 8} = Ω
(c). $K \cap \phi$ = {} = ϕ
(d). $K \cap \Omega$ = {1, 2, 3, 6} = K

4 Complement Law

Let Ω to be the sample space and M be a set defined in Ω. Ω = {1, 2, …, 8}; M = {1, 3, 6, 8}. The complement of Set M denoted as M^c is a set containing elements that do not belong to Set M but belongs to the universal set Ω.

Hence,

(a). $M \cup M^c = \Omega$
(b). $M \cap M^c = \phi$
(c) $(M^c)^c = M$
(d) $\Omega^c = \phi$
(d). $\phi^c = \Omega$

Illustration:

M^c = {2, 4, 5, 7}. Hence

(a). $M \cup M^c$ = {1, 2, 3, 4, 5, 6, 7, 8} = Ω
(b). $M \cap M^c$ = {} = ϕ
(c). $(M^c)^c = M$.
(d). $\Omega^c = \phi$.
(e). $\phi^c = \Omega$.

Example 6.14: Let Ω = {1, 2, …, 10} A = {3, 5, 8}.
∴ A^c = {1, 2, 4, 6, 7, 9, 10}.
Observe that $A \cup A^c = \Omega$.

5 De Morgan's Law

Let N and P be two sets defined in Ω

(a). $(N \cup P)^c = N^c \cap P^c$
(b). $(N \cap P)^c = N^c \cup P^c$

Illustration:

Let Ω = {1, 2, …, 8}; N = {2, 4, 5, 8}; P = {1, 2, 3, 5, 7}; N^c = {1, 3, 6, 7} and P^c = {4, 6, 8}. Hence

(a). $(N \cup P)$ = {1, 2, 3, 4, 5, 7, 8}

$(N \cup P)^c = N^c \cap P^c = \{6\}$

(b). $(N \cap P)^c = N^c \cup P^c$

$(N \cap P)$ = {2, 5}

$\therefore$ $(N \cap P)^c$ = {1, 3, 4, 6, 7, 8}

$N^c \cup P^c$ = {1, 3, 4, 6, 7, 8}

$\therefore (N \cap P)^c = N^c \cup P^c$ = {1, 3, 4, 6, 7, 8}

6. Distributive Law

Suppose we define three sets; A, B, and C in a sample space Ω, then we have the following.

(a). $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(b). $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Illustration of the above law: let A = {1, 2}; B = {2, 3} and C = {3, 6}. Then

$(A \cup B) = \{1, 2, 3\}; (A \cup C) = \{1, 2, 3, 6\}; (B \cap C) = \{3\}$

(a). $A \cup (B \cap C) = \{1, 2, 3\}; (A \cup B) \cap (A \cup C) = \{1, 2, 3\}$. Hence,

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(b). $(B \cup C) = \{2, 3, 6\}$; $A \cap (B \cup C) = \{2\}$; $(A \cap B) = \{2\}$; $(A \cap C) = \phi$

$(A \cap B) \cup (A \cap C) = \{2\}$. Hence, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

6.6. Set function (cardinality of a set): The cardinality of a set written as n(A) is the number of elements contained in a set. For example;

(a). Let A = {0, 2, 4, 6} $\therefore$ n(A) = 4

(b). Let B = {x : 0 < x < 1} $\therefore$ n(B) = ∞

(6). If A = {1, 2, 3, 4, 5, 6, 7} and B = {2, 4, 6}, then

$(A - B) = (AB^c) = (A) - (AB) = \{1, 3, 5, 7\}$

Therefore, $n(A - B) = 4$

6.7. Definitions of Probability

Probability can be defined from three perspectives, namely; classical definition, frequency definition and axiomatic definition.

6.7.1. Classical definition: If there are n mutually exclusive and equally likely outcomes, of which one outcome with a particular attribute must occur. Suppose, the occurrence of such attribute is regarded as a "success" (s) and n(s) as the number of successes, then the probability of a "success" is given by

Probability of success = $\frac{n(s)}{n}$.

Example 6.15: A box contains 6 red balls, 4 black and 5 blue balls. A ball is picked at random without replacement. What is the probability that
(i). one red is picked? (ii). 2 blue balls are picked?

Solution:
6 red balls + 4 black balls + 5 blue balls = 15 balls
(i). P(one red ball) = $\frac{6}{15} = \frac{2}{5}$
(ii). P(2 blue balls) = $\frac{5}{15} \times \frac{4}{14} = \frac{2}{21}$

Example 6.16: The probability that a student will pass a Statistics examination is $\frac{3}{4}$.

(i). what is the probability that he fails the examination?

Solution

P(a student passes) = $\frac{3}{4}$

P(that a student fails) = 1 – P(that a student passes)

= 1 - $\frac{3}{4}$ = $\frac{1}{4}$

Example 6.17: Suppose that a card is drawn at random from an ordinary deck of playing cards. Find the probability of drawing a spade.

Solution:
An ordinary deck usually contains 52 cards of which 13 are usually spaded.

P(of drawing a spade) = $\frac{13}{52} = \frac{1}{4}$

Example 6.18: What is the probability of getting a head (H) and a five if both die and a coin are tossed together?

Solution

P(H) = $\frac{1}{2}$

P(a five) = $\frac{1}{6}$

P(1 head and a 5) = $\frac{1}{2}$ x $\frac{1}{6}$ = $\frac{1}{12}$

Example 6.19: In a Department of Computer Science, students offer one of these subjects; Mathematics and English language. The probability that students offer Mathematics only is $\frac{2}{5}$ and the probability that students offer English only is $\frac{1}{5}$. Find the probability that the students offer either of the two courses; Mathematics or English if the events are mutually exclusive.

Solution

P(M) = $\left(\frac{2}{5}\right)$; P(E) = $\left(\frac{1}{5}\right)$;

The two events are mutually exclusive. Hence

$P(M \cup E) = P(M) + P(E)$

$P(M \cup E) = \frac{2}{5} + \frac{1}{5} = \frac{3}{5}$

6.7.2 Frequency Definition

The probability of an event is the proportion of the number of times that the event will occur in the long run.

Example 6.20: If records show that 294 of 300 ceramic insulators tested were able to withstand a certain thermal shock, what is the probability that any one such insulator will be able to withstand the thermal shock?

Solution:

The probability that any such insulator will withstand the thermal shock is $\frac{294}{300} = 0.98$ and this represents an estimate of the probability.

Example 6.21: A coin is tossed three times. Define the random variable X to be the number of heads. Obtain: (i). the sample space Ω using tree diagram (ii). the probability of getting at least 2 heads (iii). the probability of getting at most 2 heads.

Solution:

Tree diagram showing results of tossing a coin three times

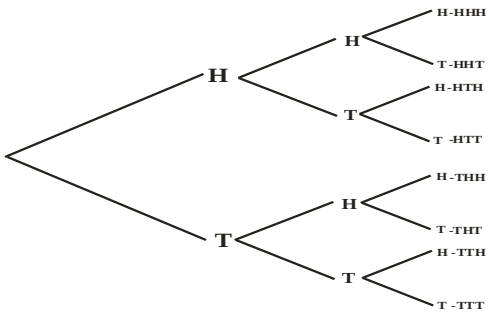


Figure 6.6

(i). Ω = {(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)}

(ii) $P(X \geq 2) = P(X = 2 \text{ or } X = 3) = P\{(HHH), (HHT), (HTH), (THH)\} = \frac{4}{8} = \frac{1}{2}$

(iii). $P(X \leq 2) = P(X = 0 \text{ or } X = 1 \text{ or } X = 2) = P\{(TTT), (HHT), (HTH), (HTT), (THH), (THT), (TTH)\}$

$= \frac{7}{8}$

6.7.3 Axiomatic Definition of Probability

One of the aims of science is to predict and describe events in the world in whichpo we live. One way in which this is done is to construct mathematical models which adequately describe the real world. Hence, axiomatic definition of probability is the mathematical representation of probability as the probability of an event or subset in a sample space. Let A be an event defined in a sample Ω. The following axioms hold:

(1) $0 \le P(A) \le 1$; That is, the probability of event A lies between zero and one.

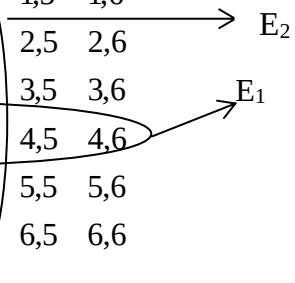
(2) $P(\Omega) = 1$.

(3) $P(A \cup B) = P(A) + P(B)$; if A and B are mutually exclusive events.

Example 6.22: Find the probability of a 4 turning up at least once in two tosses of a fair die.

Solution: **Method 1**:

	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6



Let E_1 be the event "4" on first toss and E_2 be the event "4" on second toss. $(E_1 \cup E_2)$ = event "4" on first toss or "4" on second toss or both = event that at least 4 turns up. We require $\Pr(E_1 + E_2)$.

$$E_1 = \{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}$$
$$E_2 = \{(1,4),(2,4),(3,4),(4,4),(5,4),(6,4)\}$$
$$E_1 \cup E_2 = \{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),(1,4),(2,4),(3,4),(5,4),(6,4)\}$$
$$P(E_1) = \frac{6}{36},\quad P(E_2) = \frac{6}{36},\quad P(E_1 \cup E_2) = \frac{11}{36}$$

Method 2:

$$E_1 \cap E_2 = \{(4,4)\},\quad P(E_1 \cap E_2) = \frac{1}{36}$$
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}$$

Example 6.23: Two dice are tossed once. Obtain
(i). the sample space (Ω). (ii). the probability of a sum of 3. (iii). the probability of a sum of 6.

Solution

Sample space of tossing a die twice

	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

(i). Sample space, Ω = {(1,1), (1, 2), (1, 3), . . . ,(6, 6)}; n(Ω) = 36

(ii). P(Sum of 3) =. $\frac{2}{36} = \frac{1}{18}$

(iii). P(Sum of 6) $= \frac{5}{36}$

6.8. Conditional probability

Let A and B be any two events defined on a sample space Ω. The conditional probability that event A will occur given that B has already occurred (i.e. P(A/B)) is defined as

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

If A and B are independent events, then

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

This implies that the occurrence of an event A is independent of the occurrence of event B.

But if A and B are mutually exclusive events, then $A \cap B = 0$,

$P(A \cap B) = P(\phi) = 0$

Example 6.24: Two students of A and B went for a written interview. Student A has 25% chance of passing while student B has 45% chance of passing. Both students have 18% chances of passing.

(i). If B has passed the examination, what is the probability that A will also pass?

(ii).What is the probability that neither A nor B passed the examination?

Solution

P(A) = 25% $= \frac{1}{4}$

P(B) = 45% $= \frac{9}{20}$

P$(A \cap B)$ = 18% $= \frac{9}{50}$

(i). $P(A|_{B}) = \frac{P(A \cap B)}{P(B)}$

$$= \frac{\left(\frac{9}{50}\right)}{\left(\frac{9}{20}\right)} = \frac{2}{5}$$

(ii). $P(A^c \cap B^c) = P(A \cup B)^c = 1 - P(A \cup B)$
$= 1 - [P(A) + P(B) - P(A \cap B)$
$= 1 - [0.25 + 0.45 - 0.18]$
$= 1 - 0.52 = 0.48$

Therefore, the probability that neither A nor B pass the examination = 0.48 = 48%.

Example 6.25 In a certain city, 40% of the inhabitant have black skin, 25% have one eye while 15% have both black skin and one eye. A person is chosen at random from the city.
(i). If he has black skin, what is the probability that he also has one eye?
Solution

Let the event that some of the inhabitants have one eye be A; and the events that some of the inhabitants have black skin be B.

(i) $P(A|_B) = \frac{P(A \cap B)}{P(B)} = \frac{0.15}{0.4} = 0.375$

6.9. Bayes Theorem and Partitioning

Suppose the events $A_1, A_2, \ldots, A_n$ form a sample space Ω, that is, the events A_i are mutually exclusive but collectively exhaustive and their union is Ω. Let B be any other event define on a sample space Ω. Then, the Bayes theorem states that

$$P(A_i / B) = \frac{P(A_i) \cdot P(B / A_i)}{\sum_{i=1}^{n} P(A_i).P(B / A_i)}$$

Proof:

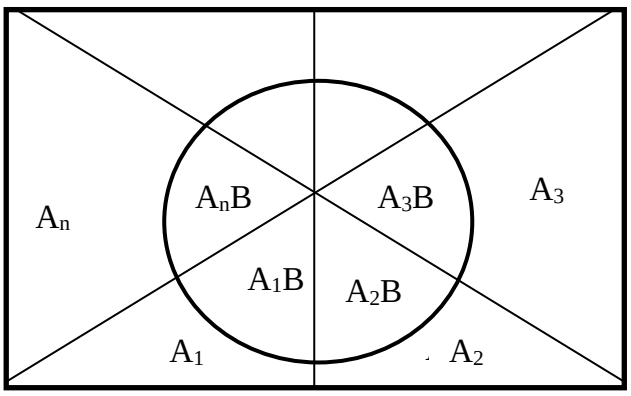


Figure 6.7

From the diagram, B is the same as
$\Omega = A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n$

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B) \cup ..... \cup (A_n \cap B)$$

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) + ... + P(A_n \cap B)$$

$$P(B) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + ... + P(A_n)P(B/A_n) = \sum_{i=1}^{n} P(A_i)P(B/A_i)$$

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B/A_i)$$

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)}$$

6.9.1. Application of Bayes' Theorem

Example 6.26. In a factory, a certain brand of chocolate is packed into boxes on four different production lines, A_1, A_2, A_3 and A_4. Records show that a small percentage of boxes are not packed properly for sales; 1% from A_1, 3% from A_2, 2.5% from A3 and 2% from A_4. If the percentages of the total output that have come from the production lines are: 35% from A_1 , 20% from A_2, 24% from A_3 and 21% from A_4.

(i).What is the probability that a box chosen at random from the whole output is faulty?

(ii).What is the probability that a faulty box comes from the first production line?

Solution

Let A_i denote the event that a box chosen at random from the whole output comes from a production line i; i = 1, 2, 3, 4. Let F denote a box chosen at random from the whole output is faulty. We want P(F) and $P(A_i/F)$

$$F = (F \cap A_1) \cup (F \cap A_2) \cup (F \cap A_3) \cup (F \cap A_4)$$
$$P(F) = P(F \cap A_1) + P(F \cap A_2) + P(F \cap A_3) + P(F \cap A_4)$$
$$P(F) = P(F/A_1)P(A_1) + P(F/A_2)P(A_2) + P(F/A_3)P(A_3) + P(F/A_4)P(A_4)$$
$$= (0.01)(0.35) + (0.03)(0.20) + (0.025)(0.24) + (0.02)(0.21)$$
$$\therefore P(F) = 0.0035 + 0.006 + 0.006 + 0.0042 = 0.0197$$

(ii). $$P(A_1/F) = \frac{P(A_1 \cap F)}{P(F)} = \frac{P(A_1/F)P(F)}{\sum_{i=1}^{n} P(A_1/F)P(F)}$$

$$\therefore P(A_1/F) = \frac{(0.35)(0.01)}{(0.0197)} = 0.1777$$

Example 6.27. In a certain factory that manufactures the lion bulb, machine A_1, A_2, and A_3 manufacture 40%, 35% and 25% respectively of the total production. The percentages of defective bulbs are 2, 4, and 5. A bulb is selected at random from a day's production and it was found to be defective. What is the probability that it is manufactured by;

(a). (i). A_1 and (ii). A_2

(b). what is the probability of getting a defective bulb in a day's production?

Solution

Let $P(A_i)$ denote the probability that a bulb is manufactured by machine A_i and let $P(D/A_i)$ denote the probability that a defective bulb comes from machine A_i ; i = 1, 2, 3.

$P(A_1)$	$40/100$	$P(D/A_1)$	2/100
$P(A_2)$	$35/100$	$P(D/A_2)$	4/100
$P(A_3)$	$25/100$	$P(D/A_3)$	5/100

(a). (i). The probability that a defective bulb is produced by machine A_1 is

$$P(D/A_1) = \frac{P(A_1).P(D/A_1)}{P(A_1)P(D/A_1)+P(A_2)P(D/A_2)+P(A_3)P(D/A_3)}$$

$$\therefore P(D/A_1) = \frac{\left(\frac{40}{100}\right)\times\left(\frac{2}{100}\right)}{\left(\frac{40}{100}\right)\times\left(\frac{2}{100}\right)+\left(\frac{35}{100}\right)\times\left(\frac{4}{100}\right)+\left(\frac{25}{100}\right)\times\left(\frac{5}{100}\right)} = \frac{16}{69}$$

(ii). the probability that a defective bulb is produced by machine A_2 is

$$P(D/A_2) = \frac{\left(\frac{35}{100}\right)\times\left(\frac{4}{100}\right)}{\left(\frac{40}{100}\right)\times\left(\frac{2}{100}\right)+\left(\frac{35}{100}\right)\times\left(\frac{4}{100}\right)+\left(\frac{25}{100}\right)\times\left(\frac{5}{100}\right)} = \frac{28}{69}$$

(b). the probability of getting a defective bulb is

$$P(D) = P(A_1)P(D/A_1)+P(A_2)P(D/A_2)+P(A_3)P(D/A_3)$$
$$= \left(\frac{40}{100}\right)\times\left(\frac{2}{100}\right)+\left(\frac{35}{100}\right)\times\left(\frac{4}{100}\right)+\left(\frac{25}{100}\right)\times\left(\frac{5}{100}\right)$$
$$\therefore P(D) = \frac{69}{2000}$$

Example 6.28. Box 1 contains 2 red, 3 white and 5 blue balls; box 2 contains 4 red, 1 white and 3 blue balls; while box 3 contains 3 red, 4 white and 3 blue balls. The three boxes are identical in appearance. A box is selected from which a ball is selected at random and it is observed to be red.
(a). what is the probability that box 2 was selected?
(b). what is the probability of selecting a white ball?

Solution

Let $P(B_i)$ be the probability of selecting box i , since the three boxes are identical, then

$$P(B_1) = P(B_2) = P(B_3) = \frac{1}{3}$$

Also let $P(R/B_i)$ denote the probability that a selected red ball is from box i; i = 1, 2, 3.

Balls	Box_1	Box_2	Box_3
Red	2	4	3

White	3	1	4
Blue	5	3	3
Total	10	8	10

(a). $P(R/B_1)=\frac{2}{10};\quad P(R/B_2)=\frac{4}{8};\quad P(R/B_3)=\frac{3}{10}$

$$P(B_2/R)=\frac{P(B_2).P(R/B_2)}{P(B_1)P(R/B_1)+P(B_2)P(R/B_2)+P(B_3)P(R/B_3)}$$

$$\therefore P(B_2/R)=\frac{\left(\frac{1}{3}\right)\times\left(\frac{4}{8}\right)}{\left(\frac{1}{3}\right)\times\left(\frac{2}{10}\right)+\left(\frac{1}{3}\right)\times\left(\frac{4}{8}\right)+\left(\frac{1}{3}\right)\times\left(\frac{3}{10}\right)}=\frac{1}{2}$$

(b). $P(W)=P(B_1)P(W/B_1)+P(B_2)P(W/B_2)+P(B_3)P(W/B_3)$

$$\therefore P(W)=\frac{1}{3}\times\frac{3}{10}+\frac{1}{3}\times\frac{1}{8}+\frac{1}{3}\times\frac{4}{10}=\frac{11}{40}$$

Example 6.29. The probability that John travels from Kaduna to Lagos by plane, car or by train are $\frac{1}{4},\frac{1}{6},\frac{2}{3}$ respectively. If the probabilities of accident when he uses these means of transport are $\frac{1}{12},\frac{1}{4},\frac{1}{8}$ respectively,

(a). what is the probability of an accident?

(b).what is the probability that John was travelling by train and it is known that an accident must have happened?

Solution

Let P(A) denote the probability of travelling by plane. Let P(C) denotes the probability of travelling by car. Let P(T) denote the probability of travelling by train.

$P(A)=\frac{1}{4};P(C)=\frac{1}{6};P(T)=\frac{2}{3}$ and

Let P(D/A) be the probability of accident by plane. Let P(D/C) be the probability of accident by car. Let P(D/T) be the probability of accident by train.

P(D/A) = $\frac{1}{12}$; P(D/C) = $\frac{1}{4}$; P(D/T) = $\frac{1}{8}$;

(a). the probability of an accident is

$$P(D)=P(A)P(D/A)+P(C)P(D/C)+P(T)P(D/T)$$

$$\therefore P(D)=\frac{1}{4}\times\frac{1}{12}+\frac{1}{6}\times\frac{1}{4}+\frac{2}{3}\times\frac{1}{8}=\frac{7}{48}$$

(b). the probability of travelling by train knowing that accident will occur is

$$P(T/D) = \frac{P(T)P(D/T)}{P(D)}$$

$$\therefore P(T/D) = \frac{\frac{2}{3}\times\frac{1}{8}}{\frac{7}{48}} = \frac{4}{7}$$

Example 6.30. In a daily production of high quality tooth brushes manufactured by PZ industry, it was discovered that machines A, B, C manufactured 50%, 30%, and 20% respectively of the total production. The percentages of defective brushes manufactured by these machines are 3, 4, and 4. A brush is selected at random from a day's production and is found to be defective.

(a). what is the probability of getting a defective brush?
(b). what is the probability that the defective brush is manufactured by machine; A, B, C.

Solution

Let $P(A_i)$ denote the probability that a bulb is produced by machine A_i ; i = A,B,C. Let $P(D/A_i)$ be the probability that a defective brush is produced by machine A_i.

$$P(A) = \frac{50}{100};\quad P(D/A) = \frac{3}{100};\quad P(B) = \frac{30}{100};\quad P(D/B) = \frac{4}{100};\quad P(C) = \frac{20}{100};\quad P(D/C) = \frac{4}{100}$$

Let P(D) denote the probability of getting a defective brush.

$$P(D) = P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C)$$

$$\therefore P(D) = \frac{50}{100}\times\frac{3}{100} + \frac{30}{100}\times\frac{4}{100} + \frac{20}{100}\times\frac{4}{100} = \frac{7}{200}$$

(b). (i) Let the probability that the defective brush is manufactured by machine A be P(A/D).

$$P(A/D) = \frac{P(A)P(D/A)}{\left(\frac{7}{200}\right)}$$

$$\therefore P(A/D) = \frac{\left(\frac{50}{100}\times\frac{3}{100}\right)}{\left(\frac{7}{200}\right)} = \frac{3}{7}$$

(ii).
$$P(B/D) = \frac{P(B)P(D/B)}{\left(\frac{7}{200}\right)}$$

$$\therefore P(B/D) = \frac{\left(\frac{30}{100}\times\frac{4}{100}\right)}{\left(\frac{7}{200}\right)} = \frac{12}{35}$$

(iii). $$P(C/D) = \frac{P(C)P(D/C)}{\left(\frac{7}{200}\right)}$$

$$\therefore P(B/D) = \frac{\left(\frac{20}{100} \times \frac{4}{100}\right)}{\left(\frac{7}{200}\right)} = \frac{8}{35}$$

6.10. Random Variable:

Random variable is a real valued function (X) that assigns values to the points in the sample space (Ω). Random variables are represented by capital or upper case letters while the values they assume in the sample space is represented by small or lower case letters. We can write $X = x_i$ to be interpreted as "the random variable X assumes the values x_i, where i = 1, 2, . . . , n". Random variables can broadly be classified into **discrete** and **continuous** random variables. A random variable is said to be discrete if it can be assigned a whole number or if it takes values on a set of integers. For example, the number of heads that appear if a coin is tossed n time(s), the number of customers that came into a restaurant at a given period of time, the number of telephone calls made by a subscriber at a given period of time, etc. On the other hand, a continuous random variable is a random variable that cannot be assigned a whole number. Such random variables take their values on a real number line. Variables such as volume of liquid, gas, lifetime of electric bulb, etc, are continuous random variables.

6.11. Probability Distributions

We have defined and classified random variables into discrete and continuous random variables. So we also have discrete probability distributions and continuous probability distributions. The probability function associated with a discrete random variable is called probability mass function (pmf) and the probability function associated with continuous random variable is called the probability density function (pdf). In this section, we shall be concerned with only three discrete probability distributions (Bernoulli; Binomial and Poisson distributions) and one continuous probability distribution (Normal distribution).

6.11.1. Probability Distribution Function of Random Variable

Let X be a random variable on a sample space Ω with a finite image set say $X(\Omega) = \{x_1, x_2, \ldots, x_n\}$.We put $X(\Omega)$ into a probability space by defining the probability of x to be P(X= x) which we write as P(x) or f(x). The function defined by P(x) = P(X= x) which assigns probability measure to every possible value of discrete random variable X is called the probability mass function (pmf). The function defined by f(x), and that assigns probability measures to every possible value of a continuous random variable X in the sample space is called the probability density function (pdf).

6.11.2. Properties of Discrete Random Variables

1. $P(x_i) \geq 0$ for all x

2. $\sum_{i=1}^{n} P(x_i) = 1$
3. $0 \le P(x_i) \le 1$

6.11.3. Properties of Continuous Random Variables

1. f(x) ≥ 0 for all values of x
2. $\int_{Rx} f(x)dx = 1$
3. $0 \le \int_{Rx} f(x)dx \le 1$

Note:
Given a random variable X with pmf / pdf as $P(x)/f(x)$, the mean and variance are given as follows:

$$\mu = E(X) = \begin{cases} \sum_x xP(x), & X \text{ is discrete} \\ \int_x xf(x)dx, & X \text{ is continuous} \end{cases}$$

$$E(X^2) = \begin{cases} \sum_x x^2 P(x), & X \text{ is discrete} \\ \int_x x^2 f(x)dx, & X \text{ is continuous} \end{cases}$$

Variance = $\sigma^2 = E(X^2) - (\mu)^2$

6.11.4. Three Discrete Probability Distribution Functions

6.11.4.1. Bernoulli Distribution

If an experiment will result in either success or failure, then, the single trial of such experiment is called Bernoulli trial and the resultant distribution from such trial is called Bernoulli distribution. It is denoted by

$$P(X = x) = p^x(1-p)^{1-x}; \quad x = 0,1$$
$$= p^x q^{1-x}; \qquad \text{where} \quad q = 1-p$$

$$P(X = 1) = p^1(1-p)^0 = p$$
$$P(X = 0) = p^0(1-p)^1 = 1-p = q$$

where 1 = success and 0 = failure.
The mean (expected value) and variance of Bernoulli distribution are p and pq respectively.

6.11.4.2. Binomial Distribution

The n repeated independent Bernoulli trial result in Binomial experiment and the resultant distribution is called Binomial distribution. It is defined as

$$P(X=x)=\begin{cases}\binom{n}{x}p^x q^{n-x} & ;\ x=0,1,2,.....,n\\ 0, \ otherwise\end{cases}$$

where p and q retain their definitions as in Bernoulli distribution; n is the number of trials and $\binom{n}{x}=\frac{n!}{x!(n-x)!}$ is a combinatorial sign called binomial coefficient. A random variable X that follows binomial with n, p parameters is denoted as b(n,p).

The mean (expected value) and variance of Bernoulli distribution are np and npq respectively.

Example 6.31: Let X be the number of successes in n independent trial of a random experiment. If the probability of success P = $\frac{3}{5}$ and n = 3, generate the probability distribution of X and hence find the mean and variance of X.

Solution

$X \sim b\left(3,\frac{3}{5}\right)$ and $P(x)=\binom{n}{x}p^x q^{n-x}$; x = 0, 1, 2,3.

$P(x)=\binom{3}{x}\left(\frac{3}{5}\right)^x\left(\frac{2}{5}\right)^{3-x}$; x = 0, 1,2, 3.

Now

$$P(X=0)=\binom{3}{0}\left(\frac{3}{5}\right)^0\left(\frac{2}{5}\right)^3=\left(\frac{2}{5}\right)^3=\frac{8}{125}$$

$$P(X=1)=\binom{3}{1}\left(\frac{3}{5}\right)^1\left(\frac{2}{5}\right)^2=3\left(\frac{3}{5}\right)\left(\frac{4}{25}\right)=\frac{36}{125}$$

$$P(X=2)=\binom{3}{2}\left(\frac{3}{5}\right)^2\left(\frac{2}{5}\right)=3\left(\frac{9}{25}\right)\left(\frac{2}{5}\right)=\frac{54}{125}$$

$$P(X=3)=\binom{3}{3}\left(\frac{3}{5}\right)^3\left(\frac{2}{5}\right)^0=\left(\frac{3}{5}\right)^3=\frac{27}{125}$$

The probability distribution of X is given thus

X	0	1	2	3	Total
$P(x)$	$\frac{8}{125}$	$\frac{36}{125}$	$\frac{54}{125}$	$\frac{27}{125}$	100

Mean $= E(X)=\sum_{x=0}^{n} x.P(x)$

$$=0\left(\frac{8}{125}\right)+1\left(\frac{36}{125}\right)+2\left(\frac{54}{125}\right)+3\left(\frac{27}{125}\right)$$

$$E(X)=\frac{36}{125}+\frac{108}{125}+\frac{81}{125}=\frac{9}{5}=1.8$$

Variance = Var(X)

$Var(X) = E(X^2) - [E(X)]^2$

But $E(X^2) = \sum_{x=0}^{n} x^2.P(x)$

$= 0^2\left(\frac{8}{125}\right) + 1^2\left(\frac{36}{125}\right) + 2^2\left(\frac{54}{125}\right) + 3^2\left(\frac{27}{125}\right)$

$E(X^2) = \frac{36}{125} + \frac{216}{125} + \frac{243}{125} = \frac{495}{125} = \frac{99}{25}$

$Var(X) = \frac{99}{25} - \frac{81}{25} = \frac{18}{25}$

Example 6.32: Assuming that the ratio of male children to female is $\frac{1}{2}$, find the probability that in a family of five children: (a). all children will be of the same sex. (b). the three oldest will be boys and the 2 youngest will be girls.

Solution

P(B) = P(G) = $\frac{1}{2}$

$P(x) = \binom{n}{x} p^x q^{n-x}$; x = 0, 1, . . . , 5.

But p = q = $\frac{1}{2}$

$P(X = x) = \binom{n}{x}\left(\frac{1}{2}\right)^n$

(a). $P(X = 5) = \binom{5}{5}\left(\frac{1}{2}\right)^5 = \frac{1}{32}$

(b). since P(B) = P(G)

P(X = 3, Y = 2) = P(X = 3).P(Y = 2)

$= \binom{5}{3}\left(\frac{1}{2}\right)^5 \cdot \binom{5}{2}\left(\frac{1}{2}\right)^5 = 0.09766$

6.11.4.3. Poisson Distribution

Poisson random variables occur naturally. Examples of Poisson random variables are: the number of telephone calls arriving at a switchboard per unit time; the number of passengers boarding a given vehicle at a given time; the number of cars passing through a roundabout at a given time; the number of customers that arrive in a bank at a given time; the number of alpha particles from a radioactive source, etc. Hence, when the probability of success is very small and the number of trial is large, then, we use Poisson distribution to approximate the

Binomial distribution. The probability mass function (pmf) of Poisson distribution is given as

$$P(x)=\begin{cases}\dfrac{\lambda^x e^{-\lambda}}{x!}; & x=0,1,2,.....\\ 0, \ otherwise\end{cases}$$

where x = 0, 1, 2, . . .; $e^{-\lambda}$ = constant; λ = the rate of occurrence or the mean occurrence.

6.11.4.3.1. Properties of Poisson Random Variable

(1). Each Poisson random variables occur naturally

(2). No two events can occur together

(3). Its mean and variance are the same $E(X)=\lambda$ and var = λ

Example 6.33: The number of errors made by a typist has a Poisson distribution with mean of 4 per sheet of paper. (i). Find the probability of at least one error in each page. (ii). Find the probability of at most three errors in each page.

Solution

(i). Let X be the number of errors

$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0)$

$$P(x)=\frac{\lambda^x e^{-\lambda}}{x!}$$

$$=1-\frac{4^0 e^{-4}}{0!}=1-e^{-4}=1-0.0183156$$

$$\therefore P(X \geq 1)=0.982$$

(ii). $P(X \leq 3) = [P(X = 0) + P(X = 1) + P(X = 2) + P(X=3)]$

$$= e^{-4} + 4e^{-4} + \frac{4^2 e^{-4}}{2!} + \frac{4^3 e^{-4}}{3!}$$

$$= e^{-4}(1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!}) = e^{-4}(1 + 4 + 8 + 10.6667)$$

$P(X \leq 3) = e^{-4}(23.6667) = 0.018316(23.6667) = 0.43347$

Example 6.34: The number of telephone calls arriving at a switchboard per minute is a Poisson random variable having parameter $\lambda = 4$. Find the probability that during any minute the switchboard will record: (i). exactly 3 calls. (ii). more than 2 calls

Solution:

Let X denotes the number of calls recorded in a minute. Then the pmf is

$P(x)=\frac{\lambda^x e^{-\lambda}}{x!}=\frac{4^x e^{-4}}{x!}$; x = 0, 1, 2, . . .

(i). $P(X=3)=\frac{4^3 e^{-4}}{3!}=10.667(0.0183156)=0.1954$

(ii). $P(X>2)=1-P(X\leq 2)$

$$= 1 - P(X = 0,1,2) = 1 - e^{-4}[\frac{4^0}{0!} + \frac{4^1}{1!} + \frac{4^2}{2!}]$$
$$= 1 - e^{-4}[1 + 4 + 8] = 0.7619$$

6.11.4.4. Normal Distribution

The Normal distribution was first published by Abraham De Moivre in 1733 as an approximation for the distribution of the sum of the binomial random variable. It is the single most important distribution in probability and statistics. If a random variable is normally distributed with parameters μ and σ^2, then the density function pdf is given as

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}; -\infty < x < \infty; -\infty < \mu < \infty; 0 < \sigma < \infty.$$

π and e are constants having the values 3.142 and 2.71828 respectively. Hence, we write $X \sim N(\mu, \sigma^2)$ where μ is the population mean and σ^2 is the population variance. But if the sample population is large ($n \geq 30$) and μ and σ^2 were unknown, a sample mean $\bar{x}$ and sample variance S^2 can be used to estimate their respective population parameters, μ and σ^2. The values of normal distribution are tabulated in the statistical table where it can be read for various statistical inferences.

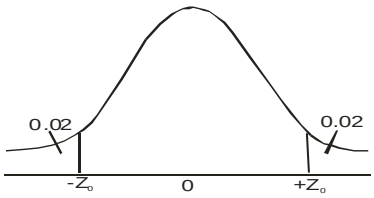
6.11.4.4. 1 Properties of Normal Distribution

1. The shape of the normal distribution curve (graph) is bell-shaped and it is symmetrical about the mean. That is, one of the half of the curve is a mirror image of the other. Its mean, mode and median coincide and are equal.
2. The area under the curve is equal to one
3. The curve approaches but never touches the horizontal axis
4. E(x) = μ and var(x) = σ^2
5. If $x \sim N(\mu, \sigma^2)$ then the random variable
6. $Z = \frac{x - \mu}{\sigma} \sim N(0,1)$

Z is called the standard normal random variable with parameters 0 and 1, where E(Z) = 0 and Var(Z) = 1

Example 6.35: Suppose that the actual amount of coffee which a filling machine puts into a food processor's 6-oz-cans varies from can to can and it can be looked upon as a random variable having a normal distribution with a standard deviation 0.04 oz. if only 2% of the cans are to contain less than 6-oz of coffee. What will be the average amount which the filling machine puts into these cans?

Solution



$P(Z) = 0.5 - 0.02 = 0.48$

The value of Z at 0.48 read from the standardized normal table is 2.05 and x = 6. Hence,

$$Z = \frac{X-\mu}{\sigma} \Rightarrow 2.05 = \frac{6-\mu}{0.04}$$
$$\therefore \mu = 6.082$$

Example 6.36: If X has a normal distribution with mean 5 and standard deviation 1, find the probability that X lies between 4.5 and 6.5.

Solution

$P(4.5 \le X \le 6.5) = P(Z_1 \le Z \le Z_2)$

$$P\left(\frac{4.5-6.5}{1} \le Z \le \frac{6.5-5}{1}\right) = P(-0.5 \le Z \le 1.5)$$
$$= P(-0.5 \le Z \le 0) + P(0 \le Z \le 1.5)$$
$$= 0.1915 + 0.4332 = 0.6247$$

Example 6.37: X ~ N(400, 40,000). Find the probability that X is at most 800.

Solution

$$Z = \frac{X-\mu}{\sigma} = \frac{800-400}{200} = 2$$
$$P(Z \le 2) = P(0 \le Z) + P(0 \le Z \le 2)$$
$$= 0.5 + 0.4772 = 0.9772$$

EXERCISE SIX

1. let x represent the number of Heads obtained from tossing a fair coin three times. Find (i). the distribution of x? (ii). Probability of getting no head; one head; two heads and three heads? (iii). Probability of getting less than two heads? (iv) Probability of getting more than two heads?

x	0	1	2	3
outcome	1	3	3	1

2. In a class of 200 Maths / Statistics students enrolled in STA 112, it was observed that 117 students offered Maths, 133 offered Physics while 95 offered both Physics and Maths.
   (i). How many of these students are not enrolled in any of the two courses?
   (ii). How many students offered only Physics?
   (iii). How many students offered only Maths?
   (iv). What is the probability of those who did not offer any of the courses?
3. In a gathering of 60 men and 40 women, it is observed that 24 men and 21 women wear glasses. What is the probability that:
   (a). a person picked at random wears glasses in the gathering
   (b). a man picked at random wears glasses.
   (c). a woman picked at random does not wear glasses.
   (d). a person picked at random is either a man who wears glasses or a woman who does not wear glasses.
4. Urn A_1 contains 8 black and 2 white balls. Urn A_2 contains 3 black and 7 white balls. Urn A_3 contains 5 black and 7 white balls. A fair die is to be cast. If the die turns up 1, 2, or 3, a ball will be selected from A_1. If the die turns up 4 or 5, a ball will be selected from A_2. Finally, a ball will be selected from A_3 if the die turns up 6. Given that the ball selected is black, what is the probability that the ball was chosen from A_2.
5. A child who cannot differentiate between wine, whisky and brandy picks a bottle from a drawer containing 2 bottles of wine, 3 bottles of whisky and 4 bottles of brandy. What is the probability that the child selects a bottle of whisky?
6. If a die is tossed once, what is the probability of obtaining an even number less than 5
7. Two children have their birthdays in the same week. Calculate the probability that the birthday is on the same day.
8. An elevator [lift] has two persons each of whom might stop on one of the five floors of a building. Calculate the probability if the elevator does not go beyond the first floor.
9. A box contains 5 red balls, 3 blue balls and 2 green balls. A man draws 1 ball from the box. Calculate the probability that the ball is either red or green.
10. In the joint toss of two dice, let Event A be such that both die show odd numbers, Event B is such that sum of the scores is 4. Obtain the probability that either A or B occurs.

11. A company ships her dish washers in lots of 24. Four are randomly selected and inspected and the lot is passed if all the four are in good condition. Find the probability that a lot will pass the inspection when four dishwashers are not in good condition.
12. A fair coin is tossed 5 times. Determine the probability of observing (i) Exactly 3 heads (ii) At least 3 heads.
13. A bag contains 3 black balls and 2 white balls all of the same size. A sample of 3 balls is selected from the bag with replacement. Find the probability of observing exactly 2 black balls.
14. A gambler pays ₦10 to play a game, which consists of tossing die 5 times and winning ₦200 if 3 or more sixes appears. (a). determine his probability of winning a single game. (b). determine his expected winning (income) per game and also his expected net profit.
15. Past records show that 40 out of 100 automobile accidents are due to over speeding. Find the probability that among 8 recent automobile accidents; (a) (i). Exactly 3 of them are due to over speeding. (ii). At least 2 of them are due to over speeding. (b) (i). Determine the mean and variance of the number of accidents due to over speeding.
16. It is known that 2 out of every 10 bolts produced by a machine are defective. In a pack of six bolts, calculate the probability that at least 2 bolts are defective.
17. The probability that a student guesses the answer to a question correctly is ¼. What is the probability that out of 5 questions he guesses 2 correctly?
18. 2% of the tools produced in a certain manufacturing process turn out to be defective. Find the probability that in a sample of 100 tools chosen at random exactly 5 will be defective.
19. A point is won if either 2 or 4 appear in a toss of a fair die. What is the probability that out of 6 tosses a girl wins 3 points?
20. If a coin is tossed once, what is the probability of
    (a) obtaining a head and a tail.
    (b) obtaining a head or a tail.
21. The probability that a patient dies as a result of surgical operation is 0.003. If 1000 patients are to undergo surgical operations, what is the probability of recording 3 or more deaths?
22. On the average, 3 vehicles pass through a university gate every 5 minutes. What is the probability that 4 or more vehicles will pass through the gate in a given 5-minute interval?
23. Vehicles arrive at a certain public motor park at the rate of one a minute. What is the probability that in a one-minute interval not more than 2 vehicles enter?
24. A Poisson distribution is given by $P(X = x) = \frac{(0.72)^x e^{-0.72}}{x!}$, Find
    (a) P(X=0) ; (b) P(X=1) : (c) P(X=2) ; (d) P(X=3)
25. Given the probability function $P(x) = \binom{5}{x}(0.6)^x(0.4)^{n-x}; 0, 1, \ldots, 5$.
    (a) Use Poisson approximation to binomial to find the mean and variance of X.
    (b). From (a), determine P(X=3)

26. The number of cars passing through FUTO roundabout in an hour is a random variable X having the probability function $P(x) = \frac{5^x \ell^{-5}}{x!}$; 0, 1, . . . Find (a) P(X = 4) (b) $P(4 \le X \le 6)$

27. In a certain book industry the number of misprints per page is a Poisson random variable X having parameter $\lambda = 0.233$. (a). Find the probability that a randomly selected page will contain no misprint. (b). If 5 pages of a book are randomly selected, find the probability that (i) Exactly 3 of them will contain no misprint. (ii). At least 3 of them will contain no misprint.

28. The following data; 2, 3, 5, 1, 4, 7 represent the number of alpha particles emitted by radioactive source at a given time. What is the probability that (a). at most three particles will be emitted at a given time. (b). at least one particle will be emitted at a given time.

29. **(a).** Suppose that the number of typographical errors on a single page of a book of 235 pages has a Poisson distribution with parameter 3. Calculate the probability that there are at least 2 errors in each page.
(**b**)**.** Find the expected number of error in the entire book.

30. The number of cars passing through a roundabout at a given time are given in the table below, where x denote different types of cars and f denote the frequency of each of the cars.

x	1	2	3	4	5	6	7	8	9
f	8	10	6	12	9	5	7	4	5

(a). fit Poisson distribution to the data
(b). what is the probability that exactly six cars pass through the roundabout at a given time?

31. The yields in kg of tomatoes have mean 39 and standard deviation 7.937. Assuming that the yields are normally distributed with those mean and standard deviation, find W such that the probability of the yield from a plot being greater than W kilograms is 5%.

32. The heights of police recruits in a country are normally distributed with mean μ and standard deviation 0.05m. If 10% of the police recruits have heights exceeding 1.8m, find the mean height of the police recruits.

33. Sacks of grains packed by an automatic machine loader have an average height of 50kg. It is found that 10% of the bags are over 51kg. Find the standard deviation.

34. The weight of 2000 cattle were normally distributed with μ = 100kg and Var (x) = 225sqkg. Find the number of cattle having weights
(i). below 90kg. (ii). between 90kg and 115kg. (iii). above 120kg.

35. Students' scores in a mathematics examination are normally distributed with mean 75% and standard deviation 10%. Find the probability that a student selected at random scores less than 60%.

36. Students' scores in a mathematics examination are normally distributed with mean 75% and standard deviation 10%. Find the probability that a student selected at random scores between 70% and 90%.

37. The breaking strength of materials from a steel rolling mill, measured in appropriate units, is normally distributed with mean 20 and standard deviation 5. Obtain the value of k such that P(X k) = 0.242

38. If a variable X is normally distributed with mean μ = 80 and variance $\sigma^2 = 25$.
Find (i). P(72.5 < X < 92) (ii). P(X < μ + 2σ).

39. Given that X~ N (50, 100). Find
(i). $p(X \leq 60)$. (ii). p(45< X < 65.3)

40. A box contains 400 beads of which 4% are defective; a second box contains 120 beads of which 6% are defective. Two other boxes contain 500 beads with 5% defectives each. If it is known that the four boxes are identical in appearance, what is the probability that a bead selected from one of the oxes is defective?

CHAPTER SEVEN
ESTIMATION

7.0 INTRODUCTION

Estimation and estimation problem is an aspect of statistical inference in which one is interested in the approximation of some numerical characteristic, say θ, of an unknown population parameter on the basis of a sample, either by using a number or interval determined by two numbers or points. The approximation is done using appropriate statistic. Hence, estimation is the process through which the value of an appropriate statistic or estimator is used to estimate corresponding numerical characteristic, say θ, of an unknown distribution or population parameter. There are two types of estimation namely point and interval estimation.

7.1 Point Estimation

This is a type of estimation in which a single numerical value or estimate, say θ computed from an appropriate statistic or estimator is used to approximate or estimate corresponding but unknown numerical characteristic of a population parameter, say θ. The single numerical value computed from an appropriate statistic or estimator is called a point estimate while, the statistic or function of the random sample used to approximate or estimate the corresponding but unknown numerical characteristic of a population parameter is called the point estimator. The three most commonly used methods of point estimation are the method of moments, Maximum Likelihood Estimation and the Method of Least Squares.

7.1.1 The Method of Moments

Moments are used in statistics to understand the various characteristics (Central tendency, dispersion, skewness and kurtosis) of a frequency distribution.

The method of moments is a type of point estimation in which sample moments are equated with corresponding observable population moments, and the resulting equations are solved simultaneously for the quantities or parameters to be approximated or estimated.

The Concept of expectation:

Let μ_r be the population moments. Then, by definition;

$$\mu_r = E(X^r) = \begin{cases} \sum_{i=1}^{n} x_i^r \ f(x); & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^r f(x)\,dx; & \text{if } X \text{ is continuous} \end{cases} \qquad (7.1)$$

where $-\infty$ and ∞ are the lower and upper limits respectively for the values of X.

μ_r is also called the rth population moment about the origin or the crude moments.

Let M_r be the sample moments. Then, by definition;

$$M_r = E(X^r) = \frac{\sum_{i=1}^{n} x_i^r}{n} \tag{7.2}$$

Then, from the law of large number as $n \to \infty$, $M_r \approx \mu_r \forall r = 1, 2, ..., k$. Then, by equating sample moments to corresponding population moments, we obtain;

$$\begin{aligned} M_1 &= \mu_1 \\ M_2 &= \mu_2 \\ &\vdots \\ M_k &= \mu_k \end{aligned} \tag{7.3}$$

The k[th] resulting equations are solved simultaneously for the quantities or parameters to be approximated or estimated.

Example 7.1: Let $X_1, X_2, X_3, ..., X_n$ be an independent identical distributed random variable from a Binomial distribution with parameter, $\theta = (n, p)^T$. Using the method of moments, evaluate the point estimate for the population parameter θ.

Solution

If X is an independent identical distributed random variable from a Binomial distribution then,

$$P(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}; x = 0, 1, ..., n; 0 \le p \le 1 \\ 0; \ otherwise \end{cases} \tag{7.4}$$

$$\mu_1 = E(X) = np \tag{7.5}$$

$$\mu_2 = E(X^2) \tag{7.6}$$

But,

$$\text{var}(X) = E(X - \mu)^2 = E(X^2) - \mu_1^2 = \mu_2 - \mu_1^2$$

$$\Rightarrow \quad \mu_2 = \text{var}(X) + \mu_1^2 = npq + n^2 p^2 \tag{7.7}$$

Given M_r as the sample moments, then;

$$M_r = E(X^r) = \frac{\sum_{i=1}^{n} x_i^r}{n}$$

$$\Rightarrow \quad M_1 = E(X) = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X} \tag{7.8}$$

$$M_2 = E(X^2) = \frac{\sum_{i=1}^{n} x_i^2}{n} \tag{7.9}$$

Equating sample moments to corresponding population moments, we obtain

$$\bar{X} = n\,p \tag{7.10}$$

$$\frac{\sum_{i=1}^{n} x_i^2}{n} = npq + n^2 p^2 \tag{7.11}$$

Solving (7.10) and (7.11) simultaneously, we obtain

$$\hat{n} = \frac{\bar{X}^2 + \sum_{i=1}^{n} x_i^2}{\bar{X}(\bar{X} + 1)} \tag{7.12}$$

and

$$\hat{p} = \frac{\bar{X}}{\hat{n}} \tag{7.13}$$

Advantages of method of moments

1. It can be easily and quickly calculated by hands.
2. Estimate obtained using the method of moments may be used as the first approximation to the solutions of the likelihood.

7.1.2 The Maximum Likelihood Estimation

A likelihood function is a function of the parameters of a statistical model, given specific observed data and can be used to estimate the distribution parameters irrespective of the distribution used.The Maximum Likelihood Estimation, denoted MLE, is a method of point estimation which involves finding the Likelihood function, the joint densities of the random variates, that maximizes the amount of information about the unknown numerical characteristic of a population parameter, say θ present in the random sample. Then, the amount of information present in the random sample, say $\hat{\theta}$, about the unknown numerical characteristic of the population parameter, say θ can be found by using various optimization algorithms or method of calculus. The procedural steps for the application of the maximum likelihood estimation are;

STEP 1: Given a random sample $X_1, X_2, X_3, ..., X_n$ from a distribution with probability density or mass function, $f(x_i/\theta)\ \forall\ i = 1,2,...,n$

STEP 2: Obtain the Likelihood function of the random sample $X_1, X_2, X_3, ..., X_n$ as

$$L(\theta) = \prod_{i=1}^{n} f(x_i/\theta) \tag{7.14}$$

Note that $L(\theta)$ is a monotonic function that contains the maximum amount of information about the unknown numerical characteristic of a population parameter, say θ present in the random sample.

STEP 3: Obtain a logarithmic likelihood function of the random sample $X_1, X_2 X_3, ..., X_n$ which is a maxima function that is unaffected by monotone transformation as;

$$\log(L(\theta)) = L^*(\theta) = \log(\prod_{i=1}^{n} f(x_i/\theta)) = \sum_{i=1}^{n} \log f(x_i/\theta) \tag{7.15}$$

STEP 4: Using the method of calculus, obtain the first partial derivative of the logarithmic likelihood function $(L^*(\theta))$ of the random sample $X_1, X_2, X_3, ..., X_n$ with respect to the unknown population parameter θ and equate to zero as;

$$\frac{\partial L^*(\theta)}{\partial \theta} = 0 \tag{7.16}$$

Solve the resulting equation in order to obtain the desired solution for θ as θ_0.

Example 7.2 Let $X_1, X_2, X_3, ..., X_n$ be an independent identical distributed random variable from an Exponential distribution with parameter, θ. The probability density function (pdf) is given as;

$$f(x_i/\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}; & \theta > 0; i = 1,2,...,n \\ 0; & otherwise \end{cases}$$

Evaluate the point estimate for the population parameter θ using the method of maximum likelihood estimation.

Solution

If X is an independent identical distributed random variable from a Exponential distribution then,

$$f(x_i/\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}; & \theta > 0; i = 1,2,...,n \\ 0; & otherwise \end{cases}$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i/\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x}{\theta}} = \theta^{-n} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}}$$

$$\log(L(\theta)) = L^*(\theta) = \log(\prod_{i=1}^{n} f(x_i/\theta)) = \sum_{i=1}^{n} \log f(x_i/\theta) = -n\log\theta - \frac{\sum_{i=1}^{n} x_i}{\theta}$$

$$\frac{\partial L^*(\theta)}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0$$

$$\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$$

Then,

$$\left.\frac{\partial^2 L^*(\theta)}{\partial \theta^2}\right|_{\hat{\theta}=\bar{x}}, = \frac{n}{\theta^2} - \left.\frac{2\sum_{i=1}^{n} x_i}{\theta^3}\right|_{\hat{\theta}=\bar{x}} = -\frac{n}{\bar{X}^2} < 0$$

Hence, $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$ is the maximum likelihood estimate for θ.

Advantages of the maximum likelihood estimation

1. Maximum likelihood provides a consistent approach to parameter estimation problems. That is the maximum likelihood estimates can be developed for a large variety of estimation situations.
2. The estimators obtained using Maximum likelihood methods become minimum variance unbiased estimators as the sample size increases.

Disadvantages of the maximum likelihood estimation

1. The likelihood equations need to be specifically worked out for a given distribution and estimation problem. The mathematics is often non-trivial, particularly if confidence intervals for the parameters are desired.
2. Maximum likelihood estimates can be heavily biased for small samples. The optimality properties may not apply for small samples.
3. Maximum likelihood can be sensitive to the choice of starting values.

7.1.3 The Method of Least Squares

The method of least squares, denoted by MLS, is a method of point estimation which involves minimizing the error sum of squares of a function of a random sample, and using various optimization algorithms or method of calculus in order to obtain the estimate $\hat{\theta}$ of the population parameter θ. The method of least squares works well in the estimation of $\hat{\theta}$ of the population parameter θ for linear models. The procedural steps for the application of MLS are;

STEP 1: Consider an estimator $\hat{Y}_i$ of the population parameter Y_i and evaluate the error (e_i), the deviation of $\hat{Y}_i$ from Y_i as;

$$e_i = Y_i - \hat{Y}_i \tag{7.17}$$

STEP 2: Square the error and sum it in order to obtain the error sum of squares as;

$$S(\theta) = \sum e_i^2 \tag{7.18}$$

STEP 3: Using the method of calculus,obtain the first partial derivative of the error sum of squares and equate to zero as;

$$\frac{\partial S(\theta)}{\partial \theta} = 0 \tag{7.19}$$

Solve the resulting equation in order to obtain the desired solution for the estimate of θ as θ_0.

Example 7.3: Obtain the estimate of the intercept and regression coefficient of the linear regression model: $Y_i = \beta_0 + \beta_1 X_i + e_i; \quad i = 1,2,...,n$

Solution

Given

$Y_i = \beta_0 + \beta_1 X_i + e_i; \quad i = 1,2,...,n$

$e_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_i)$

$$S(\theta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

$$\frac{\partial S(\theta)}{\partial \beta_0} = -2\sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)] = 0$$

$$\Rightarrow \hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n} = Y - \hat{\beta}_1 \overline{X} \qquad \text{i}$$

1. $$\frac{\partial S(\theta)}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i [Y_i - (\beta_0 + \beta_1 X_i)] = 0$$

$$\Rightarrow \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \beta_0 \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2} \qquad \text{ii}$$

Substituting (i) into (ii), we obtain

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left[\sum_{i=1}^{n} X_i\right]^2} \qquad \text{iii}$$

Then substituting (iii) into (i), we obtain

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n} = Y - \hat{\beta}_1 \overline{X} \qquad \text{iv}$$

Advantages of the method of least squares estimation

1. Simplicity: MLS is very easy to explain and to understand
2. Applicability: There are hardly any applications where least squares doesn't make sense
3. Calculations are easy and very fast.

Disadvantages of the method of least squares estimation

1. MLS estmates are sensitivity to extreme values.
2. Test statistics obtained using MLS might be unreliable when the data is not normal.

Drawback of point estimation

1. Generally, point estimation yields only a single numerical value for the approximation or estimation of the corresponding but unknown numerical characteristic of a population parameter, say θ.

2. Point estimation does not give the ability to report the accuracy of the estimate.

7.2 INTERVAL ESTIMATION

Interval estimation addressed the drawback of Point estimation. Interval estimation provides a random interval for the estimates that contains the true value of the corresponding but unknown numerical characteristic of a population parameter say θ, and also enable the proper reporting of the accuracy of the estimates.

Hence, a $100(1-\alpha)\%$ confidence interval is a random interval, say (L_1, L_2), of the estimates that contains the corresponding but unknown numerical characteristic of a population parameter, θ, with certainty regardless the value of θ. That is,

$$P(L_1 \le \theta \le L_2) = 1-\alpha$$

where α is the level of significance, L_1 and L_2 are the lower and upper limits of the random interval. Estimating the value of θ is of the form;

$$\hat{\theta} = \text{Estimator} \pm \lfloor \text{Acceptance probabilit y point} \rfloor * \lfloor \text{Standard deviation of the Estimator} \rfloor$$

In this study, we shall discuss the interval estimation of mean for one population.

7.2.1 Interval Estimation of Mean for One Population

Interval estimation of mean for one population will be discussed for three distinct cases. The three distinct cases are Interval estimation of mean for a normal population (sample size (n) is greater than or equal to thirty(30)) when the variance is known, Interval estimation of mean for a normal population (sample size (n) is greater than or equal to thirty(30)) when the variance is unknown and Interval estimation of mean for a normal population, when variance is unknown and the population size is small (sample size (n) is less than thirty(30)).

7.2.1.1 Interval estimation of mean for one large or normal population (sample size (n) is greater than or equal to thirty (30)) when variance is known

The sampling distribution for the mean for one population when variance is known is given as

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \tag{7.20}$$

where Z follow the standard normal distribution, $\bar{X}$ is the sample mean of the random smple $X_1, X_2, X_3, ..., X_n$, μ is the population mean, σ is the population standard deviation and n is the sample size of the random sample. The $100(1-\alpha)\%$ confidence interval for acceptance of μ is obtained by partitioning the standard normal curve for Z . That is;

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2}\,\sigma/\sqrt{n} \le \bar{X} - \mu \le Z_{\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha$$

$$P(\bar{X} - Z_{\alpha/2}\,\sigma/\sqrt{n} \le \mu \le \bar{X} + Z_{\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha \tag{7.21}$$

The $100(1-\alpha)\%$ confidence interval for acceptance of μ when variance is known with $(1-\alpha)$ certainty is

$$\bar{X} - Z_{\alpha/2}\,\sigma/\sqrt{n} \le \mu \le \bar{X} + Z_{\alpha/2}\,\sigma/\sqrt{n} \tag{7.22}$$

But generally, the $100(1-\alpha)\%$ confidence interval for acceptance of μ when variance is known with $(1-\alpha)$ certainty is given by

$$\hat{\mu} \in \bar{X} \pm Z_{\alpha/2}\,\sigma/\sqrt{n} \tag{7.23}$$

Example 7.4: A random sample of 30 students selected from a class yielded an average weight of 260 kg with a variance of 4900. Determine the 95% confidence interval for estimating the average weight of students in future.

Solution

Given $\bar{x}=260kg;\ \sigma^2=4900, \sigma=70$ *and* $n=30$.

$100(1-\alpha)\%$ Confidence interval $=95\%$

$\Rightarrow 1-\alpha=0.95$

Hence, $\alpha=1-0.95=0.05$

$Z_{\alpha/2}=Z_{0.05/2}=Z_{0.025}=1.96$

Substituting the values of $\bar{x}, \sigma, Z_{\alpha/2}$ *and* n in Equation 7.23, we obtain

$$\hat{\mu}=260\pm(1.96)\left(70/\sqrt{30}\right)=260\pm25.05=(234.95,285.05)$$

7.2.1.2 Interval estimation of mean for one large population (sample size (n) is greater than or equal to thirty(30)) when variance is unknown

When the variance of the population is unknown, the large sample variance is used as an approximation of the population variance. Then, the sampling distribution for the mean for one population when variance is unknown is given by

$$Z=\frac{\bar{X}-\mu}{S/\sqrt{n}} \tag{7.24}$$

$$S=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} \tag{7.25}$$

where Z follow the standard normal distribution, $\bar{X}$ is the sample mean of the random sample $X_1, X_2, X_3, ..., X_n$, μ is the population mean, S is the large sample standard deviation and n is the sample size of the random sample. The $100(1-\alpha)\%$ confidence interval for acceptance of μ is obtained by partitioning the standard normal curve for Z. That is;

$$P\left(-Z_{\alpha/2}\le Z\le Z_{\alpha/2}\right)=1-\alpha$$

$$P\left(-Z_{\alpha/2}S/\sqrt{n}\le\bar{X}-\mu\le Z_{\alpha/2}S/\sqrt{n}\right)=1-\alpha$$

$$P\left(\bar{X}-Z_{\alpha/2}S/\sqrt{n}\le\mu\le\bar{X}+Z_{\alpha/2}S/\sqrt{n}\right)=1-\alpha \tag{7.26}$$

The $100(1-\alpha)\%$ confidence interval for accesptance of μ when variance is known with $(1-\alpha)$ certainty is

$$\bar{X}-Z_{\alpha/2}S/\sqrt{n}\le\mu\le\bar{X}+Z_{\alpha/2}S/\sqrt{n} \tag{7.27}$$

But generally, the $100(1-\alpha)\%$ confidence interval for acceptance of μ when variance is known with $(1-\alpha)$ certainty is given by

$$\hat{\mu} \in \bar{x} \pm z_{\alpha/2} S/\sqrt{n} \tag{7.28}$$

Example 7.5: The nicotine content (measured in milligrams) per Cigarette stick in a random sample of 31 sticks of a particular brand of Cigarettes, selected from different packs, which was produced by certain company are:

20.1	19.6	19.2	22.1	17.0	17.9	20.1	20.0	21.2	17.4	18.2	18.2
19.7	18.5	16.9	19.4	19.1	18.3	17.9	18.7	21.4	18.7	20.2	19.7
17.7	16.8	17.9	19.7	17.5	17.4	19.1					

Determine the 95% confidence limits for assessing the average nicotine content Cigarettes produced by the company in future.

Solution

Given $n=31$

$100(1-\alpha)\%$ Confidence interval $=95\%$

$\Rightarrow 1-\alpha=0.95$

Hence, $\alpha=1-0.95=0.05$

$Z_{\alpha/2}=Z_{0.05/2}=Z_{0.025}=1.96$

$\bar{X}=\sum x/n = 585.6/31=18.8903$

$S=\sqrt{\sum(X-\bar{X})^2/n-1} \quad =\sqrt{54.8871/30} =1.3526$

Substituting the values of $\bar{X}$, S, $Z_{\alpha/2}$ *and* n in Equation 7.28, we obtain

$\hat{\mu}=18.8903\pm1.96\left(1.3526/\sqrt{31}\right)=18.8903\pm0.4762=(18.4141, 19.3665)$ milligrams

7.2.1.3 Interval estimation of mean for one population, when variance is unknown and the sample size is small (sample size (n) is less than thirty(30))

When the variance of the population is unknown, the sample variance is used as an approximation of the population variance. Then, the sampling distribution for the mean for one population when variance is unknown and the population size is small is given by

$$t=\frac{\bar{X}-\mu}{S/\sqrt{n}} \tag{7.29}$$

$$S=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} \tag{7.30}$$

where t follow the student t distribution with $(n-1)$ degree of freedom, $\bar{X}$ is the sample mean of the random sample $X_1, X_2, X_3, ..., X_n$, μ is the population mean, s is the large sample standard deviation and n is the sample size of the random sample. The $100(1-\alpha)\%$

confidence interval for acceptance of μ is obtained by partitioning the student t distribution curve for t. That is;

$$P(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1-\alpha$$

$$P\left(-t_{\alpha/2} \le t = \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2}\right) = 1-\alpha$$

$$P(\bar{X} - t_{\alpha/2} S/\sqrt{n} \le t \le \bar{X} + t_{\alpha/2} S/\sqrt{n} \quad (7.31)$$

The $100(1-\alpha)\%$ confidence interval for acceptance of μ when variance is known with $(1-\alpha)$ certainty is

$$\bar{X} - t_{\alpha/2} S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2} S/\sqrt{n} \quad (7.32)$$

But generally, the $100(1-\alpha)\%$ confidence interval for acceptance of μ when variance is known with $(1-\alpha)$ certainty is given as

$$\hat{\mu} = \bar{x} \pm t_{\alpha/2} S/\sqrt{n} \quad (7.33)$$

Example 7.6: The average life span of twenty light bulbs (measured in hours) produced by a company is given as follows;

616 622 625 575 600 605 590 600 617 619 608
613 611 599 616 617 616 620 589 608

Determine the 95% confidence limits for assessing the average life span of light bulbs produced by the company in future.

Solution

Given $n = 20$.

$100(1-\alpha)\%$ Confidence interval $= 95\%$

$\Rightarrow 1-\alpha = 0.95$

Hence, $\alpha = 1-0.95 = 0.05$

$$t_{\alpha/2} = t_{\alpha/2,n-1} = t_{0.05/2,20-1} = t_{0.025,19} = 2.09$$

$$\bar{x} = \sum X / n = 12166/20 = 608.3$$

$$S = \sqrt{\sum (X-\bar{X})^2/(n-1)} = \sqrt{3128.2/19} = 12.8313$$

Substituting the values of $\bar{x}, S, t_{\alpha/2}$ *and* n in Equation 7.33, we obtain

$$\hat{\mu} = 608.3 \pm 2.09\left(12.8313/\sqrt{20}\right) = 608.3 \pm 5.9966 = (602.3034,\ 614.2966) \text{ hours}$$

EXERCISE SEVEN

1. What is Estimation?.
2. What is the point estimate of the sample mean (μ)? (ii) Differentiate between point and interval estimation?
3. Differentiate between population parameter and sample estimate. (ii) What is an estimator?
4. Given the following scores of 10 students from Department of Statistics, FUTO 45, 67, 78,56,55,34,48,87,12 and 42. Assuming data came from a normal population. Set up a 95% confidence interval for estimating the population mean (μ).
5. Given the following set of 10 observations: 2.1, 1.2, 1.6, 0.7, 2.1, 2.7, 2.4, 0.7, 0.6 and 0.9. assuming that the data is drawn from a normal population, construct a 99% confidence interval for estimating the population mean μ.
6. A sample of size n = 100 yielded the sample mean $\bar{x}$ = 16. If the population standard deviation is 3, compute a 95% confidence interval for approximating the population mean (μ).
7. Assuming the population standard deviation is 3, how large should a sample be to estimate the population mean with a margin of error not exceeding 0.5?
8. The operations manager of a Production Plant would like to estimate the average amount of time a worker takes to assemble a new electronic component. Assume that the standard deviation of the assembly time is 3.6. (a) After observing 120 workers assembling similar devices, the manager noticed that their average time was 16.2 minutes. Construct a 92% confidence interval for approximating the average assembly time in the Production Plant. (b) How many workers should be involved in this study in order to achieve a mean assembly time of 15 seconds at 92% confidence?
9. It is necessary to estimate the mean number of concurrent users in order to ensure efficient usage of a server,. The sample mean and sample standard deviation of 100 randomly selected concurrent users at times is 37.7 and 9.2, respectively. Construct a 90% confidence interval for the mean number of concurrent users.
10. To assess the accuracy of a laboratory scale, a standard weight that is known to weigh 1 gram is repeatedly weighed 4 times. The resulting measurements (in grams) are: 0.95, 1.02, 1.01, and 0.98. Assume that the weighing by the scale when the true weight is 1 gram are normally distributed with mean. Use these data to compute a 95% confidence interval for population mean (μ).
11. Suppose we would like to estimate the mean amount of money ( ) spent on books by Computer Science (CS) students in a semester. We have the following data from 10 randomly selected CS students' $\bar{x}$ = \$249 and S = \$30. Assume that the amount spent on books by CS students is normally distributed. To compute a 95% confidence for the population mean (μ), we would use which of the following critical points: (a) $z_{0.025}$ = 1:96 (b) $z_{0.05} = 1.645$ (c) $t_{9,0.025} = 2.262$ (d) $t_{10,0.025} = 2.228$
12. Installation of a certain hardware takes a random amount of time with a standard deviation of 5 minutes. A computer technician installs this hardware on 64 different computers, with the average installation time of 42 minutes. Compute a 95% confidence interval for the mean installation time.

13. The time needed for college students to complete a certain maze follows a normal distribution with a mean of 45 seconds. The standard deviation of the collected data is 3.5 seconds in a group of nine students. Construct a 95% confidence interval for the mean exercise time for the entire students.
14. What is the difference between point and interval estimation?
15. What does 95% confidence interval mean?
16. Given the confidence interval $(1-\sigma)$ and the margin of error W, what will be the minimum number of sample size?
17. A consumer group would like to estimate the mean monthly electricity charge for a single family house in July (within $5) using a 99 percent level of confidence. Based on similar studies the standard deviation is estimated to be $20.00. (a) How large a sample is require?
18. Let the strength of a certain iron rod rolled by Ajaokuta Steel complex be a normal random variable having population standard deviation as 0.7615 unit. If the mean breaking strength of 100 such rods is 34.5 units, construct a 99% confidence interval for the mean breaking strength of all the iron rods rolled by the steel complex.
19. Consider the mathematical expression $P(L \le \theta \le U) = 1-\alpha$,Where L is called Lower confidence limit and U is called upper confidence limit. What is $1-\alpha$ called?
20. The interval $L = \bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad U = \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is called?
21. List the two conditions for using the t-student distribution for interval estimation? (b) Given two samples n_1 and n_2 with size 9 and 16 respectively. Give correct to 2 significant figures the value of $t_{(n_1+n_2-2)\frac{\alpha}{2}}$. Take $\alpha = 0.05$
22. Mention two objectives of designing an interval estimator. (ii) Mention two steps of designing an interval estimator.

CHAPTER EIGHT
TEST OF HYPOTHESIS

8.0 INTRODUCTION

From an elementary and general perspective, a hypothesis is described as an intelligent guess of a scientist. In Statistics as a discipline, a hypothesis is described as an assertion about one or more population parameters or about the distribution of the population itself which needs to be tested so that the hypothesis could be upheld or rejected at a certain level of significance. A hypothesis rejected at a certain level of significance may be upheld at another level of significance; for example, a hypothesis rejected at 1% (0.01) level of significance may be upheld (accepted) at 5% (0.05) or at 10% (0.1) level of significance. This **level of significance** is the probability that the hypothesis will be significantly rejected (or the probability that the hypothesized population parameter(s) will fall in the critical region of rejection). On the other hand, the level of significance can also be used to deduce how confident the statistician is that the hypothesis is true (or the probability that the hypothesized population parameter(s) will fall in the acceptance region. For example, a 1% level of significance would imply that the statistician is 99% confident that his hypothesis is true or that the true value of the population parameter lies in the acceptance region. Statistical hypothesis often comes in pairs. They are usually stated as **null hypothesis (** H_0 **)** and **alternative hypothesis (** H_1 **)**. The hypothesis to be tested is the null hypothesis while the alternative hypothesis is a negation of the null hypothesis. In stating the null and alternative hypotheses, it may be done in simple or composite form. This gives rise to simple and composite hypotheses. In **simple hypothesis**, the parameters or distribution of the population are completely stated, example, $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 < \theta_2$ or $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$. On the other hand, in **composite hypothesis**, the parameters or distribution of the population are not completely stated, example, $H_0 : \theta_1 > \theta_2$ versus $H_1 : \theta_1 < \theta_2$. It is more common to state statistical hypotheses in simple forms.

Given that in statistical inference, where test of hypothesis is discussed, only a sample(s) and not the actual population is used, therefore, a true hypothesis may be rejected which leads to an error in the test. On the other hand, a false hypothesis may also be accepted, this leads to another error. Rejection of the null hypothesis when it is true, is called a **Type I error**. On the other hand, acceptance of the null hypothesis when it is false, is called a **Type II error**. It is possible to reduce the probability of committing either errors in test of hypothesis, but the problem is that, reducing the probability of committing any of the errors implies increasing the probability of committing the other one. A lemma proposed by Neyman-Pearson attempts to proffer solution to this problem but this is beyond the scope of this manual. However, increasing the sample size reduces both the probabilities of committing the Type I and Type II errors simultaneously. The size of the Type I error is denoted by α

while the size of the Type II error is denoted by β. The preassigned level of significance in test of hypothesis corresponds to α while β corresponds to $1-\alpha$.

A test of hypothesis of the type, $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 < \theta_2$ or $H_1 : \theta_2 > \theta_1$ is called a **one-tailed test** while a test of hypothesis of the type, $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$ is called **two-tailed test**. In test of hypothesis involving a two-tailed test, the level of significance, α is divided by 2 before the table value or critical value is found. The table value is now found at the point where α is halved. This critical value is obtained in accordance with the test statistic for the hypothesis. The **critical value** is a point of demarcation between the acceptance region and the rejection region for the test. There is usually a table according to the test statistic containing different critical values of the test at different levels of significance and sometimes at different degrees of freedom as the case may be.

8.1 Power Function of a Test:

This is the function defined for all distributions under consideration which yields the probability that the sample point falls in the critical region of the test. The value of the power function at a parameter point is called the **power of the test** at that point.

8.2 Test Statistic

A test statistic is a function of some statistics (or sometimes, a function of some statistics and parameters, for e.g., in the case where the population variances are known and hypothesis is being tested on the population means) which is used according to some prescribed rule for a test of hypothesis. Every test statistic has a distribution. It is this distribution alongside the level of significance and sometimes with the degree of freedom as the case may be that is used to obtain the critical values (popularly known as the table values). The test statistic employed for a particular test of hypothesis depends on the parameter(s) that the hypothesis is being tested on and sometimes on the sample size.

8.3 Test of Hypothesis on One Population Mean

When the sample size is large $(n \geq 30)$ and it is desired to test if the population mean corresponds to the hypothesized value. The following hypotheses could be tested:

(i) $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$

(ii) $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$

(iii) $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$.

The appropriate test statistic for the following hypotheses when the population variance is known is given by,

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (8.1)$$

where Z shows that the statistic has a standard normal distribution and gives the calculated value of the test statistic, $\overline{X}$ gives the sample mean, μ is the hypothesized population mean, σ is the population standard deviation and n is the sample size.
If the sample size is small $(n < 30)$ and σ is known , the above statistic is only valid when the population being sampled from is not too different from the normal distribution. If on the other hand, the population variance is not known but the sample size is still large, σ is replaced with S, where S gives the sample estimate of the population standard deviation. In this case, the appropriate test statistic becomes,

$$Z = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}. \tag{8.2}$$

The critical value for this test is obtained at Z_α for a one-tailed hypothesis and at $Z_{\frac{\alpha}{2}}$ for a two-tailed hypothesis.
When the sample size is small $(n < 30)$ and the population variance is not known, the appropriate test statistic is,

$$t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}, \tag{8.3}$$

where, t implies that the test statistic has a t-distribution and gives the calculated value of the test statistic. The degree of freedom of this test statistic is $n-1$. For a one-tailed test, the critical value is obtained at $t_{\alpha.n-1}$ while, for a two-tailed test, the critical value is obtained at $t_{\frac{\alpha}{2}.n-1}$.

8.4 Test of Hypothesis on Two Population Means

Suppose, we are interested in finding out if two population means are equal or if two methods of doing something would produce the same average result, the appropriate statistic when the sample size is large and the population variances are known is given by,

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{8.4}$$

where, $\overline{X}_1$, σ_1^2 and n_1 are the sample mean, population variance and sample size from the first population; $\overline{X}_2$, σ_2^2 and n_2 are the sample mean, population variance and sample size from the second population.

The critical value for this test is obtained at Z_α for a one-tailed hypothesis and at $Z_{\frac{\alpha}{2}}$ for a two-tailed hypothesis.

On the other hand, if the sample sizes are not large and the variances have to be estimated from the samples, the appropriate test statistic for equality of the two means is given by,

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{8.5}$$

where S_p^2 is the pooled sample variance given by,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

where, S_1^2 and S_2^2 are the sample estimates of the population variances for the first and second populations respectively.

For a one-tailed test, the critical value is obtained at $t_{\alpha, n_1+n_2-2,}$ while, for a two-tailed test, the critical value is obtained at $t_{\frac{\alpha}{2}, n_1+n_2-2}$.

When it is important to test for the equality of more than two means, the F-test popularly known as the Analysis of Variance (ANOVA) is employed. This is discussed in Chapter Ten.

8.5 Test of Hypothesis on Two Means from dependent Populations

In the last section, the test of hypothesis on equality of means was discussed with the assumption that the means come from two independent populations. In the present section, we wish to discuss test of hypothesis on two means from dependent populations. The populations are said to be dependent because they have pairs of related observations. Such dependent populations may include the performance of a class of students in Joint Admissions and Matriculation Board (JAMB) and Post JAMB examination in a given university; the weights of some diabetic patients before and after treatment, to mention but a few. Therefore, the appropriate test statistic for equality of means from such population when $n < 30$ is given by,

$$t = \frac{\overline{d}\sqrt{n}}{S_d}, \tag{8.6}$$

where $\overline{d}$ is the average of the differences between the pairs of observations,

$$S_d^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}, \tag{8.7}$$

and d_i is the difference between i^{th} pair of observations.

For a one-tailed test, the critical value is obtained at $t_{\alpha.n-1}$ while, for a two-tailed test, the critical value is obtained at $t_{\frac{\alpha}{2}.n-1}$.

When $n > 30$, the test statistic is given by,

$$Z = \frac{\bar{d}\sqrt{n}}{S_d}. \tag{8.8}$$

The test statistic is obtained the same way it is obtained when the populations are independent.

For each of the test statistics stated above, the decision rule is to accept the null hypothesis if the calculated value of the test statistic is less than the critical (table) value, otherwise, the hypothesis is rejected.

8.6 Test of Hypothesis on One Population Proportion

Suppose, we wish to test if a hypothesized proportion is equal to the population, the appropriate test statistic is given by,

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}, \tag{8.9}$$

where p is the estimate of the population proportion as obtained from the sample and π is the hypothesized population proportion. In practice, p can be obtained from $p = \frac{y}{n}$, where y is the number of successes and n is the sample size.

The critical value for this test is obtained at Z_α for a one-tailed hypothesis and at $Z_{\frac{\alpha}{2}}$ for a two-tailed hypothesis.

It may happen that we wish to test hypothesis on equality of two population proportions, for example, we may wish to find out if the proportion of adult males from South East Nigeria in support of PDP presidential candidate is equal to the proportion of adult females from the same South East Nigeria in support of the PDP presidential candidate. In this case, the appropriate test statistic for the test is given by,

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}, \tag{8.10}$$

where, p_1 and p_2 are the sample estimates of the population proportions for the first and second populations respectively, n_1 and n_2 are their respective sample sizes.

The decision rule for test of hypothesis on equality of proportions is similar to that of the means, that is, accept the null hypothesis if the calculated value of the test statistic is less than the critical (table) value, otherwise, the hypothesis is rejected. The critical value for this test is obtained at Z_α for a one-tailed hypothesis and at $Z_{\frac{\alpha}{2}}$ for a two-tailed hypothesis.

8.7 Test of Hypothesis on Two Population Variances

If we desire to test if the variability of two processes are the same; this is equivalent to testing for equality of two variances. The appropriate test statistic is given by,

$$F = \frac{S_1^2}{S_2^2}, \tag{8.11}$$

where the symbols retain their usual meanings as earlier explained in this chapter. For a two-tailed test of equality of variances, the null hypothesis is accepted if, $F < F_{\frac{\alpha}{2},n_1-1,n_2-1}$. For a right-tailed test of equality of variances, the null hypothesis is accepted if, $F < F_{\frac{\alpha}{2},n_1-1,n_2-1}$.

On the other hand, for a left-tailed test of hypothesis on equality of variances, the null hypothesis is accepted if $F > F_{1-\alpha,n_1-1,n_2-1}$. The degrees of freedom are $n_1 - 1$ and $n_2 - 1$.

8.8 Numerical Illustrations

1. A teacher claims that the average performance of students in his class in Chemistry is 65%. A random sample of twenty students in the class who were given Chemistry test gave an average performance of 68.2% with a standard deviation of 1.5. Carry out a test of hypothesis to confirm the teacher's claim at 0.05 level of significance.

Solution

The hypotheses for this kind of problem can be stated as:

$H_0 : \mu = 65$

$H_1 : \mu \neq 65$

In this case, the hypothesized value of the mean is 65 with $\bar{X} = 68.2$, $s = 1.5$ and $n = 20$. Since the sample size is small, the appropriate test statistic is given by,

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{68.2 - 65}{\frac{1.5}{\sqrt{20}}} = \frac{3.2}{0.335} = 9.55.$$

Since the above test is two-tailed, $\alpha = 0.05$ is halved which gives 0.025. The degree of freedom is $n - 1 = 19$. The critical value is now found at $t_{0.025,19} = 2.093$.

Decision rule: Since, the calculated value of $t = 9.55$ is greater than the critical (table) value, we reject the null hypothesis (H_0) and conclude that the average performance of students in the teacher's class in Chemistry is not 65%.

2. A farmer claims that his new variety of cassava is better than the old one. The two varieties of cassava were planted on the same plot of land for 10 years. The average yield of the old variety was 85kg with a standard deviation of 0.32 while the average yield of the new variety was 87.2 kg with a standard deviation of 0.35. Is there a significant evidence to believe the farmer's claim at 0.01 level of significance?

Solution:

The hypotheses for this test is given by,

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 < \mu_2$

In this case, the sample estimates are given as, $\bar{X}_1 = 85$, $S_1 = 0.32$, $\bar{X}_2 = 87.2$, $S_2 = 0.35$, $n_1 = 10$ and $n_2 = 10$.

Since the sample sizes are small, the population variances are not known and the observations are from different varieties, we employ an independent sample t-statistic to solve the problem.
Therefore,

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{9(0.32)^2 + 9(0.35)^2}{10 + 10 - 2} = \frac{2.024}{18} = 0.112.$$

$$t = \frac{87.2 - 85}{\sqrt{0.112\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{2.2}{\sqrt{0.0224}} = 14.67$$

The critical value is obtained at $t_{0.01,18} = 2.552$.

Decision rule: Since the calculated value is greater than the critical value, we do not accept the null hypothesis. We therefore conclude that there is significant evidence at 0.01 level of significance to believe the farmer's claim that his new variety of cassava is better than the old one.

3. A governorship aspirant in Imo State of Nigeria claims that he will secure 80% of votes to be cast in the next governorship election. A random sample of 50 eligible voters were selected and asked whom they were going to vote for in the election. Twenty eligible voters said they were going to vote for that governorship aspirant. Is this consistent with the governorship aspirant's claim at 0.1 level of significance?

Solution:

The required hypotheses are stated as,

$H_o : \pi = 0.8$

$H_1 : \pi \neq 0.8$

From the sample, we obtain that $p = 0.4$, $n = 50$.

Thus, the test statistic is given by,

$$Z = \frac{0.4 - 0.8}{\sqrt{\frac{0.8(1 - 0.8)}{50}}} = \frac{-0.4}{\sqrt{0.0032}} = \frac{-0.4}{0.057} = -7.018$$

Since, it is a two-tailed test, the level of significance 0.1 is halved. Thus, the critical value is given by,

$Z_{0.05} = 1.645.$

If we take the absolute value of the calculated Z and compare it with the critical value of Z, or a plot of the standard normal curve with -1.645 as the critical point, we will observe that,

the value -7.018 is in the critical region. Therefore, the null hypothesis is rejected. We therefore disregard the governorship aspirant's claim.

4. From two chemical processes, two samples each of size 10 were taken and the following results were obtained: $S_1^2 = 16.8$ and $S_2^2 = 4$. Test the hypotheses,

$H_0 : \sigma_1^2 < \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$ at 0.05 level of significance. $F = \frac{16.8}{4} = 4.2$.

The critical value is obtained at $F_{9,9,0.05} = 3.8$.

Decision: Since $F > F_{,9,9,0.05}$, we reject H_0 and accept H_1 and conclude that the first process varies more than the second process.

EXERCISE EIGHT

1. The following is true about a statistical hypothesis except:
   A. it is an assertion about one or more population parameters **B**. it is a conjecture about a population parameter(s) **C**. It is distribution free **D**. Its acceptance or rejection depends on the size of the critical region.
2. Which of the following is true about a null hypothesis?

   A. it always indicates that the parameter is in the critical region B. It always indicates that the parameter is not in the critical region C. it is a negation of itself D. It is a right-tailed test.
3. When a statistical hypothesis completely specifies the value of the parameters or the distribution of the random variables, it is called,
   A. null hypothesis B. alternative hypothesis C. left-tailed hypothesis D. simple hypothesis.
4. Given the hypothesis, $H_0 : \theta = 10$ versus $H_1 : \theta < 10$. The test is called,

   A. composite test B. left-tailed test C. right-tailed test D. power function.
5. A test of statistical hypothesis is,
   A. a rule which on consideration of the sample values leads to rejection or acceptance of the null hypothesis B. the size of the critical region C. a statistical statement with equality and inequality signs D. an intelligent guess of a statistical scientist.
6. A subset of the sample space which in accordance with a prescribed test leads to the rejection of the null hypothesis is called,

   A. type I error B. Type II error C. acceptance region D. critical region.
7. The test $H_0 : \theta = 5$ versus $H_1 : \theta \neq 5$ is called,

   A. right-tailed B. left-tailed test C. composite test D. two-sided test.
8. A statistical function which yields the probability that the sample point falls in the critical region of the test is called,

   A. parabolic function B. quadratic function C. power function D. linear function.
9. The value of the power function at a parametric point is called,
   A. mantissa B. power of the test C significance level D. critical value.
10. the error committed when the null hypothesis is rejected in error is called,
    A. Type II error B. Type I error C. mean square D. absolute error.
11. The error committed when the alternative hypothesis is rejected in error is called,
    A. Type II error B. Type I error C. mean square D. absolute error E.
12. The maximum value of the power function when the null hypothesis H_0 is true is called,

    A. critical value B. level of significance C. calculated value D. table value.
13. A test of hypothesis on one or two population means when the sample size is small and the population variances are not known is most appropriately done with,

A. Z-statistic B. chi-square C. t-statistic D. F-statistic.

14. A test of hypothesis on one or two means when the sample size is large ($n \geq 30$) is most appropriately done with,

A. Z-statistic B. chi-square C. t-statistic D. F-statistic.

15. A test of hypothesis on two population means when the samples are drawn from normally distributed populations with unknown variances assumed to be equal is most appropriately done with,

A. t-statistic B. Z-statistic C. chi-square D. F-statistic

16. The appropriate test statistic on means of two dependent observations when the sample sizes are small is,

A. $Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ B. $t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ C. $t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ D. $t = \frac{\overline{d}\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{(n-1)}}}$.

17. A student claims that he gets correctly at least 80% of questions he solves in any Physics examination. If in a given Physics examination, he got 39 questions correctly out of 60 questions. Identify the appropriate test statistic for this claim.

A. $Z = \frac{p - \pi_0}{\frac{\pi_0(1-\pi_0)}{n}}$ B. $t = \frac{p - \pi_0}{\frac{\pi_0(1-\pi_0)}{n}}$ C. $\chi^2 = \frac{p - \pi_0}{\frac{\pi_0(1-\pi_0)}{n}}$ D. $\chi^2 = \frac{p - p_0}{\frac{\pi_0(1-\pi_0)}{n}}$.

18. In a test of hypothesis, if the table value and the calculated value are used to take decision, what should be the appropriate decision rule?

A. reject H_0 if the calculated value is greater than the table value B. reject H_0 if the calculated value is less than the table value C. reject H_0 if the calculated value is equal to the tabulated value D. reject H_1 if the calculated value is greater than the tabulated value.

19. The appropriate test for equality of two variances is given by,

A. $F = \frac{s_1^2}{s_2^2}$ B. $\chi^2 = \frac{ns_1^2}{s_2^2}$ C. $t = \frac{s_1^2/n_1}{s_2^2/n_2}$ D. $F = \frac{\sigma_1^2}{\sigma_2^2}$.

20. A teacher claims that his new method of teaching trigonometry improves the understanding of the students in the topic. In order to test this claim, 60 students from the same class were randomly chosen and given tests on trigonometry before and after the introduction of the new method. The average performance of the students before the introduction of the new method was 65.8 with a standard deviation of 0.23 while the average performance of the students after the examination was 81.2 with a standard deviation of 0.5. Denoting the new method by 2 and before the new method by 1, formulate the appropriate null and alternative hypothesis for the claim.

A. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_2 > \mu_1$ B. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_2 < \mu_1$ C. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ D. $H_0 : \mu_1 > \mu_2$ versus $H_1 : \mu_2 < \mu_1$.

21. The average blood pressure of 28 patients was 125 with average deviation of 2.5. The average blood pressure of 26 other patients was 130 with average deviation of 2.01. Is there evidence to conclude that there is significant difference in their average blood pressure?

Denoting the first group by 1 and the second group by 2, you are required to formulate the appropriate null and alternative hypotheses for this test.

A. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_2 > \mu_1$ B. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_2 < \mu_1$ C. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ D. $H_0 : \mu_1 > \mu_2$ versus $H_1 : \mu_2 < \mu_1$.

22. In using a Z-statistic to test the hypothesis, $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ at 0.05 level of significance, at what value should the critical value of Z be obtained?
A. 0.05 B. 0.025 C. 0.01 D. 0.25 E. 0.5.

23. In using a Z-statistic to test the hypothesis $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 > \mu_2$ at 0.01 level of significance, at what value should the critical value of Z be obtained?
A. 0.001 B. 0.01 C. 0.025 D. 0.0001.

24. In test of hypothesis with Z-statistic, t-statistic or F-statistic, one of the following assumptions is common,
A. the data must be continuous B. homogeneity of variances C. homogeneity of means D. the data can be discrete or continuous.

25. In test of hypothesis involving chi-square, the data must be,
A. continuous data B. frequency data C. dichotomous data D. binary data.

26 Test the hypotheses for the problems in questions 25 and 26.

CHAPTER NINE

CORRELATION AND REGRESSION ANALYSES

9.0 CORRELATION ANALYSIS

INTRODUCTION

So far we have considered problems in which only one variate is measured on a random sample of observational units. In practice two or more variates are measured on the randomly selected observation units e.g (Birth length and Head circumference of a new baby, height of father and height of son e.t.c.). If two quantities vary in such a way that changes in one are accompanied by changes in the other, these quantities are said to be correlated. The degree of relationship between the variables under consideration is measured through the correlation analysis. It refers to the techniques used in measuring the closeness of the relationship between the variables. This method of analysis does not make any assumptions about the nature of the relationship between variables, that is, it does not assume that a given variable is the dependent variable while others are independent variable. Rather, it assumes that all the variables are random and only seeks to measure the strength of the relationships among them. The measure of correlation is called the correlation coefficient or correlation index which summarizes in one figure the direction and degree of correlation. Correlation is classified in several different ways. Three of the most important ways of classifying correlation are

- Positive (Direct) or Negative (Inverse)
- Simple, partial and multiple
- Linear and non linear

The various methods of ascertaining whether two variables are correlated or not are through

i. Graphic method
ii. Product moment correlation coefficient due to Karl Pearson
iii. Concurrent deviation method (Spearman rank correlation coefficient)

9.1 SCATTER DIAGRAM

The scatter diagram is known by many names, such as scatter plot, scatter graph, and correlation chart. This diagram is drawn with two variables, usually the first variable is independent and the second variable is dependent on the first variable.

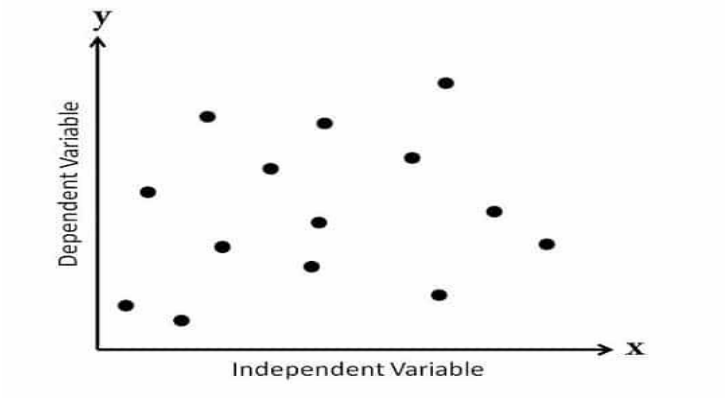


Figure 9.1: An example of scatter diagram

The scatter diagram is used to ascertain the nature of relationship between two variables. This diagram helps you determine how closely the two variables are related. After determining the correlation between the variables, you can then predict the behavior of the dependent variable based on the measure of the independent variable.

9.1.1 Type of Scatter Diagrams

The scatter diagram can be categorized into several types. However, let us discuss the two types that will cover most scatter diagrams used in study of relationship between two variables. The first type is based on the type of correlation, and the second type is based on the slope of trend.

According to the type of correlation, scatter diagrams can be divided into following categories:

- Scatter Diagram with No Correlation
- Scatter Diagram with Moderate Correlation
- Scatter Diagram with Strong Correlation

9.1.1.1 Scatter Diagram with No Correlation

This type of diagram is also known as "Scatter Diagram with Zero Degree of Correlation".

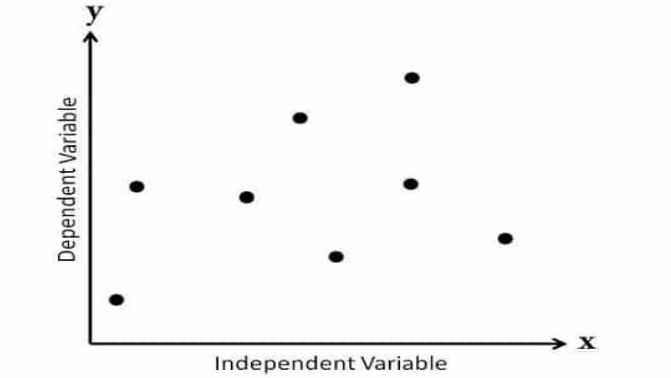


Figure 9.2: An example of scatter diagram with No Correlation

In this type of scatter diagram, data points are spread so randomly that you cannot draw any line through them.

In this case you can say that there is no relation between these two variables.

9.1.1.2 Scatter Diagram with Moderate Correlation

This type of diagram is also known as "Scatter Diagram with Low Degree of Correlation".

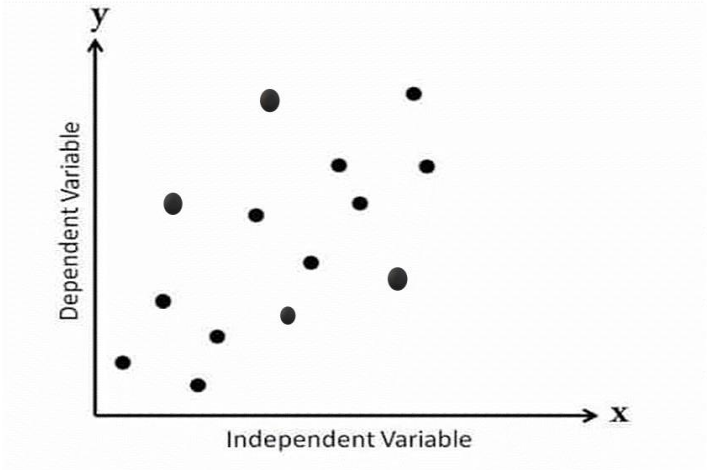


Figure 9.3: An example of scatter diagram with moderate Correlation

Here, the data points are little closer together and you can feel that only a little kind of relationship exists between these two variables.

9.1.1.3 Scatter Diagram with Strong Correlation

This type of diagram is also known as "Scatter Diagram with High Degree of Correlation".
In this diagram, data points are grouped very close to each other such that you can draw a line by following their pattern.

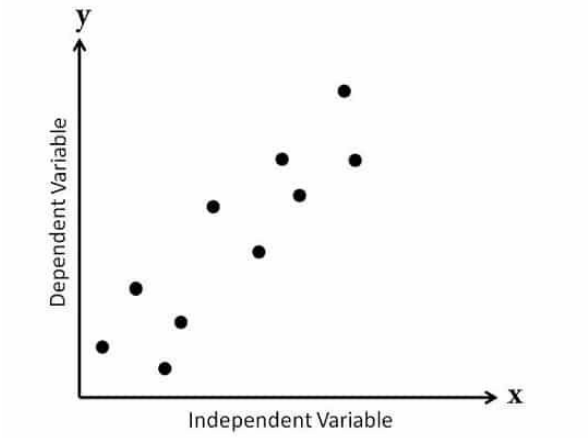


Figure 9.4: An example of scatter diagram with high Correlation

In this case you will say that the variables are closely related to each other.
As discussed earlier, you can also divide the scatter diagram according to the slope, or trend, of the data points:

- Scatter Diagram with Strong Positive Correlation
- Scatter Diagram with Weak Positive Correlation
- Scatter Diagram with Strong Negative Correlation
- Scatter Diagram with Weak Negative Correlation
- Scatter Diagram with Weakest (or no) Correlation

Strong positive correlation means there is a clearly visible upward trend from left to right; a strong negative correlation means there is a clearly visible downward trend from left to right. A weak correlation means the trend, up or down, is less clear. A flat line from left to right is the weakest correlation, as it is neither positive nor negative and indicates the independent variable does not affect the dependent variable.

The correlation coefficient between two variables X and Y is usually denoted by ρ_{XY} or simply ρ and its sample estimate by r_{xy}. In this manual, emphasis is on product moment correlation and Speraman rank correlation coefficient for two variates X and Y measured on a sample of n observation units.

9.2 PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT

The Pearson's product moment correlation coefficient can be calculated using the formula for a sample of size n $(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$

$$r_{xy}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{9.1}$$

It can be shown that

$$r_{xy}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=\frac{n\sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{\sqrt{\left(n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right)\left(n\sum_{i=1}^{n}Y_i^2-\left(\sum_{i=1}^{n}Y_i\right)^2\right)}} \tag{9.2}$$

The Karl Pearson's method is based on the assumption that the population from which the sample was collected is normally distributed. When the population is not normal or when the shape of the distribution is not known, there is need for a measurement of correlation that involves no assumption about the parameter of the population. Instead of using raw scores, we may now operate with the ranks of the measurements of each variable.

9.3 SPEARMAN RANK CORRELATION COEFFICIENT

A frequently used and simple procedure developed by the British Psychologist Charles Edward Spearman is the so-called Spearman rank correlation coefficient.
$(X_1,Y_2),(X_2,Y_2),\cdots,(X_n,Y_n)$ are pairs of values on the two variates

$$r_R=1-\frac{6\sum_{i=1}^{n}D_i^2}{n(n^2-1)}=1-\frac{6\sum_{i=1}^{n}D_i^2}{n^3-n} \tag{9.3}$$

where

r_R = rank coefficient of correlation

$D_i = (R_{xi} - R_{yi})$ the differences between ranks of the two variates on an observation unit

R_{xi} = rank on X_i variate

R_{yi} = rank on Y_i variate

n = total number of observations

It should be noted that for equal ranks, it is customary to give each individual an average rank.

9.4 REGRESSION ANALYSIS

When two variables are closely related (especially when one variable is random and the other is fixed), it may be of interest to estimate the value of one variable given the value of the another. Regression analysis reveals this relationship between two variables and this makes possible, estimation or prediction. Thus, regression analysis is a statistical tool which helps to predict one variable (the dependent variable) from the other variable or variables (the independent), on the basis of assumed nature of the relationship between the variables. The variable being predicted is usually referred to as the unknown or dependent variable because its values are dependent on the values of the other variable or variables referred to as the independent variables, explanatory variables, predetermined variables or predictor variables. The independent variable is denoted by X and the dependent variable by Y. A regression model may be simple or multiple, linear or non linear. It is simple if there is only one independent variable and multiple if there are more than one independent variables in the model. A regression model is linear, if its parameters do not contain any exponents and are not multiples of other parameters in the model, otherwise the model is said to be non-linear. In this manual, we consider only simple linear regression.

Regression equations are algebraic expressions of the regression lines. In simple linear regression, the model describing the relationship between X and Y can be expressed as given

$$Y_i = \alpha + \beta X_i + e_i \qquad (9.4)$$

where

α = intercept,

β = slope,

e_i = random error,

X_i = independent variable and

Y_i = dependent variable

We use $\hat{\alpha}$ and $\hat{\beta}$ to estimate α and β, hence the regression line becomes $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$

There are many methods of obtaining estimates of α and β. However, the method usually employed is least squares method because it yields line of best fit (that is line with the least error sum of square) and the line so obtained is called the least squares line.

According to the method of least squares, the best estimators $\hat{\alpha}$ and $\hat{\beta}$ of α and β are

$\hat{y}_i = \alpha + \beta x_i + e_i \ , (i = 1,2,...,n)$

$e_i = \hat{y}_i - \alpha - \beta x_i$

The direct regression approach minimizes the sum of squares $S(\alpha,\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ with respect to α and β

The partial derivatives of $S(\alpha,\beta)$ with respect to α is

$$\frac{\partial S(\alpha,\beta)}{\partial \alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)$$

and the partial derivative of $S(\alpha,\beta)$ with respect to β is

$$\frac{\partial S(\alpha,\beta)}{\partial \beta} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x$$

the solution of α and β are obtained by setting equation A and B to zero respectively

To obtained α

$$-2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$$

$$\sum_{i=1}^{n} y_i - n\alpha - \beta\sum_{i=1}^{n} x_i = 0$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} \qquad (9.5)$$

To obtained β

$$-2\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)X_i = 0$$

$$\sum_{i=1}^{n}\left(Y_i X_i - \alpha X_i - \beta X^2{}_i\right) = 0$$

Since $\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$, then

$$\sum_{i=1}^{n}\left(Y_i X_i - (\bar{Y} - \beta\bar{X})X_i - \beta X^2{}_i\right) = 0$$

$$\sum_{i=1}^{n}\left(Y_i X_i - \bar{Y}X_i + \beta\bar{X}X_i - \beta X^2{}_i\right) = 0$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y} + \beta\bar{X} - \beta X_i)X_i = 0$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y} + \beta\bar{X} - \beta X_i) = 0$$

$$\sum_{i=1}^{n}(Y - \bar{Y}) = -\hat{\beta}\sum_{i=1}^{n}(\bar{X} - X_i)$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{9.6}$$

It can be shown that

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{n\sum_{i=1}^{n}X_iY_i - \sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{n\sum_{i=1}^{n}X_i^2 - \left(\sum_{i=1}^{n}X_i\right)^2} \tag{9.7}$$

Example 9.1

Calculate the correlation coefficient from the data given below

X	9	8	7	6	5	4	3	2	1
Y	15	16	14	13	11	12	10	8	9

Solution:

Calculation of correlation coefficient

X	Y	X^2	Y^2	XY
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$

$n = 9$,

$$r_{xy} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left(n\sum X^2 - (\sum X)^2\right)\left(n\sum Y^2 - (\sum Y)^2\right)}}$$

$$r_{xy} = \frac{9\times 597 - 45\times 108}{\sqrt{\left(9\times 285 - (45)^2\right)\left(9\times 1356 - (108)^2\right)}}$$

$= 0.95$

Interpretation: Since the correlation coefficient (r) is positive and close to 1, then the variables are positively strongly related.

Example 9.2

The ranking of 10 students in two subjects A and B are as follows:

A	6	5	3	10	2	4	9	7	8	1
B	3	8	4	9	1	6	10	7	5	2

Calculate the rank correlation coefficient.

Solution

R_A	R_B	$D=(R_A-R_B)$	D^2
6	3	3	9
5	8	-3	9
3	4	-1	1
10	9	1	1
2	1	1	1
4	6	-2	4
9	10	-1	1
7	7	0	0
8	5	3	9
1	2	-1	1
			$\sum D^2 = 36$

$$r_R = 1 - \frac{6\sum D^2}{n^3 - n} \quad 、$$

$$= 1 - \frac{6 \times 36}{10^3 - 10} = 0.782$$

Note: in the above illustration, actual ranks were given.

Example 9.3

Calculate the Spearman's coefficient of rank correlation between marks assigned to ten students by Judges X and Y in a certain competition test as shown below:

S/N	1	2	3	4	5	6	7	8	9	10
Marks by Judge X	52	53	42	60	45	41	37	38	25	27
Marks by Judge Y	65	68	43	38	77	48	35	30	25	50

Solution

Firstly, we assign ranks and then calculate rank correlation coefficient

S/N	X	Y	R_X	R_Y	$D=(R_X-R_Y)$	D^2
1	52	65	8	8	0	0
2	53	68	9	9	0	0
3	42	43	6	5	1	1
4	60	38	10	4	6	36
5	45	77	7	10	-3	9
6	41	48	5	6	-1	1
7	37	35	3	3	0	0
8	38	30	4	2	2	4
9	25	25	1	1	0	0
10	27	50	2	7	-5	25
						$\sum D^2 = 76$

$$r_R = 1 - \frac{6\sum D^2}{n^3 - n} = 1 - \frac{6 \times 76}{10^3 - 10} = 0.539$$

Note: Here, the actual data were given and ranks assigned. Ranks can be assigned by taking either highest value as 1 or the lowest value as 1. For equal ranks, the rank assigned is the average of the ranks which these individuals would have got had they differed slightly from each other.

Example 9.4

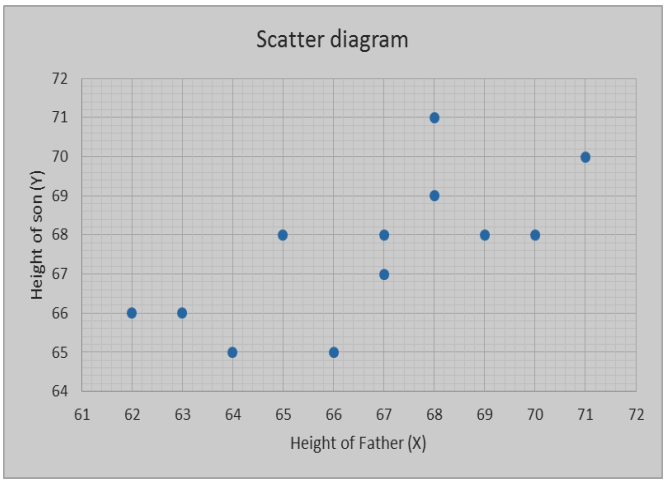
The table below shows the respective heights X and Y of a sample of 12 fathers and their eldest sons.

- Construct a scatter diagram and comment
- Find the least-squares regression line of Y on X

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

Solution

i.



Comment: The Scatter diagram revealed that the two variables are moderately correlated with positive value.

ii. Regression line of Y on X is given $Y = a + bX$

$n = 12$

$$\bar{Y} = \frac{\sum_{i=1}^{12} Y_i}{n} = \frac{811}{12} = 67.58$$

$$\bar{X} = \frac{\sum_{i=1}^{12} X_i}{n} = \frac{800}{12} = 66.67$$

$$\hat{\beta} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{12 \times 54107 - 800 \times 811}{12 \times 53418 - 800^2} = 0.476$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 67.58 - 0.476 \times 66.67 = 35.82$$

Therefore, the regression line of Y on X is

$$\hat{Y}_i = 35.82 + 0.476X_i$$

Height of Father (X)	Height of Son (Y)	X^2	Y^2	XY
65	68	4225	4624	4420
63	66	3969	4356	4158
67	68	4489	4624	4556
64	65	4096	4225	4160
68	69	4624	4761	4692
62	66	3844	4356	4092
70	68	4900	4624	4760
66	65	4356	4225	4290
68	71	4624	5041	4828
67	67	4489	4489	4489
69	68	4761	4624	4692
71	70	5041	4900	4970
$\sum X = 800$	$\sum Y = 811$	$\sum X^2 = 53418$	$\sum Y^2 = 54849$	$\sum XY = 54107$

EXERCISE NINE

1. Obtain the rank correlation coefficient between the variables X and Y from the following pairs of observed values

X	50	55	65	50	55	60	50	65	70	75
Y	110	110	115	125	140	115	130	120	115	160

2. Two ladies were asked to rank 7 different types of lipsticks. They gave the following ranks

LIPSTICKS	A	B	C	D	E	F	G
CHIOMA	2	1	4	3	5	7	6
NKECHI	1	3	2	4	5	6	7

Calculate Spearman's rank correlation coefficient.

3. Quotations of index numbers of security prices of a certain joint stock company are given below

Year	1	2	3	4	5	6	7
Debenture Price	97.8	99.2	98.8	98.3	98.4	96.7	97.1
Share Price	73.2	85.8	78.9	75.8	77.2	87.2	83.8

Using rank correlation method, determine the relationship between debenture prices and share prices

4. Calculate the coefficient of rank correlation from the following data

Price of Tea	75	88	95	70	60	80	81	50
Price of Coffee	120	134	150	115	110	140	142	100

5. The table below shows how 10 students were ranked according to their achievements in both the laboratory and lecture portions of a Chemistry course. Find the coefficient of rank correlation.

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

6. The average prices of stocks and bonds listed on country XYZ during the years 2000 – 2009 are given in the table below. Find the correlation coefficient and interpret the results.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Average Price of Stocks	35.22	39.87	41.85	43.23	40.06	53.29	54.14	49.12	40.71	55.15
Average price of bonds	102.43	100.93	97.43	97.81	98.32	100.07	97.08	91.59	94.85	94.65

7. Using the data below, find the correlation coefficient between the two sets of quiz grades

Grade On First Quiz(X)	6	5	8	8	7	6	10	4	9	7
Grade On Second Quiz (Y)	8	7	7	10	5	8	10	6	8	6

8. The table below shows the ages X and systolic blood pressures Y of 12 women. (i) Find the correlation coefficient between X and Y (ii) Determine the least squares regression line of Y on X (iii) Estimate the blood pressure of a woman whose age is 45 years .

Age (X)	56	42	72	36	63	47	55	49	38	42	68	60
Blood pressure (Y)	147	125	160	118	149	128	150	145	115	140	152	155

9. Refer to the table on question (7), find the least – squares regression line of (i) Y on X and

10. The ranks of the same 15 students in two subjects A and B are given below. The two numbers within brackets denote the ranks of the same student in A and B respectively. (1, 10), (2, 7), (3, 2), (4, 6), (5, 4), (6, 8), (7, 3), (10, 1), (9, 1), (10, 1), (9, 1), (10, 15), (11, 19), (12, 5), (13, 14), (14, 12), (15, 13). Find the Spearman's rank correlation coefficient.

11. Fit a Least-squares regression line of Y on X to the data in the table below

X	0.5	1.0	1.5	2.0	2.5	3.0	3.5
Y	10	12	14	16	17	18	21

12. Given the data below, find the linear correlation coefficient between the variables X and Y

X	1.2	1.8	3.1	4.9	5.7	7.1	8.6	9.8
Y	4.5	5.9	7	7.8	7.2	6.8	4.5	2.7

13. Two judges in a contest, who were asked to rank 8 candidates A, B, C, D, E, F, G and H, in order of their preference, submitted the choices shown in the table below. Find the coefficient of rank correlation.

CANDIDATE	A	B	C	D	E	F	G	H
JUDGE 1	5	2	8	1	4	6	3	7
JUDGE 2	4	5	7	3	2	8	1	6

14. The following data represent scores on a math (X) test and a reading (Y) test given to a class of 14 sixth grade students.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14
(X)	97	68	85	74	92	92	100	63	85	87	81	93	77	82
(Y)	89	57	87	76	97	79	91	50	85	84	91	91	75	77

i. Calculate the Pearson r correlation between math and reading scores for this group of students.

15. The following data represent test scores and the sophomore average of 7 Students

Student	1	2	3	4	5	6	7
Test score (X)	70	90	75	85	80	70	90
Sophomore average (Y)	2.5	4.0	3.5	3.0	3.0	2.0	3.0

i. Calculate the Pearson r correlation between the sophomore average (Y) and the test score (X).
ii. Compute the linear regression equation for predicting the sophomore average (Y) from the test score (X).

16. The time, X, in years that an employee spent at a company and the employee's hourly pay, Yy, for 5 employees are listed in the table below. Calculate and interpret the correlation coefficient r.

TIME (X)	5	3	4	10	15
HOUR (Y)	25	20	21	35	38

17. The table below shows the number of absences, X, in Statistics course and the final exam grade, Y, for 7 students. Find the correlation coefficient and interpret your result.

X	1	0	2	6	4	3	3
Y	95	90	90	55	70	80	85

18. The table below shows the height, X, in inches and the pulse rate, Y, per minute, for 9 people. Find the correlation coefficient and interpret your result.

X	68	72	65	70	62	75	78	64	68
Y	90	85	88	100	105	98	70	65	72

19. Five children aged 2, 3, 5, 7 and 8 years old weigh 14, 20, 32, 42 and 44 kilograms respectively.
   i. Find the equation of the regression line of age on weight.
   ii. Based on this data, what is the approximate weight of a six year old child?

20. The success of a shopping center can be represented as a function of the distance (in miles) from the center of the population and the number of clients (in hundreds of people) who will visit. The data is given in the table below:

No. Customer (x)	8	7	6	4	2	1
Distance (y)	15	19	25	23	34	40

   i. Calculate the linear correlation coefficient.
   ii. If the mall is located 2 miles from the center of the population, how many customers should the shopping center expect?
   iii. To receive 5 customers, at what distance from the center of the population should the shopping centre be located?

21. The grades over 10 of five students in STA 211 and MTH 201 classes are:

STA 211	6	4	8	5	3. 5
MTH 201	6. 5	4. 5	7	5	4

Determine the regression lines and calculate the expected grade in MTH 201 for a student who has a 7.5 in STA 211.

22. The heights (in centimeters) and weight (in kilograms) of 10 basketball players on a team are:

Height (X)	186	189	190	192	193	193	198	201	203	205
Weight (Y)	85	85	86	90	87	91	93	103	100	101

Calculate:
   i. The regression line of y on x.
   ii. The coefficient of correlation.
   iii. The estimated weight of a player who measures 208 cm.

23. From the following data of hours worked in a factory (X) and output units (Y), determine the regression line of y on x, the linear correlation coefficient and determine the type of correlation.

Hours (X)	80	79	83	84	78	60	82	85	79	84	80	62
Production (Y)	300	302	315	330	300	250	300	340	315	330	310	240

24. The following table summarizes the results of an aptitude test given to six clerks to determine the correlation between test scores (X) and sales in the first month (Y) in hundreds of dollars.

X	25	42	33	54	29	36
Y	42	72	50	90	45	48

i. Find the correlation coefficient and interpret the results.
ii. Calculate the regression line of y on x and predict the sales of a vendor who obtains 47 on the test.

25. The number of offenses committed in the past year by four drivers of a transport company and their respective experience in years is represented by the following table:

Years (X)	3	4	5	6
Offenses (Y)	4	3	2	1

Calculate the linear correlation coefficient and interpret it.

CHAPTER TEN
ANALYSIS OF VARIANCE

10.0 INTRODUCTION

The **Analysis of Variance (ANOVA)** is a technique of decomposing the total variability of a response variable into: variability due to the experimental factor(s) and variability due to error (i.e., factors that are not accounted for in the experimental design). The basic purpose of ANOVA is to test the equality of several means.

For example, a cereal field is divided into a number of plots, each plot 'treated' with a different manure to see which produces the most cereal. Consider, for example, a sample which consists of p sub-samples, we wish to know whether the total sample may be regarded as homogeneous or alternatively there is some indication that the sub-samples were drawn from different populations. To take a more complex case, we may have a number of observations taken by p different observers each on a sample affected by q different effects, as for instance, if p laboratory assistants carry out an assay on samples of a drug from q different suppliers. Our classification here is two-fold and we wish to discuss whether there are any significant differences between the q sources of supply and, independently if possible, whether there are any differences between the results obtained by the p assistants.

10.1 Basic Concepts

Let us define some of the important concepts in design of experiments. We have already seen the concepts treatment, experimental unit, and response, but we define them again here for completeness.

Observation units, sometimes also called**statistical units**, are the entities on which information is received and statistics are compiled in the process of collecting statistical data. An **observation**contains information, at a particular period, of a particular variable, such as the individual price of an item at a given outlet. Observation units vary according to the specific survey or data collection:

Treatments are the different procedures we want to compare. These could be different kinds or amounts of manures in agronomy, different long-distance rate structures in marketing, or different temperatures in a reactor vessel in chemical engineering. For example, a cereal field is divided into a number of plots, each plot 'treated' with a different manure to see which produces the most cereal; a teacher practices different teaching methods on different groups in her class to see which yields the best results; a doctor treats a patient with a skin condition with different creams to see which is most effective.

A factor of an experiment is a controlled independent variable; a variable whose levels are set by the experimenter. Factors combine to form treatments. For example, the baking treatment for a cake involves a given time at a given temperature. The treatment is the

combination of time and temperature, but we can vary the time and temperature separately. Thus we speak of a time factor and a temperature factor.

Levels of a Factor are individual settings for a factor. Different treatments constitute different levels (imply amounts or magnitudes) of a factor. For example, three different groups of runners are subjected to different training methods. The runners are the experimental units, the training methods, the treatments; where the three types of training methods constitute three levels of the factor 'type of training'.

Treatment effects are deviations from the grand (overall) mean. Treatment effect is a unique parameter to treatments (see Equation 10.2).

Randomizationis the use of a known, understood probabilistic mechanism for the assignment of treatments to units. Other aspects of an experiment can also be randomized: for example, the order units are evaluated for their responses.

Two forms of ANOVA can be discussed in this manual. These are:

a. One-way ANOVA
b. Two-way ANOVA

Each of these two forms of ANOVA is discussed separately.

10.2 MODEL FOR ONE-WAY ANOVA

A brief introduction of the model for one-way ANOVA and its assumptions is presented here. Recall that we have *p* treatments or different levels of a single factor that we wish to compare. The observed response from each of the *p* treatments is a random variable. The data would appear as in Table 10.1. An entry in Table 10.1, say x_{ij}, represents the *i*th observation taken under treatment *i*. There will be in general n_i observations under the *i*th treatment.

Model Definition:

A linear statistical formula

$$x_{ij} = \mu + \mu_i + \varepsilon_{ij}, \quad i = 1, 2, ..., p, \; j = 1, 2, ..., n_i \tag{10.1}$$

that describes the observations in Table 10.1, where x_{ij} is the (i,j)th observation, μ is a parameter common to all treatments called the grand mean, μ_i is a parameter unique to the *i*th treatment effect, and ε_{ij} is a random error component. Our objective would be to test appropriate hypotheses about the treatment effects and to estimate them.

Assumptions of the model

In building the model in Equation 10.1, there are certain basic assumptions that are made.

- Observations x_{ij} are independent

- ε_{ij} are normally distributed with mean zero and constant variance.
- $\sum_{i=1}^{p} n_i \mu_i = 0$

Assumption (ii) implies that the response variable for each group is normal.

Note: the model in Equation 10.1 represents two situations.

The p treatments could have been specifically chosen by the experimenter. In this situation we wish to test hypotheses about the treatment means, and conclusions would apply to the factor levels considered in the analysis. The conclusions cannot be extended to similar treatments that were not explicitly considered. We may also wish to estimate the model parameters (μ, μ_i, σ^2). This situation is called the fixed effects model.

The p treatments could be a random sample from a larger population of treatments. In this situation we should like to be able to extend the conclusions (which are based on the sample of treatments) to all treatments in the populations, whether they were explicitly considered in the analysis or not. Here the μ_i are random variables, and knowledge about the particular ones investigated is useless. Instead, we test hypotheses about the variability of the μ_i and try to estimate this variability. This is called the random effects, or components of variance, model.

10.2.1 One-Way Classification

Assume there are a number of populations of interest each of which is comprised of a number of experimental observations as shown in Table 10.1.

Table 10.1 Layout of One – Way Classification

$$\begin{bmatrix} \underline{t_1} & \underline{t_2} & \cdots & \underline{t_p} \\ x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ x_{13} & x_{23} & \cdots & x_{p3} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{x_{1n_1}} & \underline{x_{2n_2}} & \cdots & \underline{x_{pn_p}} \\ \bar{x}_{1\bullet} & \bar{x}_{2\bullet} & \cdots & \bar{x}_{p\bullet} \end{bmatrix}$$

Required equation

$x_{ij} = \mu + \mu_i + \varepsilon_{ij}, \quad i = 1, 2, ..., p, \ j = 1, 2, ..., n_i$

where

μ = the grand mean

μ_i =the effect of ith treatment mean and

ε_{ij} =error term in the (i,j)th cell.

Each population is referred to as a treatment. Suppose there are p treatments, and let $\mu_1,\ldots,\mu_p$ be the mean of the experimental observations in each of the p treatments. Suppose there are n_i experimental observations in treatment i

The technique of ANOVA involves testing the null hypothesis

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$

against the alternative hypothesis

H_1 : at least one is significantly different from the others.

The total number of observations in all treatments is given as

$$n = \sum_{i=1}^{p} n_i \tag{10.2}$$

The sample mean of treatment i as shown in Table 10.1 is given as:

$$\bar{x}_{i\bullet} = \frac{1}{n_i}\left(x_{1j} + x_{2j} + \ldots\ldots + x_{2n_i}\right) = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \tag{10.3}$$

$\bar{x}_{i.}$ is the estimate of the ith population mean, μ_i

where x_{ij} is the j^{th} observation in the ith treatment. The sample mean of all treatments or the grand mean is given as:

$$\bar{x}_{\bullet\bullet} = \frac{\sum_{i=1}^{p} n_i \bar{x}_{i\bullet}}{p} \tag{10.4}$$

$\bar{x}_{..}$ is the estimate of the common mean μ, under the null hypothesis, each $\bar{x}_{i.}$ is estimating the grand mean μ.

The sample variance of observations within each treatment (i) is given as:

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\bullet})^2 \tag{10.5}$$

The total sum of squares (SST), treatment sum of squares (SSTR) and error sum of squares (SSE) are given as:

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\bullet\bullet})^2 \tag{10.6}$$

$$SSTR = \sum_{i=1}^{p} n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 \tag{10.7}$$

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\bullet})^2 \tag{10.8}$$

It can be shown that the three sums of squares are related by:

$$SST = SSTR + SSE$$

i.e. $$\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_{..}\right)^2 = \sum_{i=1}^{p} n_i\left(\bar{x}_{i.}-\bar{x}_{..}\right)^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_{i.}\right)^2 \quad (10.9)$$

The computational formulae for the SST, SSTR and SSE are given below

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{n_i}\sum_{j=1}^{p} x_{ij}\right)^2 \quad (10.10)$$

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n}G^2 \quad (10.11)$$

where $$G = \sum_{i=1}^{n_i}\sum_{j=1}^{p} x_{ij}$$

$$SSTR = \sum_{i=1}^{p}\frac{x_{i\bullet}^2}{n_i} - \frac{1}{n}G^2 \quad (10.12)$$

$$SSE = SST - SSTR \quad (10.13)$$

The mean square for treatments (MSTR) and the mean square for error (MSE) are given as:

$$MSTR = \frac{SSTR}{p-1} \quad (10.14)$$

$$MSE = \frac{SSE}{n-1} \quad (10.15)$$

To test the null hypothesis, the test statistic is defined as:

$$F_C = \frac{MSTR}{MSE} \quad (10.16)$$

Note that the statistic in (10.11) is zero if $\bar{x}_{i.} = \bar{x}_{..}$ for all i i.e. when there is no treatment effect. The more it differs from zero the more it indicates that there is treatment effect.

This ratio in (10.11) is distributed as F with $df1 = p-1$ (for numerator degrees of freedom) and $df2 = n-p$ (for denominator degrees of freedom) respectively, if the hypothesis of homogeneity of means is true. If the observed value (F_C) of this ratio exceeds the tabulated $F_{1-\alpha,df1,df2}$, the null hypothesis, H_0 is rejected at α level of significance.

The calculations shown above can be summarized in what is commonly referred to as ANOVA Table given in Table 10.2

Table 10.2: ANOVA Table for One-way Classification

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
Treatments: Error:	$p-1$ $n-p$	SSTR SSE	$MSTR$ MSE	MSTR/MSE -
Total:	$n-1$	SST	-	-

10.2.2 Multiple Contrasts

When the null hypothesis indicates that there is treatment effect, an analyst can identify the treatment(s) that are different from the others using many tests. However, the test most commonly in use are the Fisher least significant difference (LSD) test, the Tukey Kramer test and the Scheffe test. Critical values for these tests are as given in Table 10.3.

Table 10.3: Test Statistics for multiple Comparison and their Critical Values

Test	Critical Value	Notes
Fisher (LSD)	$t_{\alpha/2,n-p}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)}$	$t_{\alpha/2,n-p}$ is the critical value of the Students' t distribution.
Tukey-Kramer	$q_{\alpha}(p,n-p)\sqrt{\frac{MSE}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$q_{\alpha}(p,n-p)$ is the critical value of the Studentised range distribution.
Scheffe	$\sqrt{\left(\frac{2(p-1)F_{(\alpha/2,\,p-1,n-p)}MSE}{n}\right)}$	$F_{\alpha/2,\,p-1,n-p}$ is the critical value of the Fisher's F distribution.

If $|\bar{x}_i - \bar{x}_{i'}|$ < critical value then it is concluded that there is no significant difference between the means of the two treatments being tested.

Note: For this chapter, only the Fisher's LSD was used for the multiple comparison.

Example 10.1: In an experiment to compare the safety of compact cars, midsize cars, and full-size cars. The NTSB collected a sample of three each, of pressure applied to the driver's head during a crash (the treatments) Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each type of cars. Use α = 5%.

Table 10.4: Sample data on car types

	Compact cars	Midsize cars	Full-size cars	Total Sum
	643	469	484	
	655	427	456	
	702	525	402	
$x_{i\bullet}$	2000	1,421	1,342	4,763
$\bar{x}_{i\bullet}$	666.6667	473.6667	447.333	
$\sum_{j=1}^{n_i} x_{ij}^2$	1,335,278	677,915	603,796	2,616,989

$$x_{\bullet\bullet} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij} = 4{,}763$$

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij}^2 = 2{,}616{,}989$$

Solution:

STEP 1: State the model for the data
The model for the data is

$x_{ij} = \mu + \mu_i + \varepsilon_{ij}, \quad i = 1, 2, ..., p, \; j = 1, 2, ..., n_i$

STEP 2: State the null and alternative hypotheses

The null hypothesis for an ANOVA always assumes the population means are equal. Hence, we may write the null hypothesis as:

$H_0 : \mu_1 = \mu_2 = \mu_3$ - (i.e. the mean head pressure is the same for the three types of cars).

And, the alternative hypothesis is:

$H_1 : \mu_i \neq \mu_{i'}, \; i \neq i'$

i.e. the alternative hypothesis states that at least mean head pressure for one of the car types is different from the others.

STEP 3: Find the Critical Value $F_{\alpha, df1, df2}$

To obtain the critical value $F_{\alpha, df1, df2}$ from an F distribution, we determine the numerator and denominator degrees of freedom df1 and df2 , along with the significance level (α).

$F_{\alpha, df1, df2}$ has df1 and df2 degrees of freedom, where df1 is the numerator degrees of freedom, which is equal to p-1 and df2 is the denominator degrees of freedom and equals n - p.

In our example, df1 = 3 - 1 = 2 and df2 = 9 - 3 = 6. Hence we need to find $F_{0.05,2,6}$. Using the F tables in your text $F_{0.05,2,6}$ corresponding to α = 5%. Level of significance is $F_{0.05,2,6} = 5.14$.

The decision rule is therefore, to reject H_0 if $F_c > 5.14$ or do not reject it otherwise.

STEP 4: Calculate the appropriate test statistic

The test statistic in ANOVA is the ratio of the between and within mean treatment variation in the data. The SST, SSTR and SSE are computed using Equations (10.5) to (10.7), with p=3 and $n_i = 3$ for all *i*

Using

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n}G^2$$

$$SST = 2{,}616{,}989 - \frac{1}{9}(22{,}686{,}169)$$

$$\text{SST} = 2{,}616{,}989 - 2{,}520{,}685.44 = 96{,}303.56$$

Between Sum of Squares (or Treatment Sum of Squares) – variation in the data between the different types of cars (or treatments).

Treatment Sum of Squares:

$$SSTR = \sum_{i=1}^{p}\frac{x_{i\bullet}^2}{n_i} - \frac{1}{n}G^2$$

$$SSTR = \frac{1}{3}\left(2000^2 + 1421^2 + 1342^2\right) - \frac{1}{9}(22{,}686{,}169)$$

$$\text{SSTR} = 2{,}606{,}735 - 2{,}520{,}685.44 = 86{,}049.56$$

Within variation (or Error Sum of Squares) – variation in the data from each individual treatment.

$$SSE = SST - SSTR$$

$$SSE = 96{,}303.56 - 86{,}049.56 = 10{,}254.00$$

The next step in an ANOVA is to compute the "average" sources of variation in the data using SST, SSTR and SSE.

Mean Square Treatment: $MSTR = \frac{SSTR}{p-1}$, i.e., "average between variation" (with p = 3)

$$MSTR = \frac{86049.55}{(3-1)} = 43024.78$$

Mean Square Error: $MSE = \frac{SSE}{n-p}$, i.e., "average within variation"

$$MSE = \frac{10254}{(9-3)} = 1709$$

Note: Mean squares are not additive, i.e., $MST \neq MSTR + MSE$

The test statistic may now be calculated. For a one-way ANOVA the test statistic is equal to the ratio of $MSTR$ and MSE. This is the ratio of the "average between variation" to the "average within variation." In addition, this ratio is known to follow an F distribution.

Hence,

$$F_c = \frac{MSTR}{MSE} = \frac{43024.78}{1709} = 25.17$$

The summary of all the computations are given as ANOVA Table.

Table 10.2: ANOVA Table for One-way Classification

SV	DF	SS	MS	F
Car types: Error:	3-1=2 9-3=6	86049.55 10254.00	43024.78 1709	25.17
Total:	9-1=8	26303.55	-	-

The intuition here is relatively straightforward. If the average between variations rises relative to the average within variation, the F statistic will rise and so will our chance of rejecting the null hypothesis.

STEP 4: Make the decision

The null hypothesis is rejected if: $F_c > F_{0.05,2,6}$. In our example 25.17 > 5.14, so we reject the null hypothesis.

STEP 5: Multiple Comparison

Since the null hypothesis is rejected at $\alpha = 0.05$ level of significance, it indicates that the mean head pressure of at least one of the cars is significantly different from the others with 95% confidence, we do not yet know which mean(s) is/are different. The ANOVA test will tell us that at least one mean is different from others, but does not tell us the ones that is different. Therefore, in order to determine the mean(s) that are significantly different from the others, the methods in Table 10.3 are used. This is illustrated here with the most commonly used test, the Fisher LSD Test. For a balanced design the LSD is given as

$t_{\alpha/2,n-p}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)}$, where MSE is the mean square error and n_i is the number of rows in each treatment. In the example above, MSE = 1709, p = 3, n = 9, $n_i = 3, \forall i = 1, 2, 3.$

for all *i* and $t_{0.025,6} = 2.447$

Hence, the $LSD = t_{\alpha/2,6}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)} = 2.447\sqrt{\frac{2(1709)}{3}} = 82.60$

Thus, if the absolute value of the difference between any two treatment means is greater than 82.61, we may conclude that they are not statistically equal.

Compact cars vs. Midsize cars: 666.67 − 473.67 = 193. Since 193 > 82.61, i.e., mean head pressure is statistically higher for compact than midsize cars. There is no need comparing compact and full size since it is less than midsize cars.

Note: a design is balanced if $n_i = m, \forall i$

Midsize cars vs. Full-size cars: 473.67 − 447.33 = 26.34. Since 26.34 < 82.61, i.e., mean head pressure is statistically higher for equal between midsize and full-size cars.

The comparison can be summarized in a tabular form as:

Table 10.6: Differences between means of pairs of Treatments or car types.

$\bar{x}_i$	Compact 666.67	Midsize 473.67	Full size 447.33
Compact	-	193	219.34
Midsize	193	-	-26.34
Full size	219.34	-26.34	-

Example 10.2: A researcher wishes to try three different techniques to lower blood pressure. The subjects are randomly assigned to three groups; the first group takes medication, the second group exercises, and the third group follows a special diet. After four weeks, the reduction in each person's blood pressure is recorded. At $\alpha = 0.05$ test the claim that there is no difference among the methods producing blood pressure. The data are shown in Table 10.7.

Table 10.7: Data on three techniques to lower blood pressure

	Medication	Exercise	Diet	Total Sum
	10	6	5	
	12	8	9	
	9	3	12	
	15	0	8	
	13	2	4	
$x_{i\bullet}$	59	19	38	116
$\sum_{j=1}^{n_i} x_{ij}^2$	719	113	330	1,162

$$x_{\bullet\bullet} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij} = 116$$

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij}^2 = 1{,}162$$

Solution:

STEP 1: State the model for the data
The model for the data is
$x_{ij} = \mu + \mu_i + \varepsilon_{ij}$, $i = 1, 2, ..., p$, $j = 1, 2, ..., n_i$.

STEP 2: State the null and alternative hypotheses
The null hypothesis for an ANOVA always assumes the population means are equal. Hence, we may write the null hypothesis as:

$H_0: \mu_1 = \mu_2 = \mu_3$ - (i.e. the different techniques are equally effective).

And, the alternative hypothesis is:

$H_1: \mu_i \neq \mu_{i'}, \ i \neq i'$

i.e. the alternative hypothesis states that at least one mean is different from the others).

STEP 3: Find the Critical Value $F_{\alpha,df1,df2}$

In our example, df1 = 3 - 1 = 2 and df2 = 15 - 3 = 12. Hence we need to find $F_{0.05,2,12}$ using the F table. $F_{0.05,2,12}$ = 3.89.

The decision rule is therefore, to reject H_0 if $F_c > 3.89$ or do not reject it otherwise.

STEP 4: Calculate the appropriate test statistic

The test statistic in ANOVA is the ratio of the between and within mean treatment variation in the data. The SST, SSTR and SSE are computed using Equations (10.5) to (10.7), with p=3 and $n_i = 3$ for all *i*

Using

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n}G^2$$

$$SST = 1{,}162 - \frac{1}{15}(13456)$$

$$\text{SST} = 1{,}162 - 897.0667 = 265$$

Between Sum of Squares (or Treatment Sum of Squares) – variation in the data between the different types of cars (or treatments).

Treatment Sum of Squares:

$$SSTR = \sum_{i=1}^{p} \frac{x_{i\bullet}^2}{n_i} - \frac{1}{n}G^2$$

$$SSTR = \frac{1}{5}\left(59^2 + 19^2 + 38^2\right) - \frac{1}{15}(13456)$$

$$\text{SSTR} = 1057.2 - 897.0667 = 160$$

Within variation (or Error Sum of Squares) – variation in the data from each individual treatment.

$$SSE = SST - SSTR$$

$$SSE = 265 - 160 = 105$$

The next step in an ANOVA is to compute the "average" sources of variation in the data using SST, SSTR and SSE.

Mean Square Treatment: $MSTR = \frac{SSTR}{p-1}$, i.e., "average between variation" (with p = 3)

$$MSTR = \frac{160}{3-1} = \frac{160}{2} = 80$$

Mean Square Error: $MSE = \frac{SSE}{n-p}$, i.e., "average within variation"

$$MSE = \frac{105}{15-3} = \frac{105}{12} = 8.75$$

$$F_C = \frac{MSTR}{MSE} = \frac{80}{8.75} = 9.14$$

Table 10.2: ANOVA Table for One-way Classification

SV	DF	SS	MS	F
Blood Pressure Technique	3-1= 2	160	80	9.14
Error:	15-3=12	105	8.75	
Total:	15-1=14	265	-	-

STEP 4: Make the decision

In our example 7.14 > 3.89, so we reject the null hypothesis and conclude that the blood pressure techniques differs.

Example 10.3: A state employee wishes to see if there is significant difference in the number of employees at the interchanges of three state toll roads. The data are shown. At $\alpha = 0.05$ can it be concluded that there is a significant difference in the average number of employees at each interchange?

Table 10.8: Data on employees at toll roads

	Toll Road A	Toll Road B	Toll Road C	Total Sum
	7	10	1	
	14	1	12	
	32	1	1	
	19	0	9	
	10	11	1	
	11	1	11	
$x_{i\bullet}$	93	24	35	152
$\sum_{j=1}^{n_i} x_{ij}^2$	1851	224	349	2424

Solution:

$$x_{\bullet\bullet} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij} = 152$$

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij}^2 = 2424$$

Solution:

STEP 1: State the model for the data

The model for the data is

$$y_{ij} = \mu + t_i + \varepsilon_{ij}, \quad i = 1,2,\ldots,p,\ j = 1,2,\ldots,n_i$$

STEP 2: State the null and alternative hypotheses

The null hypothesis for an ANOVA always assumes the population means are equal. Hence, we may write the null hypothesis as:

$H_0 : \mu_1 = \mu_2 = \mu_3$ - (i.e. the state toll roads have equal number of employees).

And, the alternative hypothesis is:

$$H_1 : \mu_i \neq \mu_{i'},\ i \neq i'$$

i.e. the alternative hypothesis states that at least one mean is different from the others).

STEP 3: Find the Critical Value $F_{\alpha,df1,df2}$

In our example, df1 = 3 - 1 = 2 and df2 = 18 - 3 = 15. Hence we need to find $F_{0.05,2,15}$ using the F table. $F_{0.05,2,15}$ = 3.68.

The decision rule is therefore, to reject H_0 if $F_c > 3.68$ or do not reject it otherwise.

STEP 4: Calculate the appropriate test statistic

The test statistic in ANOVA is the ratio of the between and within mean treatment variation in the data. The SST, SSTR and SSE are computed using Equations (10.5) to (10.7), with p=3 and $n_i = 3$ for all *i*

Using

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n}G^2$$

$$SST = 2424 - \frac{1}{18}(23104)$$

$$\text{SST} = 2424 - 1283.556 = 1140.44$$

Between Sum of Squares (or Treatment Sum of Squares) – variation in the data between the different types of cars (or treatments).

Treatment Sum of Squares:

$$SSTR = \sum_{i=1}^{p} \frac{x_{i\bullet}^2}{n_i} - \frac{1}{n} G^2$$

$$SSTR = \frac{1}{6}\left(93^2 + 24^2 + 35^2\right) - \frac{1}{18}(23104)$$

$$\text{SSTR} = 1741.667 - 1283.556 = 458.11$$

Within variation (or Error Sum of Squares) – variation in the data from each individual treatment.

$$SSE = SST - SSTR$$

$$SSE = 1140.44 - 458.11 = 682.33$$

The next step in an ANOVA is to compute the "average" sources of variation in the data using SST, SSTR and SSE.

Mean Square Treatment: $MSTR = \frac{SSTR}{p-1}$, i.e., "average between variation" (with p = 3)

$$MSTR = \frac{458.11}{3-1} = 229.06$$

Mean Square Error: $MSE = \frac{SSE}{n-p}$, i.e., "average within variation"

$$MSE = \frac{682.33}{18-3} = \frac{682.33}{15} = 45.49$$

$$F_C = \frac{MSTR}{MSE} = \frac{229.06}{45.49} = 5.04$$

Table 10.2: ANOVA Table for One-way Classification

SV	DF	SS	MS	F
Blood Pressure Technique	3-1= 2	458.11	229.06	5.05
Error:	18-3=15	682.33	45.49	
Total:	18-1=17	1140.44	-	-

STEP 4: Make the decision

In our example 5.05 > 3.68, so we reject the null hypothesis and conclude that the the number of employees in the state toll roads are significantly different.

10.3 Two-way Classification

We now discuss the case when the data are cross classified by two factors A at p levels and B at q levels making pq sub-classes in all. When these factors impact upon an experimental

observation, a two factor ANOVA table can be used to test the significance of the effect of the factors. Two sets of hypothesis are of interest:

a. A test of the hypothesis that the means of all the columns are equal.

iii. A test of the hypothesis that the means of all the rows are equal.

Table 10.9: Layout for Two – way Classification

B / A	B_1	B_2	B_3	…..	B_j	…….	B_q	Total $X_{i.}$	Average $\bar{X}_{i.}$	Sum of Squares $\sum_{j=1}^{q} X_{ij}^2$
A_1	X_{11}	X_{12}	X_{13}	….	X_{1j}	….	X_{1q}	$X_{1.}$	$\bar{X}_{1.}$	$\sum_{j=1}^{q} X_{1j}^2$
A_2	X_{21}	X_{22}	X_{23}	…..	X_{2j}	…..	X_{2q}	$X_{2.}$	$\bar{X}_{2.}$	$\sum_{j=1}^{q} X_{2j}^2$
A_3	X_{31}	X_{32}	X_{33}	…..	X_{3j}	…..	X_{3q}	$X_{3.}$	$\bar{X}_{3.}$	$\sum_{j=1}^{q} X_{3j}^2$
. .	. .	. .	. .	…..	. .	…..	. .	. .	. .	. .
A_i	X_{i1}	X_{i2}	X_{i3}	…..	X_{ij}	…..	X_{iq}	$X_{i.}$	$\bar{X}_{i.}$	$\sum_{j=1}^{q} X_{ij}^2$
. .	. .	. .	. .	….	. .	….	. .	. .	. .	. .
A_p	X_{p1}	X_{p2}	X_{p3}	…..	X_{pj}	…..	X_{pq}	$X_{p.}$	$\bar{X}_{p.}$	$\sum_{j=1}^{q} X_{pj}^2$
Total $X_{.j}$	$X_{.1}$	$X_{.2}$	$X_{.3}$	…..	$X_{.j}$	…..	$X_{.q}$	$X_{..}$	——	
Average $\bar{X}_{.j}$	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}_{.3}$	…..	$\bar{X}_{.j}$	…..	$\bar{X}_{.q}$	——	$\bar{X}_{..}$	
Sum of Squares $\sum_{i=1}^{p} X_{ij}^2$	$\sum_{i=1}^{p} X_{i1}^2$	$\sum_{i=1}^{p} X_{i2}^2$	$\sum_{i=1}^{p} X_{i3}^2$	…..	$\sum_{i=1}^{p} X_{ij}^2$	…..	$\sum_{i=1}^{p} X_{iq}^2$			$\sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij}^2$

10.3.1 Model for Two-way ANOVA

The ANOVA model is defined as

$$X_{ij} = \mu + \alpha_i + \beta_j + e_{ij},\ i = 1, 2, ..., p,\quad j = 1, 2, ..., q \qquad (10.17)$$

where X_{ij} is the observation made in (i, j)th cell or plot; μ is the constant or universe effect; α_i is the mean effect of the ith level of a factor A; β_j is the mean effect of the jth level of factor B and ε_{ij} is the random error associated with the (i, j)th cell or plot.

Assumptions

(1) $e_{ij} \approx N(0, \sigma^2)$

(2) $\sum_{i=1}^{p} \alpha_i = \sum_{j=1}^{q} \beta_j = 0$

The hypotheses for the two – way classification are:

$H_0^{(1)}: \alpha_1 = \alpha_2 = \alpha_3 = ... = \alpha_p$

$H_1^{(1)}: \alpha_i \neq \alpha_j,\ i \neq j$

and

$H_0^{(2)}: \beta_1 = \beta_2 = \beta_3 = ... = \beta_q$

$H_1^{(2)}: \beta_i \neq \beta_j,\ i \neq j$

As shown in Table 10.9, if we denote the value of the dependent variable (response) when factor A is at ith level and factor B is at jth level as x_{ij}. Let $\bar{X}_{\cdot j}$ be the jth column mean and let $\bar{X}_{i\cdot}$ be the ith row mean. Let $\bar{X}_{\cdot\cdot}$ be the overall mean of the observations.

The total sum of squares (SST), the column sum of squares (SSC), the row sum of squares (SSR) and the error sum of squares (SSE) are given as:

$$\sum_{i=1}^{p}\sum_{j=1}^{q}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2$$

$$SST = SSE + SSR + SSC$$

Computational formulae

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij}^2 - \frac{G^2}{n} \tag{10.18}$$

where $G = \sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij},\ n = pq$

$$SSR = q\sum_{i=1}^{p} \bar{X}_{i\cdot}^2 - \frac{G^2}{n} \tag{10.19}$$

$$SSC = p\sum_{j=1}^{q} \bar{X}_{\cdot j}^2 - \frac{G^2}{n} \tag{10.20}$$

To compute the SSE, we use

$$SSE = SST - SSR - SSC \tag{10.21}$$

The mean square of columns (MSC), the mean square of rows (MSR) and the mean square error (MSE) are defined as:

$$MSC = \frac{SSC}{p-1} \tag{10.22}$$

$$MSR = \frac{SSR}{q-1} \tag{10.23}$$

$$MSE = \frac{SSE}{(p-1)(q-1)} \tag{10.24}$$

Under the null hypothesis that the mean of all columns are equal, the test statistic is given as $F_C = \frac{MSC}{MSE}$ where F_C is F-distributed with $df1 = (p-1)$ and $df2 = (p-1)(q-1)$. Under the null hypothesis that the mean of all rows are equal, the test statistic is given as $F_R = \frac{MSR}{MSE}$ where F_C is F-distributed with $df1 = (q-1)$ and $df2 = (p-1)(q-1)$. The ANOVA Table for a Two – way classification is displayed in Table 10.10.

Decision Rule:
Reject $H_0^{(1)}$ *when* $F_R > F_{\alpha,p-1,(p-1)(q-1)}$, where $F_{\alpha,p-1,(p-1)(q-1)}$ is the tabulated F- distribution value located at p -1 and (p -1)(q – 1) degrees of freedom.

Similarly,
Reject $H_0^{(2)}$ *when* $F_C > F_{\alpha,q-1,(p-1)(q-1)}$, where $F_{\alpha,q-1,(p-1)(q-1)}$ is the tabulated F- distribution value located at q -1 and (p -1)(q – 1) degrees of freedom.

Example 10.4: Suppose a farmer wishes to determine what the optimal amount of irrigation is for his fields, and also what the best type of fertilizer is for his fields. He has five brands of fertilizer to test and three levels of irrigation. By planting one acre of crops at each level of irrigation and with each fertilizer, he obtains the yields listed in Table 10.11.

Table 10.10: Analysis of variance for two-way classification with one member in each sub-class

Source variation	Degrees of freedom	Sum of squares	Mean squares	F - Ratio
Row	$p-1$	$SSR = q\sum_{i=1}^{p}\bar{X}_{i.}^2 - \frac{G^2}{n}$	$MSR = \frac{SSR}{q-1}$	$F_R = \frac{MSR}{MSE}$
Column	$q-1$	$SSC = p\sum_{j=1}^{q}\bar{X}_{.j}^2 - \frac{G^2}{n}$	$MSC = \frac{SSC}{p-1}$	$F_C = \frac{MSC}{MSE}$
Residual	$(p-1)(q-1)$	$SSE = SST - SSR - SSC$	$MSE = \frac{SSE}{(p-1)(q-1)}$	———
Total	$pq-1$	$SST = \sum_{i=1}^{p}\sum_{j=1}^{q}X_{ij}^2 - \frac{G^2}{n}$	———	

Table 10.11: Data on farm yields

	$B_1 = Low$	$B_2 = Moderate$	$B_3 = Heavy$	$X_{i\bullet}$	$\bar{X}_{i\bullet}$	$\sum_{j=1`}^{q} X_{ij}^2$
$A_1 = A$	33.7	37.2	35.6	106.5	35.5000	3786.89
$A_2 = B$	35.2	39.1	38.1	112.4	37.4667	4219.46
$A_3 = C$	30	32.1	35.1	97.2	32.4000	3162.42
$A_4 = D$	34.6	37.1	30.1	101.8	33.9333	3479.58
$A_5 = E$	29.5	32.8	35.1	97.4	32.4667	3178.1
$X_{\bullet j}$	163	178.3	174	515.30		
$\bar{X}_{\bullet j}$	32.60	35.66	34.80		34.3533	
$\sum_{i=1}^{p} X_{ij}^2$	5342.14	6395.31	6089.00			17826.45

Note: Each acre is classified according to two attributes or variables, namely: the type of fertilizer and the level of irrigation that were used.

Solution

$$X_{ij} = \mu + \alpha_i + \beta_j + e_{ij},\ i = 1, 2, ..., 5,\quad j = 1, 2, 3$$

$$H_0^{(1)}: \mu_A = \mu_A = \mu_A = \mu_D = \mu_E$$
$$H_1^{(1)}: \mu_j \neq \mu_{j'},\ j \neq j' \text{ (not all column means re equal)}$$

$$H_0^{(2)}: \mu_L = \mu_M = \mu_H$$
$$H_1^{(2)}: \mu_i \neq \mu_{i'},\ i \neq i' \text{ (not all row means are equal)}$$

$$G = \sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij} = 515.30,\quad \bar{X}_{\bullet\bullet} = 34.3533,\quad \sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij}^2 = 17826.45$$

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij}^2 - \frac{G^2}{pq}$$
$$= 17826.45 - \frac{(515.30)^2}{15}$$
$$= 17826.45 - 17702.27 = 124.18$$

$$SSR = q\sum_{i=1}^{p} \bar{X}_{i\bullet}^2 - \frac{G^2}{pq}$$
$$= 3\left[35.5^2 + 37.4667^2 + ... + 32.4667^2\right] - \frac{(515.30)^2}{15}$$
$$= 3(5919.3167) - 17702.27$$
$$= 55.68$$

$$MSR = \frac{SSR}{p-1} = \frac{55.68}{4} = 13.92$$

$$SSC = p\sum_{i=1}^{q} \bar{X}_{\bullet j}^2 - \frac{G^2}{pq}$$

$$= 5\left[32.60^2 + 35.66^2 + ... + 34.80^2\right] - \frac{(515.30)^2}{15}$$

$$= 5(3545.4356) - 17702.27$$

$$= 17727.18 - 17702.27$$

$$= 24.9082$$

$$MSC = \frac{SSC}{q-1} = \frac{24.9082}{2} = 12.45$$

$$SSE = SST - SSR - SSC$$

$$= 124.18 - 55.68 - 24.91$$

$$= 43.59$$

$$MSE = \frac{SSE}{(p-1)(q-1)} = \frac{43.59}{4 \times 2} = 5.45$$

ANOVA Table for Example 10.4

Source of variation	df	SS	MS	F-Ratio
Row	4	55.68	13.92	2.55
Column	2	24.91	12.45	2.28
Residual	8	43.59	5.45	————
Total	14	124.18	————	————

$F_{4,8,0.05} = 3.8379$

$F_{2,8,0.05} = 4.4590$

Conclusion

Row: We do not reject the null hypothesis and conclude that the five brands of fertilizers used are of equal effects since the $F_R = 2.55 < F_{4,8,0.05} = 3.8379$ at 0.05 level of significance.

Column: We do not reject the null hypothesis and conclude that the three levels of irrigation used are of equal effects since the $F_R = 2.28 < F_{2,8,0.05} = 4.4590$ at 0.05 level of significance.

EXERCISE TEN

Use Problem I to answer questions 1-6.

1. An educational psychologist interested in different methods for teaching introductory calculus wish to compare: Calculus taught by (1) a Seminar method, ((2) a tutorial, a question-and-answer method, or (3) a straight lecture method. To see whether the three methods differ from one another, he randomly select $n=27$ university year-one-students, and randomly assign $n_1=n_2=n_3=9$ of them to three experimental group. Each student then learns calculus by the method corresponding to his or her group. At the end of the course, all 27 students receive the same standardized final examination. The grades received by students in all three teaching methods are shown in the table below.

   Table 10.10: Data on teaching method

Seminar	94	90	95	89	88	92	92	97	91
Q & A	83	86	89	87	85	86	85	81	83
Lecture	80	85	81	81	79	83	78	80	82

   Calculate a measure of variability for the Seminar method

2. Calculate a measure of variability for the Q&A method
3. Calculate a measure of variability for the Lecture method
4. What is the sum of squares for the between teaching method?
5. What is the sum of squares for the within teaching method?
6. At $\alpha=0.05$, find the critical value for testing the homogeneity of teaching method means?
7. Which test is used to compare three or more means?
8. State three reasons why multiple t tests cannot be used to compare three or more means?
9. What are the assumptions in ANOVA for the one way classification?
10. What is the F test formula for comparing three or more means?
11. State the hypotheses used in the one way ANOVA test.

Use Problem II to answer questions 12-21.

Problem II: A researcher wishes to see whether there is any difference in the weight gains of athletes following one of these special diets. Athletes are randomly assigned to three groups and placed on the die for six weeks. The weight gains (I pounds) are shown here.

Diet A	Diet B	Diet C
3	10	8
6	12	3
7	11	2
4	14	5
	8	
	6	

12. Obtain the means of Diets A, B and C

13. Calculate the variance of Diet A.
14. Calculate the variance of Diet B.
15. Calculate the variance of Diet C.
16. Determine the number of degrees of freedom for between diets.
17. Determine the number of degrees of freedom for within diets.
18. Obtain the grand mean of diets.
19. Determine the test statistic value for testing that there is a difference in the diets?
20. Estimate the population variance of diets.
21. At $\alpha = 0.05$, can the researcher conclude that there is a difference in the diets?

Use Problem III to answer questions 22-25.

Problem III: Suppose that a random sample of n = 5 was selected from the vineyard properties for sale in Owerri, Imo state, in each of three years. The following data are consistent with summary information on price per acre for disease-resistant grape vineyards in Owerri. Carry out an ANOVA to determine whether there is evidence to support the claim that the mean price per acre for vineyard land in Owerri was not the same for each of the three years considered. Test at the 0.05 level and at the 0.01 level.

1996: 30000 34000 36000 38000 40000
1997: 30000 35000 37000 38000 40000
1998: 40000 41000 43000 44000 50000

22. State the hypotheses and identify the claim.
23. Calculate the sum of squares of the price per acre for vineyard land in Owerri for the three years.
24. Calculate the degrees of freedom for total variation.

Use Problem IV to answer questions 26-29.

Problem IV: A partially completed ANOVA table for a one-way cross classification is shown:

SOURCE	Df	SS	MS	F
Treatments	4	24.7	-	-
Error	-	-	-	
Total	34	62.4		

25. How many treatments are involved in the experiment?
26. What are the degrees of freedom for the error?
27. From the table, what is the sum of squares for error?
28. Determine the critical value for a difference between the population means.
29. Do the data provide enough evidence to indicate a difference between the population means?
30. What is the formula for computing the column sum of squares for a two-way ANOVA?

CHAPTER ELEVEN

CONTINGENCY TABLES AND CHI-SQUARE METHODS

11.0 INTRODUCTION

A contingency table, sometimes called a two-way frequency table, is a tabular mechanism with at least two rows and two columns used in statistics to present categorical data in terms of frequency counts. The intersection of a row and a column of a contingency table is called a cell. They are used in survey research, business intelligence, engineering and scientific research. They provide a basic picture of the interrelation between two variables and can help find interactions between them.

Our interest is in assessing whether one factor depends (or contingent) on another or that a set of factors influence another set of factors ,ie, the two categories are independent. When the factors are cross tabulated it is called contingency table and the analysis is called the contingency analysis.

Under this analysis, the n observations are classified in r rows of a factor (A) and c columns of another factor (B) so that the observed frequency in each ijth cell forms the matrix layout as illustrated in Table 11.1

Table 11.1: Matrix Layout of r x c Contingency Table.

A \ B	B_1	B_2	----	B_j	----	B_c	Total $n_{i.}$
A_1	n_{11}	n_{12}	----	n_{1j}	----	n_{1c}	$n_{1.}$
A_2	n_{21}	n_{22}	----	n_{2j}	----	n_{2c}	$n_{2.}$
:		:	:	:	:	:	:
:	:	:	:	:	:	:	:
A_i	n_{i1}	n_{i2}		n_{ij}	.....	n_{ic}	$n_{i.}$
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
A_r	n_{r1}	n_{r2}	-----	n_{rj}	----	n_{rc}	$n_{r.}$
Total $n_{.j}$	$n_{.1}$	$n_{.2}$	-----	$n_{.j}$	----	$n_{.c}$	n

$$n_{.j} = \sum_{i=1}^{r} n_{ij} \tag{11.1}$$

$$n_{i.} = \sum_{j=1}^{c} n_{ij} \tag{11.2}$$

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j} \tag{11.3}$$

Contingency table has a characteristics that each element or outcome belong to one and only one of mutually exclusive and exhaustive events in the row and one and only one of mutually exclusive and exhaustive events in the columns .For instance, if we classify

individuals by sex and age, any of the individuals randomly picked must be either a male or female belonging to one and only one of the age categories

Uses of Contingency Table

1 Assessment of joint and marginal probabilities.
2 Computation of expected frequencies.
3 Testing for independence of the two characteristics.

Example 11.1

Suppose there are two variables, sex (male or female) and handedness (right or left handed). Further suppose that 100 individuals are randomly sampled from a very large population as part of a study of sex differences in handedness. A contingency table can be created to display the numbers of individuals who are (male right handed and left handed),(female right handed and left handed). Such a contingency table is shown in Table 11.2.

Table11.2: Distribution of 100 individuals by sex and handedness.

SEX / HANDEDNESS	RIGHT HANDED	LEFT HANDED	TOTAL
MALE	43	9	52
FEMALE	44	4	48
TOTAL	87	13	100

The total number of males who are right-handed and left-handed individuals are 52. The total number of females who are right handed and left-handed individuals are 48. The total number of individual who are right-handed are 87 and left-handed individuals are 13. The marginal totals are $n_{1.} = 52,\ n_{2.} = 48, n_{.1} = 87,$ *and* $n_{.2} = 13$. The grand total (the total number of individuals represented in the contingency table) is the number in the bottom right corner ie $n = n_{1.} + n_{2.} = n_{.1} + n_{.2} = 100$

The example of Table 11.2 is the simplest kind of contingency table, a table in which each variable has only two levels; this is called a 2×2 contingency table. In principle, any number of rows and columns may be used. There may also be more than two variables, but higher order contingency tables are difficult to represent visually.

Example 11.2

The contingency Table of 11.3 has two rows and five columns (2 x 5 contingency table) showing the results of a random sample of 2200 adults classified by two variables, namely gender and favorite way to eat ice cream (Larson and Farber 2014). All computations done in Example 11.1 are shown in Table 11.3.

Table 11.3: Distribution of individual by gender and favorite way to eat ice cream .

PREFERENCE / GENDER	Cup	Cone	Sundae	Sandwich	Other	Total
Male	592	300	204	24	80	1200
Female	410	335	180	20	55	1000
Total	1002	635	384	44	135	2200

11.1 Joint and Marginal Probability Table.

Contingency tables are used for the computation of joint and marginal probabilities as illustrated in Table 11.4. One benefit of having data presented in a contingency table is that it allows one to easily perform basic probability calculations, especially the joint and marginal probabilities, a feat made easier still by augmenting a summary row and column to the table as illustrated in Table 11.4.

Table 11.4: Matrix Layout of r x c Joint and Marginal probability table.

A / B	B_1	B_2	----	B_j	----	B_c	Total
A_1	p_{11}	p_{12}	----	p_{1j}	----	p_{1c}	$p_{1.}$
A_2	p_{21}	p_{22}	----	p_{2j}	----	p_{2c}	$p_{2.}$
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
Ai	p_{i1}	p_{i2}	:	p_{ij}	:	p_{ic}	$p_{i.}$
:	:	:	:	:	:	:	:
A_r	p_{r1}	p_{r2}	-----	p_{rj}	----	p_{rc}	$p_{r.}$
Total	$p_{.1}$	$p_{.2}$	-----	$p_{.j}$	----	$p_{.c}$	1.00

$$p_{ij} = \frac{n_{ij}}{n} \tag{11.4}$$

$$p_{.j} = \sum_{i=1}^{r} p_{ij} \tag{11.5}$$

$$p_{i.} = \sum_{j=1}^{c} p_{ij} \tag{11.6}$$

NOTE:

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = \sum_{j=1}^{c} p_{.j} = \sum_{i=1}^{r} p_{i.} = 1.0 \tag{11.7}$$

Example 11.3

The joint and marginal probabilities of contingency Table 11.3 are given in Table 11.5. For example, the probability of male that preferred to eat their ice cream in cup is

$$p_{11} = \frac{592}{2200} = 0.27$$

The marginal probability of female members is

$$p_{2.} = \frac{1000}{2200} = 0.45$$

The marginal probability of those who eat ice cream as Sandwich is

$$p_{.4} = \frac{44}{2200} = 0.01$$

Table 11.5. Joint and Marginal Probability Distribution of individual by gender and favorite way to eat ice cream.

Preference / Gender	Cup	Cone	Sundae	Sandwich	Others	Total
Male	0.27	0.14	0.09	0.01	0.04	0.55
Female	0.19	0.15	0.08	0.00	0.03	0.45
Total	0.46	0.29	0.17	0.01	0.07	1.00

Example 11.4

The joint and marginal probabilities of contingency Table 11.2 are given in Table 11.6. For example, the probability of male that are left handed is $p_{12} = \frac{9}{100} = 0.09$

The marginal probability of male members is

$$p_{2.} = \frac{52}{100} = 0.52$$

The marginal probability of those who are right handed

$$p_{.4} = \frac{87}{100} = 0.87$$

Table 11.6. Joint and Marginal Probability Distribution of 100 individuals by sex and handedness

SEX / HANDEDNESS	RIGHT HANDED	LEFT HANDED	TOTAL
MALE	0.43	0.09	0.52
FEMALE	0.44	0.04	0.48
TOTAL	0.87	0.13	1.00

11.2 Expected Frequency

In Chi-square test of independence, the expected frequency for each cell is calculated and the assumption suggests that no cell should have expected frequency less than 5 but if any

expected counts are less than 5, then some other test should be used such as Fisher's exact test which is not covered in this text.

Once we have the observed counts we need to compute the expected counts under the null hypothesis that the two categorical variables are independent. This is done using the marginal totals and overall total to compute expected counts for each cell of the table. In words, to find the expected count for each cell in the table we multiply the marginal row and column totals for that cell and divide by the overall total. That is

The expected frequency value is given as

$$E_{ij} = \frac{sum\ of\ row i\ x\ sum\ of\ column\ j}{sample\ size} \tag{11.14}$$

$$E_{ij} = \frac{n_{i.}\ x\ n_{.j}}{n} \tag{11.15}$$

Assuming independence

$$p(A_i \cap B_j) = p(A_i)p(B_j) = \left(\frac{n_{i.}}{n}\right)\left(\frac{n_{.j}}{n}\right) = \frac{n_{i.}\ x n_{.j}}{n^2} \tag{11.16}$$

Therefore the expected number of the ijth cell is

$$E_{ij} = n\left(\frac{n_{i.}}{n}\right)\left(\frac{n_{.j}}{n}\right) = \frac{n_{i.}\ x\ n_{.j}}{n} \tag{11.17}$$

Example 11.5

From Table 11.3. Computing E_{11} says that the value one would expect at cell 11 ie., the expected number of male who prefer to eat ice cream from a cup--is approximately

$$E_{11} = \frac{1200\ x\ 1002}{2200} = 546.55$$

Computing E_{21} says that the value one would expect at cell 21 that is, the expected number of female who prefer to eat ice cream from a cup--is approximately

$$E_{11} = \frac{1000\ x\ 1002}{2200} = 455.45$$

Table 11.7: Distribution of expected frequency of individual by gender and favorite way to eat ice cream for Table 11.3

Preference / Gender	Cup	Cone	Sundae	Sandwich	Other	Total
Male	546.55	346.36	209.45	24	73.64	1200
Female	455.45	288.64	174.55	20	61.36	1000
Total	1002	635	254.45	44	135	2200

Example 11.6

Table 11.8: Distribution of the party affiliation and opinion for 500 surveyed individuals on Tax Reform.

	Favour	Indifferent	Opposed	Totals
Democrat	138	83	64	285
Republican	64	67	84	215
Totals	202	150	148	500

Table 11.8 represents the observed counts of the party affiliation and opinion for 500 individuals.. The question becomes, "How would this table look if the two variables were not related?" That is, under the null hypothesis that the two variables are independent, what would we expect to find in our data if the two variables (e.g. Party Affiliation and Opinion) were not related? We need to find what is called the **Expected Counts Table** or simply the **Expected Table.** This table displays what the counts would be for our sample data if there were no association between the variables.

To demonstrate, we will use the Party Affiliation and Opinion on Tax Reform.

Table 11.9. Calculating Expected Counts from Observed Counts

	Favor	**Indifferent**	**Opposed**	Total
Democrat	(285x202)/(500) = 115.14	(285x150)/(500) = 85.50	(285x148)/(500) = 84.36	285
Republican	(215x202)/(500) = 86.86	(215x150)/(500) = 64.50	(215x148)/(500) = 63.64	215
Total	202	150	148	500

One of the major benefits of computing expected frequencies is the ability to test whether the two variables being examined (in this case, gender and favorite way to eat ice cream) are actually independent as they have been assumed throughout. This is done by computing, for each cell, the expected frequency E_{ij} and comparing it with the observed frequency O_{ij}.

11.3 Chi-square Tests

Chi-square Tests can be conducted on contingency tables to test whether or not a relationship exists between variables. The chi square test statistic for contingency is given as

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \ldots + \frac{(O_k - E_k)^2}{E_k} \qquad (11.12)$$

$$\chi^2 = \sum_{i=k}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (11.13)$$

where "O_i" is the Observed value, "E_i" is the expected value, and the sum is taken over all cells in the contingency table ie (k = rc). When the **chi- square value is small it** means that

there is little relationship between the categorical variables. A **large chi square value** means that there is a definite relationship between the two variables.

11.4 Chi-square Test of Independence

How do we test the independence of two categorical variables? It will be done using the Chi-square test of independence. In this Chapter, we are interested in researching if two categorical variables are related or associated in the population from which the sample was drawn. Therefore, until we have evidence to suggest that they are dependent, we must assume that they are independent. This is the motivation behind the hypothesis for the Chi-square Test of Independence:

H_0 : In the population, the two categorical variables are independent.
H_1 : In the population, two categorical variables are dependent.

NOTE: The are several ways to phrase these hypotheses. Instead of using the words "independent" and "dependent" one could say "there is no relationship between the two categorical variables" versus "there is a relationship between the two categorical variables". The important part is that the null hypothesis refers to the two categorical variables not being related while the alternative is trying to show that they are related.

The statistical question becomes, "Are the observed counts so different from the expected counts that we can conclude a relationship between the two variables?". To conduct this test we compute a Chi-square test statistic (Equation 11.13) where we compare each cell's observed count to it's respective expected count.

As we have done with other statistical tests, we make our decision by either comparing the value of the test statistic to a critical value. The critical value for our Chi-square test is

$$\chi^2_{\alpha,(r-1)(c-1)}$$

Calculating the Chi-square test statistic for Table 11.8, we obtain

$$\chi^2 = \frac{(138-115.4)^2}{115.4} + \frac{(83-85.50)^2}{85.50} + \frac{(64-84.36)^2}{84.50} + \frac{(64-86.86)^2}{86.86} + \frac{(67-64.50)^2}{64.50} + \frac{(84-63.64)^2}{63.64}$$
$$= 22.152$$

with degrees for freedom equal to $\chi^2_{0.05,(2-1)(3-1)} = \chi^2_{0.05,2} = 5.99$ computed from the table.

Since $\chi^2 = 22.152 > \chi^2_{0.05,2} = 5.99$, we reject the null hypothesis and conclude that the party affiliations and opinions are different.

11.4.1 Steps to calculate the contingency table

1. State the hypotheses

2. State the assumptions

3. State the test statistic $\chi^2_{cal} = \sum_{i=k}^{k} \frac{(O_i - E_i)^2}{E_i}$

4. Set your decision rule ie reject Ho if $\chi^2{}_{cal} > \chi^2{}_{tab} = \chi^2_{\alpha,(r-1)(c-1)}$ otherwise do not reject.

5. Compute the test statistic

(i) Get the marginal totals

(ii) Get the expected frequency for each cell

(iii) Compute the Chi-square test statistic denoted by χ^2_{cal}

(iv) Obtain the tabulated Chi-square value $\chi^2_{tab} = \chi^2_{\alpha,(r-1)(c-1)}$

6. Take your decision using step three above.

EXERCISE ELEVEN

1. To determine whether the age of a driver aged 21 years or older has any effect on the number of motor accidents he is involved in a survey was conducted and following information obtained. Test the hypothesis that the number of accidents is independent of the age of the driver at 5%

Number of accidents	Age of Driver				
	21-30	31-40	41-50	51-60	61-70
0	148	221	186	120	72
1	44	30	21	36	20
2	19	13	10	4	3
More than 2	4	5	2	1	3

2. The relationship between cigarette smoking and development of lung cancer was investigated and the following results were obtained. Is smoking responsible for development of lung cancer? take α =5%

	Developed lung cancer	No lung cancer
Smokers	180	7
Non-smokers	3	104

3. To test effectiveness of a drug, the following result was obtained from an experiment at 5%.

	Recover	Did not recover
Used drug	75	25
No drug	65	35

4. A survey to investigate the relationship between exposure to unprotected sex and HIV infection and the information below was obtained.

	HIV Infection	NO HIV
Unprotected sex	800	10
Protected sex	5	720

Is HIV infection independent of exposure to sex? take α at 1%.

5. In an experiment on the immunization of goats from anthrax, the following results were obtained .Derive your inference on the efficacy of the vaccine.

	Died of Anthrax	Survived
Inoculated with vaccine	2	10
Not inoculated	6	6

6. Suppose that in a public opinion survey answers to the questions

(a) Do you drink ? (b) Are you in favour of local option on sale of liquor? were given in a table below

	Question (a)	
Question (b)	Yes	NO
Yes	56	31
NO	18	6

Can you infer that opinion on local option is dependent on whether or not an individual drinks at 1%?

7. In a certain sample of 2000 families,1400 families consume tea. Out of 1800 Imolites families,1236 families consume tea. Test if there is any significant difference between consumption of tea among Imolites and non Imolite families
8. Test at 5% if the serum has no effect on the recovery?

	Recover	Do not Recover
Group A(Using Serum)	75	25
GroupB(NotUsing Serum)	65	35

9. From the table below the results of an investigation of the effect of vaccination of laboratory animals against a particular disease. Using 0.01 sig. level test the hypothesis that there is no difference between the vaccinated and unvaccinated group.

	Got disease	Do not get disease
Vaccinated	9	42
Non Vaccinated	17	28

10. Using the table above that shows the number of student in each of two classes. A and B, who passed and failure an examination given to both groups. Using the 0.05 significance level. Test the hypothesis that there is no difference between the two classes.

	Passed	**Failed**
Class A	72	17
Class B	64	23

11. Test the hypothesis that there is no difference between sleeping pill and sugar pills at a significance level of 0.05.

	Slept well	Did not sleep well
Took sleeping pills	44	10
Took sugar pills	81	35

12. In a survey of 200 boys of which 75 were intelligent.40 had educated fathers while 85 of the unintelligent boys had uneducated fathers. Do these figure support that hypothesis that educated fathers have intelligent boys at 1% .
13. Test if the hypothesis the opinion bout autonomous college is independent of the level of classes. at 5%

	Favouring	Opposing
Undergraduate	290	100
Post graduate	310	90

14. Two researchers adopted different methods sampling techniques while investigating the same group of students to find the number of students falling in different intelligence level. At 5% would you say that the sampling techniques adopted by two researchers are significantly different.

	Below average	average	Above average	Genius
X	86	60	44	10
Y	40	33	69	12

15.The following table gives for a sample of married women,the level of education and the degree of adjustment in marriage score

		Marriage	adjustment	score	
		Very low	low	high	Very high
Level of	College	24	97	62	58
education	High school	22	28	30	41
	Middle school	32	10	11	20

16. The following **Contingency Table** shows the number of Females and Males who each have a given **eye color**. Note that, for example, the table show that 20 Females have Black eyes and that 10 Males have Gray eyes. Also notice that the Table shows the totals. We have 85 Females in the dataset. We have 82 Males in the dataset. We have a total of 167 People in the dataset. Finally, this table also shows the totals for **eye color**. For example, 45 People have Black eyes

EYE COLOR	Black	Brown	Blue	Green	Gray	Total
Female	20	30	10	15	10	85
Male	25	15	12	20	10	82
Total	45	45	22	35	20	167

17. A speech pathologist might want to know whether the proportion of males among stammerers and the proportion of males among lispers are the same. Test if the claim is true at 5% level of significant.

	Stammer	Lisp
Male	32	28
Female	18	22

18. A coach examines the records at his school over 31 years. And he wants to know whether a game is won or lost is independent of whether the game is played at home or away. At 1% level of sign.

	Home	Away
Won	97	69
Lost	42	83

19. A serum thought to be effective in preventing cold is given to 300 persons. Their records for one year are compared with those of 200 untreated persons with the following results:

	No cold	One cold	More than one cold
Treated	145	80	75
Untreated	80	70	50

20. It is reported that offspring of users of a certain recreational drug may have a higher incidence of birth defects than the general population .To obtain information about a possible relationship between this drug and birth defects,100 offspring of female rats fed the drug and 100 offspring from untreated female rats are examined. The results are given below analyzed at 1%

	Progeny	
female	Birth defects	Normal
Treated	30	70
Untreated	20	80

21.Test the null hypothesis that there is no difference between the quality of the two diets

	Excellent	Average	Poor
Diet A	37	24	19
Diet B	17	33	20

22. Construct a test to determine if the incidence rate of alcoholism appears to be the same in all four groups at 5% level of significant

	Alcoholic	Non-alcoholic
Clergy	32	268
Educators	51	199
Executives	67	233
Merchants	83	267

23. Analyze the data concerning if credit fraud and release of personal information is harmful or not at 1% level of significant

GOVT		FINANCIAL	OTHER COMMERCIAL
HARM	30	24	21
NO HARM	182	104	34

24. A study of college experience of students who were the first generation in their family to attend college. These students were compared to other students who were not the first generation in their family to go to college. Whether or not the students dropped out during their first semester was measured and represented in the table below. Test whether or not

students drop out of college is independent of whether or not they are first generation in their family to go to college.

	Generation to go to college	
	First	Other
Dropped Out	73	89
Did Not Drop Out	657	1226

25. The data below shows strong gender differences in teenagers' approaches to deal.ing with mental health issues. Do the data show a significant relationship between gender and willingness to seek mental health assistance at 5%?

	Willingness to Use	Mental Health	services
	Probably No	Maybe	Probably Yes
Males	17	32	11
Females	13	43	34

26. A researcher would like to know which factors are most important to people who are buying a new car. A sample of n =200 customers between the ages of 20 and 29 are asked to identify the most important factor in the decision process: Performance, Reliability or style. The researcher would like to know whether there is a difference between the factors identified by women compare to those identified by men at 1%.

	Performance	Reliability	Style
Male	21	33	26
Female	19	67	34

27. From the table below the results of an investigation of the effect of vaccination of laboratory animals against a particular Eye problem disease. Using 0.01 sig. level test the hypothesis that there is no difference between the vaccinated and unvaccinated group.

	EYE PROBLEM	**NO EYE PROBLEM**
Vaccinated	9	42
Non Vaccinated	17	28

28. Test at 5% if the serum has no effect on the number alive?

	Number Alive	**Number Death**
Group 1 (Using Serum)	75	25
Group2 (NotUsing Serum)	65	35

29. The test conducted in order to ascertain whether a given set of observation is drawn from a specified probability distribution is called ----------
    A. Chi-square test
    B. Chi-square test of independence.
    C. Chi-square test of goodness of fit.
    D. Chi-square test of probability distribution.

30. Out of 150 computer systems that crashed ,5% were due to hard ware failure. Calculate the expected number of computer systems that crashed because of hardware failure.
    A 0.05 B. 7.5 C.15 D.750

CHAPTER TWELVE

RATIOS, RATES AND INDEX NUMBERS

12.0 INTRODUCTION

Ratios and Rates are mathematical concepts describing the Numerator – Denominator relationship between two numbers. In Statistics, Ratios are used to discuss the relative values of the frequencies (a) two categories of a characteristic of an observation unit (eg male and female sexes, Broad classes of Age of a population etc), (b) two characteristics of an observation unit (eg population and Land Area; GDP; Electricity Supply; water supply and motorable roads in a country, Patients and Doctors; Nurses and other paramedical personnel in a Hospital, Students/Pupils and teachers; and Classrooms in a school etc).

Rates, on the other hand, are used in Statistics to discuss frequency of events in a specified time interval in relation to the population exposed to the risk of the event. The incidence of such events as success or failure in an examination, births, deaths, migration, Marriage, Divorce, unemployment, economic activity, disability, occurrence of diseases etc are usually measured by comparing the frequencies of such events with the population at risk of the events. When the exposed population is the population at the midpoint of the interval, then the rate is referred to as central rate. If the exposed population is the population at the beginning of the interval, the rate is referred to as probability. In what follows we discuss the ratios and rates commonly used to discuss statistical data.

Ratios and rates provide better indices for comparison of two categories of a characteristic or two characteristics of an observation unit when the populations are not equal.

Furthermore, an index number is a specialized average designed to measure the change in a group of related variables over a period of time. It is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, or other characteristics such as income, profession etc. If we wish to compare several series of figures, it is more than likely that their complexity will render direct comparison meaningless. Index numbers, in effect, relates a variable or variables in a given period to the same variable or variables in another period called the base period.

Index numbers measure the effect of changes over a period of time. Thus, index numbers are most widely used for measuring changes over a period of time.

An index, the simplified name for index numbers, which is computed from a single variable is called a univariate index while an index which is constructed from a group of variables is considered a composite index. Index numbers are indispensable tools of economic and business analysis. They help in framing suitable policies. Many of the economic and business policies are guided by index numbers.

Index numbers are highly useful in deflating, that is, they are used to adjust the original data for price changes or to adjust wages for cost of living changes and thus transform nominal wages into real wages.

Trends and tendencies are revealed by index numbers. The time series so formed enable us to study the general trend of the phenomenon understudy. Hence, they are important in forecasting future economic activity.

Index numbers can be used to compare food or other living costs in a city during one year with those of a previous year or we can compare cocoa production during a given year in one part of a country with that in another part.

It can be applied in education. For instance, it can be used to compare the relative intelligence of students in different locations or for different years.

Many private agencies and government are engaged in computing indexes (index numbers) for purposes of forecasting business and economic conditions providing general information. We can have wage indexes, unemployment indexes, production indexes etc. The best known is the cost of living index or consumption price index.

12.1 Ratios

Ratios commonly encountered in statistics include;

(a) **Population Density** defined as ratio of population of a well defined territory at a specified time to the total land Area of the territory, ie

$$\text{Density} = \frac{\text{Population}}{\text{Land Area}} \tag{12.1}$$

Example: The population of Nigeria was 140,431790 in 2006 while the total land area is 923768square kilometers. So the density is

$$\text{Density} = \frac{140431790}{923768} = 152.02 \text{ persons per km}^2$$

This indicates that in the 2006 Nigeria census about 152 persons were reported to living in area of one km^2.

(b) **Per Capita Income** (PCI) is defined for a country as the ratio of Gross Domestic Product (GDP) to the total population of the country, ie

$$\text{PCI} = \frac{\text{GDP}}{\text{Total Population}} \tag{12.2}$$

Example: If the GDP of Nigeria with a population of 177101301 in 2014 was ₦ $6.716\text{x}10^{10}$, the Per Capita Income is

$$\text{PCI} = \frac{67160000000}{1771101301} = \text{N } 379.22 \text{ per person}$$

Per Capita Income (PCI) of a country is measure of level of economy of the country.

Table 12. 1: GDP (x ₦10^{10}) and total population of a country

SN	Activity Sector	Year				
		2010	2011	2012	2013	2014
1	Agriculture	13.05	13.43	14.33	14.75	15.38
2	Industry	12.03	12.87	13.03	13.02	13.79
3	Construction	1.57	1.82	1.99	2.27	2.57
4	Trade	8.99	9.64	9.85	10.51	11.13
5	Services	18.97	19.75	20.73	22.67	24.29
	GDP					
	Population	**157704321**	**162344706**	**167121633**	**172039120**	**177101301**

(c) **Students or Pupils – Teacher Ratio**, used in assessing the adequacy of number of teachers in a school system, is defined as the ratio of number of students (or pupils) to Teachers available to them.

Example; The number of teachers in a state in 2008 is 2670 while the corresponding students enrolment is 134208. The student- Teacher ratio is 134208 / 2670 = 50.27

This implies that out of every one person reported as a teacher more than 50 persons were reported as students.

(d) **Student or pupils – Classroom Ratio,** defined as the ratio of Students (or Pupils) to the number of classes available to them is a measure of adequacy of educational facilities in a school system. In Hospitals we use (e) **Patients** – **Doctors** Ratio; (f) **Patients** – **Nurses** Ratio and (g) **Patients** – **Hospital bed** ratio to assess the adequacy of hospital facilities in a given environment.

Other ratios commonly in use are those required for assessment of quality of Demographic data, dependency burden and quality of life. Demographic indicators include

(h) **Sex Ratio (SR)** of a population, defined as the ratio of number of males (M) to the number of females (F) in an enumeration, ie

$$SR = \frac{M}{F} \times 100 \qquad (12.3)$$

Example: From the 2006 Nigeria census, the number reported as males is 71345488 while the number reported as females is 69086302. Therefore, the sex ratio of the population is

$SR = \frac{71345488}{69086302} \times 100 = 103.27$ That is out of every 100 persons enumerated as females in the 2006 Nigeria census; about 103 were reported as males. Sex ratios can also be defined for the total number reported as dead (Sex Ratio at Death SRD) and the number reported as births (Sex Ratio at Birth SRB), specific sex ratios. Just like the population sex ratio, sex ratio at death and sex ratio at Birth are used to evaluate the quality of data on death and birth records from vital registration, censuses and sample surveys.

(i) **Dependency Ratio (DR).** For a country, the dependency ratio is the ratio of total number of persons reported as aged either under 15years or at least 65 years to population aged 15 to 64 years

$$DR = \frac{\text{popn under 15 yrs} + \text{popn 65 yrs and above}}{\text{population aged 15 to 64 years}} \times 100 \qquad (12.4)$$

Example: From the census records of a country, the total number of persons reported as aged under 15 years is 9941, as aged 15 to 64 years is 33248 and at least 65 years is 8222. The corresponding Dependency ratio is given by

$$DR = \frac{9941+8222}{33248} \times 100 \quad = 54.63\%$$

That is, out of every 100 persons reported as aged 15 to 64 years, about 55 (54.63) were reported as either under 15 years or 65 years and above. Dependency ratio is used as a measure of dependency burden in a country. This because most of those under 15 years are either of pre-school age or in school while most of the people aged at least 65 years are either aged or retirees and are assumed not to be working and/ or not earning income. They depend on the population aged 15 to 64 years who are assumed to be working and earning income for their sustenance.

Table 12. 2: Distribution of populations of four countries in the same year by broad age groups

Age Group	Country			
	A	B	C	D
0 – 14	21901	13084	22855	870
15 – 64	23096	11346	86127	3457
65 +	2031	700	14475	193

12.2 Rates

Rates commonly used to discuss demographic and socio-economic indicators in statistics and to assess the success or otherwise of the sustainable development goals include;

(i) **Birth Rate**: This is the ratio of total number of births (B) in country in a year to the population exposed to the risk of birth. When the population assumed exposed to the risk of birth is the midyear population (P) of the country, irrespective of differences in the degree of exposure to the risk of childbearing, then the birth rate is referred to as Crude Birth Rate (CBR).

$$CBR = \frac{B}{P} \times 1000 \qquad (12.5)$$

(ii) **Death Rate:** Death rate is the ratio of the total number of deaths (D) in country in a year to the population exposed to the risk of dying. When the population exposed to the risk of dying is the midyear population (P) of the country, the rate is referred to as Crude Death Rate (CDR)

$$CDR = \frac{D}{P} \times 1000 \qquad (12.6)$$

Examples:
The total number of registered births in a country in a year is 432443, while the total number of registered death is 94840. If the midyear population of the country is 21169051, then the

$$\text{CBR} = \frac{432443}{21169051} \text{ x } 1000 \quad = 20.43\%0$$

and the

$$\text{CDR} = \frac{94840}{21169051} \text{ x } 1000 \quad = 4.48\%0$$

Thus, out of every 1000 persons reported in the midyear population, about 20 (20.43) were reported as births and about 4 (4.48) were reported as dead. Crude birth rate and crude death rate are used to assess the impacts of births and deaths on population.

Table 12. 3: Distn of total popn, number dead and number of Births in a country by sex

Age Group	Population		Death		Birth	
	Male	Female	Male	Female	Male	Female
Total	10823484	10345567	54920	39920	222937	209506
Both	21169051		94840		432443	

(iii) **Marriage Rate:** This is the ratio of number marriages in a country in a year to the midyear population exposed to the risk of marriage. In most countries marriage is restricted to population aged 10 or 15 years and above. However, if the total midyear population is used, the rate is referred to as Crude Marriage Rate (CMR)

$$\text{`CMR} = \frac{\text{Popn of Ever Married}}{\text{Total popn 10 yrs and above}} \text{ x } 1000 \qquad (12.7)$$

(iv) **Divorce Rate:** This is the ratio of number divorces in a country in a year to the midyear population exposed to the risk of divorce. The population exposed to the risk of divorce is the population of ever-married persons in the country.

$$\text{CRD} = \frac{\text{Popn of Divorced}}{\text{Total popn Ever married}} \text{ x } 1000 \qquad (12.8)$$

Examples: From Table 12 .4,

Crude Marriage Rate (CMR) is given by

$$\text{CMR} = \frac{52393770}{97831443} \text{ x } 100 = 53.56\%$$

This indicates that out of every 100 persons reported as aged 10 years and above, about 54 (53.56) were reported as ever-married. The Crude rate of Divorce (CRD) is given as

$$\text{CRD} = \frac{703614}{52393770} \text{ x } 100 = 1.34\%$$

This indicates that out of every 100 persons reported as ever-married, about one (1.34) was reported as divorced.

Table 12.4: Distn of popn aged 10 years and above by marital status in the 2006 Nig Census

Age Group	Population		Ever married		Divorced	
	Male	Female	Male	Female	Male	Female
Total	49387659	48443784	23417936	28975834	229627	473987
Both	97831443		52393770		703614	

12.3 Index Numbers

12.3.1 Methods of Constructing Index Numbers

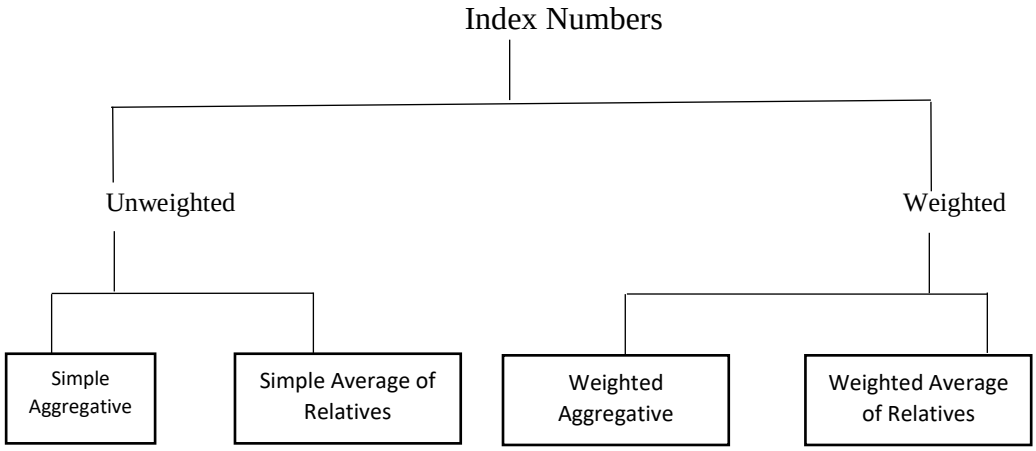
A large number of formulae have been devised for constructing index numbers. They can be grouped under two major heads namely

a) Unweighted indices and
b) Weighted indices

In the unweighted indices, weights are not expressly assigned whereas in the weighted indices, weights are assigned to the various items. Each of these types may be further divided into two heads:

i) Simple aggregative and
ii) Simple average of relatives

The chart below illustrates the various methods



12.3.2 Unweighted Index Numbers

i) Simple Aggregative Method.

This is the simplest method of constructing index numbers. When this method is used to construct a price index, the total of current year prices for the various commodities in question is divided by the total base year prices and the quotient is multiplied by 100. It should be noted that the base year period of an index number (also called the reference

period) is the period against which comparisons are made. It may be a year, a month or a day. The index for base period is always taken as 100.

Hence, the Simple Aggregative Price index (P_{01}) is given by

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100 \tag{12.9}$$

where $\sum p_0$ = sum of all commodity price in the base year

$\sum p_1$ = sum of corresponding commodity prices in the current year.

ii) Simple Average of Price Relative Method

When we use this method to construct a price index, we first of all obtain price relatives for the various items included in the index and the average of those relatives is obtained using any one of the measures of central value such as arithmetic mean, median, mode, geometric or harmonic mean.

Using the arithmetic mean, the simple arithmeticmean of relative price index P_{01} is given by

$$P_{01} = \frac{\sum\left(\frac{p_1}{p_0} \times 100\right)}{N} \tag{12.10}$$

where N = number of items (commodities) whose price relatives are averaged

p_1 = commodity price during the current year

p_0 = commodity price during the base year

Also, using the Geometric mean, the formula for obtaining the index becomes

$$\log(P_{01}) = \frac{\sum \log\left(\frac{p_1}{p_0} \times 100\right)}{N} \tag{12.11}$$

or

$$\log(P_{01}) = \frac{\sum \log(k)}{N} \tag{12.12}$$

where $k = \frac{p_1}{p_0} \times 100$

or

$$P_{01} = anti\log\left[\frac{\left(\sum \log\left(\frac{p_1}{p_0} \times 100\right)\right)}{N}\right] = anti\log\frac{\sum \log(k)}{N} \tag{12.13}$$

Merits

i) Extreme items do not influence the index; equal importance is given to all the items.

ii) The index is not influenced by the units in which prices are quoted or by the absolute level of individual prices.

Limitations

i) The relatives are assumed to have equal importance
ii) Difficulty is faced with regards to the selection of an appropriate average. It should be noted that the use of the arithmetic mean is considered as questionable sometimes because it has an upward bias. The use of geometric mean involves difficulties of computation. Other averages are almost never used while constructing index numbers.

12.3.3 Weighted Index Numbers

In this case, we weigh the price of each commodity by a suitable factor. Weighted index numbers are of two types namely

i) Weighted Aggregative Indices and
ii) Weighted Average of Relatives

12.3.3.1 Weighted Aggregative Indices

Here, the indices are of the simple aggregative type with the main difference that weights are assigned to the various items included in various methods of assigning weights.

We use one of the following formulae:

a) Laspeyres Method (The base year method)

The Laspeyres price index is a weighted aggregate price index where the weights are determined quantities in the base period. Let p_0 and p_n represent the prices of the item purchased by a consumer in the base and current years respectively. Let q_0 and q_n respectively represent the quantities of the items purchased by a consumer in the base year and the current year.

$$\text{The Laspeyres'index} = \frac{\sum(p_n q_0)}{\sum(p_0 q_0)} \times 100 \qquad (12.14)$$

Merits

i) It takes into account the relative importance of each of the items under consideration
ii) It makes for easy year – to- year (period-to-period) comparison of changes in prices

Demerits

i) It does not take into consideration the consumption pattern, that is, it assumes constant consumption.
ii) The different units of measurements affect the value of the index.

b) Paasche Method.

The Paasche Price Index is a weighted aggregate price index in which the weights are determined by quantities in the given year.

$$\text{The Paasche index} = \frac{\sum(p_n q_n)}{\sum(p_0 q_n)} \times 100 \qquad (12.15)$$

Merits

i) It takes into account the relative importance of each item.
ii) It uses the current quantity of various items.

Demerits

i) Measurement units are not uniform
ii) It cannot make for easy year-to-year comparison in the changes in prices.

An important feature of Laspeyres and Paasche indices is that the former is expected to overestimate (shows an upward bias) whereas the latter underestimates (shows a downward bias).

c) Dorbish and Bowley's Method

Dorbish and Bowley suggested simple arithmetic mean of the Laspeyres and Paasche indices. The formula is given by

$$\text{Dorbish and BowleyIndex} = \frac{\frac{\sum p_n q_0}{\sum p_0 q_0} + \frac{\sum p_n q_n}{\sum p_0 q_n}}{2} \times 100 = \frac{L+P}{2} \qquad (12.16)$$

where L = Laspeyres index
P = Paasche index

d) Fisher's Ideal Method

This index is the geometric mean of Laspeyres and Paasche indices. The Fisher's ideal index is given by

$$\text{Fisher's Ideal Index} = \sqrt{\left(\frac{\sum p_n q_0}{\sum p_0 q_0}\right)\left(\frac{\sum p_n q_n}{\sum p_0 q_n}\right)} \times 100 = \sqrt{(L \times P)} \times 100 \qquad (12.17)$$

The index satisfies both time-reversal and factor-reversal tests which gives it certain theoretical advantages over the other index numbers. It should be noted that the index is free from bias. It is not however a practical index to compute because it is excessively laborious.

e) The Marshall-Edgeworth Method

This method considers both the current year as well as the base year prices and quantities.

$$\text{The Marshall-Edgeworth index} = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100 \qquad (12.18)$$

It is a simple readily constructed measure, giving a very close approximation to the results obtained by the ideal formula.

f) Kelly's Method (Fixed Weight Aggregative Index)

The index is given by

$$\text{Kelly's Index} = \frac{\sum p_n q}{\sum p_0 q} \times 100 \qquad (12.19)$$

where $q = \frac{q_0 + q_1}{2}$

Similarly, the average of the quantities of 3 or more years can be used as weights.

g) Walsch Price Index

Instead of taking the arithmetic mean of the base year and the current year quantities as weights, we take their geometric mean, then we obtain the index given by

$$\text{WalschIndex} = \frac{\sum p_n \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100 \qquad (12.20)$$

12.3.3.2 Weighted Average of (Price) Relatives

In this method, we weigh each (price) relatives by the total value of the commodity in terms of some monetary units, such as Naira and Dollar. Since the value of a commodity is obtained by multiplying the price p of the commodity by the quantity q, the weights are given by pq.

Depending on whether we are using the base year, given year or typical year values denoted by $\left(p_0 q_0, p_n q_n, \quad and \quad p_1 q_1\right)$ respectively.

We use one of the following formulae

(i) Weighted Arithmetic Mean of Price Relatives Using Base Value Weights.

$$\frac{\sum\left(\frac{p_n}{p_0}\right)(p_0 q_0)}{\sum p_0 q_0} = \frac{\sum p_n q_0}{\sum p_0 q_0} = \text{Laspeyres formula} \qquad (12.21)$$

(ii) Weighted Arithmetic Mean of Price Relatives Using Given Year Value Weights.

$$\frac{\sum\left(\frac{p_n}{p_0}\right)(p_0 q_0)}{\sum p_n q_n} \qquad (12.22)$$

(iii) Weighted Arithmetic Mean of Price Relatives Using Typical Year Value Weights

$$\frac{\sum\left(\frac{p_n}{p_0}\right)(p_1 q_1)}{\sum p_1 q_1} \qquad (12.23)$$

12.4 Quantity or Volume Index Numbers

Quantity index numbers measure the physical volume of production, employment or construction. Though price indices are widely used, production indices are highly significant as indicators of the level of output in the economy or in parts of it.

The formulae for obtaining price index numbers are easily modified to obtain quantity (or volume) index numbers by simply interchanging p and q. Thus, when Laspeyres method is used, the index say Q_{01} is given by

$$Q_{01} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100 \qquad (12.24)$$

Using Paasche method

$$Q_{01} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100 \qquad (12.24)$$

Using Fisher's Formula

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100 \qquad (12.25)$$

These formulae represent the quantity index in which quantities of the different commodities are weighted by the prices.

Value Index numbers

The value of a single commodity is the product of its price and quantity. Let v be the value index which is the sum of the values of a given year divided by the sum of the values of the base year.

Value index = V = $\frac{\sum p_n q_n}{\sum p_0 q_0} \times 100$ (12.26)

where $\sum p_0 q_0$ = total value of all commodities in the base year

$\sum p_n q_n$ = total value of all commodities in the given period

This is a simple aggregate index, since the values have not been weighted.

The formula above can be stated more simply as

$$V = \frac{\sum V_1}{\sum V_0} \qquad (12.27)$$

In this type of index, both price and quantity are variable in the numerator. It should be noted that weights do not have to be applied, since they are inherent in the value figures.

Therefore, a value index is an aggregate of values. It measures the change in actual values between the base and the given period.

12.5 Test of Consistency of Index Number Formulae

Various formulae have been suggested for constructing index numbers. None of the formulae measures the price changes or quantity changes with perfection and has some bias.

The problem is to choose the most appropriate formulae in a given situation. The following tests are suggested for choosing an appropriate index.

a) Unit Test

This test requires that the formula for constructing an index should be independent of the units in which or for which prices and quantities are quoted.

b) Time Reversal Test (TRT)

This test requires the index number formula to possess time consistency by working both forward and backward with respect to time. In other words, time reversal test is a test to determine whether a given method will work both ways in time, forward and backward. Hence, when the data for any two years are treated by the same method, but with the bases reversed the two index numbers secured should be reciprocals of each other. So that their product is unity ie

$$P_{01} \times P_{10} = 1 \qquad (12.27)$$

where

P_{01} = the index for time "1" on time "0" as base and

P_{10} = the index for time "0" on time "1" as base.

If the product is not unity, there is said to be time bias in the method. This test is not satisfied by Laspeyres and Paasche method.

c) Factor Reversal Test (FRT)

The test holds that the product of price index and the quantity index should be equal to the corresponding value index. In other words, the test entails that the change in price multiplied by the change in quantity should be equal to the total change in value. However, the test is satisfied only by the Fisher's Ideal Index

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \qquad (12.28)$$

d) Circular Test

The test is an extension of the Time Reversal Test. Westergaard suggested that the circular test is an extension of TRT for more than two periods and is based on the shiftability of the base period. This requires the index to work in a circular manner and this property enables us to find the index numbers from period to period without referring back to the original base each time. Let a,b and c be 3 periods, the test requires that if an index is constructed for the year 'a' on base year 'b' and the year 'b' on base year 'c', we ought to get the same result as if we calculated directly an index for 'a' on base year 'c' without going through 'b' as an intermediary ie

$P_{ab} \times P_{bc} \times P_{ca} = 1, \quad a \neq b \neq c$

where P_{ij} = the price index (without factor 100) for period j with period i as base.

For instance, using Laspeyres' method, the index will be

$$\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_2 q_1}{\sum p_1 q_1} \times \frac{\sum p_2 q_1}{\sum p_2 q_2} \neq 1$$

Hence, Laspeyeres index does not satisfy the circular test. It can be shown that none of Paasche, Marshal Edgeworth's, Walsch's and Fisher's indices satisfies this test. In fact, circular test is not satisfied by any of the weighted aggregative formulae with changing weights.

This test is satisfied only by the index number formulae based on

a) Simple geometric mean of the price relatives and
b) Kelly's fixed base method.

e) **Chain Base Index numbers (Chain Indices)**

The chain base method consists in computing a series of index numbers (by a suitable method) for each year with the preceding year as the base year. In its simplest form, the chain index is one in which the figures for each year (or sub-period) are first expressed as percentages are then chained together by successive multiplication to form a chain index.

For instance, if P_{ab} denotes the price index for current period b with respect to the base year a, then we compute series of indices $P_{01}, P_{12}, P_{23}, \ldots\ldots\ldots P_{r-1,r}$, if we are given the data for (r+1) periods.

These indices are called Link Index numbers or Link Relatives. The basic chain indices are obtained from these link relatives by successive multiplication as given below:

P_{01} = First Link

$P_{02} = P_{01} \times P_{12}$

$P_{03} = (P_{01} \times P_{12}) \times P_{23} = P_{02} \times P_{23}$

. .

. . .

$P_{or} = P_{0,r-1} \times P_{r-1,r}$

12.6 Steps in Constructing a Chain Index

i) Express the figures for each year as percentages of the preceding year. This gives the Link relatives (LR). Hence

$$\text{LR} = \frac{\text{current year's price}}{\text{Previous year's price}} \times 100 \qquad (12.29)$$

ii) Chain together these percentages by successive multiplication to form a chain index. Chain index of any year is the average link relatives of that year multiplied by chain index of previous year divided by 100.

Conversion of Chain index to Fixed Index

To convert the Chain Base Index (CBI) numbers into Fixed Base Index (FBI), we follow the following procedure:

(i) For the first year, the FBI will be taken the same as the CBI. If the index numbers are to be constructed by taking first year as the base, in that case, the index for the first year is taken as 100.

(ii) To calculate the indices for other years, we use the following formula

$$\text{Current year's FBI} = \frac{(Current\ year's\ CBI)*(Previous\ year's\ FBI)}{100} \quad (12.30)$$

12.7 Base Shifting

Most times, it frequently becomes necessary to change the reference base of an index number series from one time period to another without returning to the original raw data and re-computing the entire series.

This change of reference base period is usually referred to as 'shifting" the base.

Reasons for shifting the base

(i) The period base has become too old and is almost useless for purposes of comparison. By shifting the base, it is possible to state the series in terms of a more recent time period.

(ii) It may be desired to compare several index number series which have been computed on different base periods.

Mathematically speaking, this method is strictly applicable only if the index numbers satisfy the circular test.

12.8 Splicing

The process of splicing is very simple and is akin to that used in shifting the base.

Thus,

$$\text{Spliced index} = \frac{(Index\ number\ of\ current\ year)*(Index\ number\ of\ new\ base\ year)}{100} \quad (12.31)$$

12.9 Deflating

By deflating of the price index numbers, it implies adjusting them after making allowance for the effect of changing price levels. A rise in price level means a reduction in the purchasing power of money. Hence, the purchasing power of money is the reciprocal of the price index.

Also, the real wages (income) is obtained by dividing the money or nominal income by the corresponding appropriate price index and multiplying the result by 100. The real income is also known as the deflated income.

12.10 Consumer Price Index (CPI) Numbers (cost of living Index Numbers)

They are generally intended to represent the average change over time in the prices paid by the ultimate consumer of a specified basket of goods and services.

The need for constructing CPI's arises because the general index numbers fail to give an exact idea of the effect of the change in the general price level on the cost of living of different classes of people in different manners. It should be noted that different classes of people consume different types of commodities and even the same type of commodities are not consumed in the same proportion by different classes of people.

For instance, the consumption pattern of rich, poor and middle class people varies widely. It should also be noted that the consumption habits of the people of the same class varies from place to place.

12.10.1 Methods of constructing the CPI

(i) Aggregate Expenditure Method (AEM) or Aggregate method.

Here, the quantities of commodities consumed by the particular group in the base year are estimated which constitute the weights.

The CPI gives the situation where the aggregate expenditure of the current year is divided by the aggregate expenditure of the base year and the quotient is multiplied by 100 ie

$$\text{CPI} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \qquad (12.32)$$

which is the Laspeyres method and thus is the most popular method for constructing CPI

(ii) Family Budget Method (FBM)

Here, the weights are obtained which are the family budgets of a large number of people for whom the index is meant and the aggregate expenditure of an average family on which various items are estimated. The weights are the value weights obtained by multiplying the prices by quantities consumed (ie $p_0 q_0$). The price relatives for each commodity are obtained and these price relatives are multiplied by the value weights for each item and the product is divided by the sum of the weights ie

$$\text{FBM} = \frac{\sum PV}{\sum V} \qquad (12.33)$$

where

P = $\frac{p_1}{p_0}\times 100$ for each item

V = value weights ie $p_O q_o$. This method is the same as the weighted average of price relatives method.

12.11 Limitations of Index Numbers

(i) Since index numbers are generally based on a sample, it is not possible to take into account each and every item in the construction of the index.
(ii) While taking the sample, random sampling is seldomly used.
(iii) It is often difficult to take into account changes in the quantity of products.
(iv) A large number of methods are designed for constructing index numbers and different methods of computations give different results.
(v) Index numbers can also be misused in such a manner as to draw the desired conclusion.
(vi) Lack of adequate and accurate data often becomes a serious limitation of the index itself.

12.12 Empirical Examples

1. For the data given below, calculate the index number by taking
   i. 2000 as the base year
   ii. 2007 as the base year
   iii. 2000 to 2002 as the base period

Year	Price of commodity (X)
2000	4
2001	5
2002	6
2003	7
2004	8
2005	10
2006	9
2007	10
2008	11

Solution

(i) Taking 2000 as the base year

Year	Price of commodity (X)	Index Number
2000	4	$\frac{4}{4}\times 100 = 100$
2001	5	$\frac{5}{4}\times 100 = 125$
2002	6	$\frac{6}{4}\times 100 = 150$
2003	7	175
2004	8	200
2005	10	250
2006	9	225
2007	10	250
2008	11	$\frac{11}{4}\times 100 = 275$

(ii) Index number taking 2007 as the base year

Year	Price of commodity (X)	Index Number
2000	4	$\frac{4}{10}\times 100 = 40$
2001	5	$\frac{5}{10}\times 100 = 50$
2002	6	$\frac{6}{10}\times 100 = 60$
2003	7	70
2004	8	80
2005	10	100
2006	9	90
2007	10	100
2008	11	$\frac{11}{10}\times 100 = 110$

(iii) Taking 2000 to 2002 as the base period. When 2000 to 2002 is to be taken as base, it means we have to take as average of 2000, 2001 and 2002

Average = $\frac{4+5+6}{3} = 5$

Hence 2001 will be taken as 100

Year	Price of commodity (X)	Index Number
2000	4	$\frac{4}{5}\times 100 = 80$
2001	5	$\frac{5}{5}\times 100 = 100$
2002	6	$\frac{6}{5}\times 100 = 120$
2003	7	140
2004	8	160
2005	10	200
2006	9	180
2007	10	200
2008	11	$\frac{11}{5}\times 100 = 220$

2. Construct index numbers of price from the following data by using
   i. Laspeyres method
   ii. Paasche method
   iii. Bowley's method
   iv. Fisher's Ideal method
   v. Marshall – Edgeworth method

Commodity	2007		2008	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution

Calculation of various indices

Commodity	2007		2008					
	Price P_0	Quantity q_0	Price P_1	Quantity q_1	p_1q_o	p_0q_o	p_1q_1	p_0q_1
A	2	8	4	6	32	16	24	12
B	5	10	6	5	60	50	30	25
C	4	14	5	10	70	56	50	40
D	2	19	2	13	38	38	26	26
					$\sum p_1q_0 = 200$	160	130	103

i. Laspeyre's method

$$= \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = 125$$

ii. Paasches method

$$\frac{\sum p_1q_1}{\sum p_0q_1} \times 100 = 126.21$$

iii. Bowley's method

$$\frac{\frac{\sum p_1q_0}{\sum p_0q_0} + \frac{\sum p_1q_1}{\sum p_0q_1}}{2} \times 100 = 125.6$$

iv. Fisher's Ideal Method

$$\sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100 = 126.6$$

v. Marshall – Edgeworth Method

$$\frac{\sum (q_0 + q_1)p_1}{\sum (q_0 + q_1)p_0} \times 100 = 125.47$$

3. The following table gives the prices of some food items in the base year and current year and the quantities sold in the base year. Calculate i) the arithematic mean and ii) geometric mean

Commodity	P_0 (₦)	q_0 (Kg)	P_1 (₦)
Sugar	3.0	20	4.0
Flour	1.5	40	1.6
Milk	1.0	10	1.5

Solution

(i) Index number using weighted arithematic mean of price relatives

Commodity	P_0	q_0	P_1	$P_0q_0 = (V)$	$\frac{P_1}{P_0} \times 100 = (P)$	**PV**
Sugar	3.0	20	4.0	60	$\frac{4}{3} \times 100$	8,000
Flour	1.5	40	1.6	60	$\frac{1.6}{1.5} \times 100$	6,400
Milk	1.0	10	1.5	10	$\frac{1.5}{1.0} \times 100$	1,500
					$\sum V = 130$	$\sum PV = 15{,}900$

which implies that there has been a 22.31% increase in prices over the base level

(ii) Index number using Geometric mean of price relatives

Commodity	P_0	q_0	P_1	**V**	**P**	**Log (P)**	**V Log (P)**
Sugar	3.0	20	4.0	60	133.3	2.1249	127.494
Flour	1.5	40	1.6	60	106.7	2.0282	121.692
Milk	1.0	10	1.5	10	150.0	2.1761	21.761
				$\sum V = 130$			$\sum VLog(P) = 270.947$

$$P_{01} = anti\log\left[\frac{\sum VLog(P)}{\sum V}\right] = anti\log\left[\frac{270.947}{130}\right] = 120.9$$

4. By using the average of the quantities of two years as weights , compute a price index

Commodity	**Quantities**		**Prices**	
	2002	**2003**	**2002**	**2003**
A	10	16	20	25
B	9	7	25	28
C	20	24	40	40

Solution

Computation of price index using average of the two year quantities as weights

Commodity	**Quantities**			**Prices**			
	2002 (q_0)	**2003** (q_1)	$\frac{q_0 + q_1}{2}$ (q)	**2002** (P_0)	**2003** (P_1)	P_1q	P_0q
A	10	16	13	20	25	325	260
B	9	7	8	25	28	224	200
C	20	24	22	40	40	880	880

Applying Kelly's method

$$P_{01} = \frac{\sum p_1 q}{\sum p_0 q} \times 100 = 106.64$$

where $\sum p_1 q = 1{,}429$

$\sum p_0 q = 1{,}340$

5. From the following data of the wholesale prices of wheat for 10 years, construct index numbers taking 2004 as base using chain base method.

Year	Price of Wheat
2004	50
2005	60
2006	62
2007	65
2008	70
2009	78
2010	82
2011	84
2012	88
2013	90

Solution

Construction of chain indices

Year	Price of Wheat	Link Relative	Chain Index
2004	50	100	100
2005	60	$\frac{60}{50} \times 100 = 120.00$	$\frac{120 \times 100}{100} = 120$
2006	62	$\frac{62}{50} \times 100 = 103.33$	$\frac{103.33 \times 124}{100} = 120$
2007	65	104.84	130
2008	70	107.69	140
2009	78	111.43	156
2010	82	105.13	164
2011	84	102.44	168
2012	88	104.76	176
2013	90	102.27	180

EXERCISE TWELVE

(1) For the data in Table 12.1 calculate and interpret the (a) Domestic Product (GDP) and (b) Per Capita Income (PCI)

(2) For the data in Table 12.2 (a) calculate and interpret the dependency ratios for populations (i) A (ii) B (iii) C (iv) D (b) What is your impression about the dependency burden in these countries.

(3) For the data in Table 12.3 calculate and interpret (a) the Sex Ratio (SR)

(i) of the population (ii) at Birth (SRB) (iii) at Death (SRD) (b) (i) Crude Birth Rate (CBR)
(ii) Crude Death Rate (CDR).

(4) For the data in Table 12.4 calculate and interpret (a) Crude Marriage Rate (CMR) for the (i) males and (ii) females, (b) Crude rate of Divorce (CRD) for the (i) males and (ii) females

(5) From the data given below, compute a quantity index

Commodity	**Quantities**		**Price**
	2002	**2003**	**2003**
A	30	25	30
B	20	30	40
C	10	15	20

(6) Calculate Fisher's Ideal index from the following data and prove that it satisfies both Time Reversal and Factor Reversal tests.

Commodity	**Year 2013**		**Year 2014**	
	Price	**Expenditure**	**Price**	**Expenditure**
A	8	80	10	120
B	10	120	12	96
C	5	40	5	50
D	4	56	3	60
E	20	100	25	150

(7) Using data of Example 5, construct index numbers taking 2004 as the base

Year	Price of Wheat	Index nos (2004 =100)
2004	50	100
2005	60	$\frac{60}{50}\times 100 = 120$
2006	62	$\frac{62}{50}\times 100 = 124$
2007	65	$\frac{65}{50}\times 100 = 130$
2008	70	$\frac{70}{50}\times 100 = 140$
2009	78	156
2010	82	164
2011	84	168
2012	88	176
2013	90	180

It should be noted that from 2004 to 2005, there is a 20% increase. From 2005 to 2006, there is a 24% increase etc

(8) Compute the chain index number with 2009 prices as base from the following table giving the average wholesale prices of the commodities A, B, and C for the year 2010.

Commodity	2009	2010	2011	2012	2013
A	20	16	28	35	21
B	25	30	24	36	45
C	20	25	30	24	30

(9) Convert the following FBI (Fixed base index) numbers into chain base index.

Year	2000	2001	2002	2003	2004	2005
FBI	376	392	408	380	392	400

(10) Construct the CPI number for 2013 on the basis of 2012 from the following data using (i) The aggregate expenditure method.

Commodity	Quantity consumed in 2012	Price in 2012	Price in 2013
A	6	5.75	6
B	6	5	8
C	1	6	9
D	6	8	10
E	4	2	1.50
F	1	20	15

(11) Repeat example 10 using the Family Budget method

CHAPTER THIRTEEN

ANALYSIS OF COVARIANCE

13.0 INTRODUCTION

Just like the analysis of variance, analysis of covariance refers to the partitioning of total cross – product into recognized sources of variation. In order to increase the accuracy of estimates, experiments are usually planned in such a way that effects of environmental factors are eliminated from estimates of treatment effects. An example is when an experiment is conducted to compare the effects of different feeds on the growth of certain species of birds. If it is suspected that the result may be affected by the initial weight of the birds, then the experiment may be conducted in such a way as to control the effect of the initial weight of the birds. Blocking and stratification are sometimes adopted to achieve this and increase the accuracy of estimates. Another approach is to treat the initial weight as covariate to the final weight and apply the method of analysis of covariance. The method of analysis of covariance makes use of both analysis of variance and regression analysis to achieve improved accuracy.

In analysis of covariance, there are two measures: the main measure; designated by Y and the supplementary measure; designated by X. Interest is in comparing the means ($\overline{Y}_{.j}$) of the main variable Y obtained after treatments, t_j, $j=1,2,...s$, have been applied. However, if Y_{ij} is substantially correlated with (affected by) a supplementary measure; X_{ij}, then the analysis of covariance will result in a smaller estimate of experimental error

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} \qquad (13.1)$$

The layout of a typical problem for analysis of covariance is given in Table 13.1

Table 13.1: Experimental Layout for Analysis of covariance problem

I	Group (j)				Sum	Mean
	1	2	...	s	$X_{i.}$ $Y_{i.}$	$\overline{X}_{i.}$ $\overline{Y}_{i.}$
1	X_{11} Y_{11}	X_{12} Y_{12}	...	X_{1s} Y_{1s}	$X_{1.}$ $Y_{1.}$	$\overline{X}_{1.}$ $\overline{Y}_{1.}$
2	X_{21} Y_{21}	X_{22} Y_{22}	...	X_{2s} Y_{2s}	$X_{2.}$ $Y_{2.}$	$\overline{X}_{2.}$ $\overline{Y}_{2.}$
. . .			...			
M	X_{m1} Y_{m1}	X_{m2} Y_{m2}	...	X_{ms} Y_{ms}	$X_{m.}$ $Y_{m.}$	$\overline{X}_{m.}$ $\overline{Y}_{m.}$
Sum	$X_{.1}$ $Y_{.1}$	$X_{.2}$ $Y_{.2}$	...	$X_{.s}$ $Y_{.s}$	$X_{..}$ $Y_{..}$	-
Mean	$\overline{X}_{.1}$ $\overline{Y}_{.1}$	$\overline{X}_{.2}$ $\overline{Y}_{.2}$	...	$\overline{X}_{.s}$ $\overline{Y}_{.s}$	-	$\overline{X}_{..}$ $\overline{Y}_{..}$

where $\bar{X}_{i.} = \frac{1}{s}\sum_{j=1}^{s} X_{ij}$, $\bar{X}_{.j} = \frac{1}{m}\sum_{i=1}^{m} X_{ij}$, $\bar{X}_{..} = \frac{1}{ms}\sum_{j=1i=1}^{s}\sum^{m} X_{ij}$,

$\bar{Y}_{i.} = \frac{1}{s}\sum_{j=1}^{s} Y_{ij}$, $\bar{Y}_{.j.} = \frac{1}{m}\sum_{i=1}^{m} Y_{ij}$, $\bar{Y}_{..} = \frac{1}{ms}\sum_{j=1i=1}^{s}\sum^{m} Y_{ij}$

13.1 Models for Analysis of Covariance

The model for analysis of covariance is given by

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + e_{ij} \tag{13.2}$$

or more conveniently,

$$Y_{ij} = \mu + \alpha_j + \beta(X_{ij} - \bar{X}_{..}) + e_{ij} \tag{13.3}$$

where Y_{ij} is the ith response (main effect) at the jth treatment in the presence of the supplementary variable,

X_{ij} is the ith effect of the supplementary variable in the jth treatment,

μ is the grand mean (mean in the absence of any treatment),

α_j is the mean effect of the jth treatment,

β is the coefficient of regression of Y_{ij} on X_{ij}

e_{ij} is the error associated with Y_{ij}

13.1.1 Assumptions of Analysis of Covariance

(i) The X's are fixed, measured without error and independent of the treatments,

(ii) In the absence of the treatment effect, the regression of Y on X is linear and independent of treatment. (Hence, the use of β not β_j).

(iii) The residuals are normally and independently distributed with zero mean and common variance.

13.1.2 Parameter Estimates

From (11.2)

$$e_{ij} = Y_{ij} - (\mu + \alpha_j + \beta X_{ij}) \tag{13.4}$$

$$\sum_{j=1i=1}^{s}\sum^{m} e_{ij}^2 = \sum_{j=1i=1}^{s}\sum^{m}\left[Y_{ij} - (\mu + \alpha_j + \beta X_{ij})\right]^2 \tag{13.5}$$

The ordinary Least Squares (OLS) estimates of the parameters are:

$$\hat{\mu} = \bar{Y}.. - \hat{\beta}\bar{X}_{..} \tag{13.6}$$

$$\hat{\alpha}_j = \overline{Y}_{.j} - \overline{Y}_{..} - \hat{\beta}(\overline{X}_{.j} - \overline{X}_{..}) \tag{13.7}$$

$$\hat{\beta} = \frac{\sum_{j=1}^{s}\sum_{i=1}^{m}(Y_{ij} - \overline{Y}..)(X_{ij} - \overline{X}_{..})}{\sum_{j=1}^{s}\sum_{i=1}^{m}(X_{ij} - \overline{X}_{..})^2} \tag{13.8}$$

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}X_{ij} \tag{13.9}$$

$$\hat{Y}_{ij} = (\overline{Y}_{..} - \hat{\beta}\overline{X}_{..}) + [(\overline{Y}_{.j} - \overline{Y}_{..}) - \hat{\beta}(\overline{X}_{.j} - \overline{X}_{..})] + \hat{\beta}X_{ij} \tag{13.10}$$

or

$$\hat{Y}_{ij} = \overline{Y}_{.j} + \hat{\beta}(X_{ij} - \overline{X}_{.j}) \tag{13.11}$$

The component $\hat{Y}_{ij} - \overline{Y}_{..}$ is the total effect, $(\overline{Y}_{.j} - \overline{Y}_{..})$ is the treatment effect while $\hat{\beta}(X_{ij} - \overline{X}_{..})$ is regression effect;.

From (13.1) and (13.11)

$$\hat{e}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \left[\overline{Y}_{.j} + \hat{\beta}(X_{ij} - \overline{X}_{.j})\right]$$

$$= (Y_{ij} - \overline{Y}_{.j}) - \hat{\beta}(X_{ij} - \overline{X}_{..})$$

So that the error sum of squares becomes

$$\sum_{j=1i=1}^{s}\sum^{m}\hat{e}_{ij}^2 = \sum_{j=1i=1}^{s}\sum^{m}\left[Y_{ij} - (\hat{\mu} + \hat{\alpha}_j + \hat{\beta}X_{ij})\right]^2$$

$$= \sum_{j=1i=1}^{s}\sum^{m}\left[(Y_{ij} - \overline{Y}_{.j}) - \hat{\beta}(X_{ij} - \overline{X}_{..})\right]^2$$

$$= \sum_{j=1i=1}^{s}\sum^{m}\left[(Y_{ij} - \overline{Y}_{.j})^2 - 2\hat{\beta}(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{..}) + \hat{\beta}^2(X_{ij} - \overline{X}_{..})^2\right]$$

$$= \sum_{j=1i=1}^{s}\sum^{m}(Y_{ij} - \overline{Y}_{.j})^2 - 2\hat{\beta}\sum_{j=1i=1}^{s}\sum^{m}(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{..}) + \hat{\beta}^2\sum_{j=1i=1}^{s}\sum^{m}(X_{ij} - \overline{X}_{..})^2$$

$$= \sum_{j=1i=1}^{s}\sum^{m}(Y_{ij} - \overline{Y}_{.j})^2 - 2\hat{\beta}\sum_{j=1i=1}^{s}\sum^{m}(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{..}) + \hat{\beta}\sum_{j=1i=1}^{s}\sum^{m}(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{..})$$

$$= \sum_{j=1i=1}^{s}\sum^{m}(Y_{ij} - \overline{Y}_{.j})^2 - \hat{\beta}\sum_{j=1i=1}^{s}\sum^{m}(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{..})$$

$$= \sum_{j=1i=1}^{s}\sum^{m}(Y_{ij} - \overline{Y}_{.j})^2 - \frac{\sum_{j=1}^{s}\sum_{i=1}^{m}(Y_{ij} - \overline{Y}..)(X_{ij} - \overline{X}_{..})}{\sum_{j=1}^{s}\sum_{i=1}^{m}(X_{ij} - \overline{X}_{..})^2}\sum_{j=1i=1}^{s}\sum^{m}(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{..})$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{m}(Y_{ij} - \overline{Y}_{.j})^2 - \frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m}(Y_{ij} - \overline{Y}..)(X_{ij} - \overline{X}_{..})\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m}(X_{ij} - \overline{X}_{..})^2} \tag{13.12}$$

Recall, in analysis of variance, the total deviation from the grand mean, $X_{ij} - \overline{X}_{..}$ is partitioned as

$$X_{ij} - \overline{X}_{..} = X_{ij} - \overline{X}_{.j} + \overline{X}_{.j} - \overline{X}_{..} \tag{13.13}$$

and the total sum of squares is computed as

$$\sum_{j=1}^{s}\sum_{i=1}^{m}(X_{ij} - \overline{X}_{..})^2 = \sum_{j=1}^{s}\sum_{i=1}^{m}[(X_{ij} - \overline{X}_{.j}) + (\overline{X}_{.j} - \overline{X}_{..})]^2 \tag{13.14}$$

Similarly, for analysis of covariance, the total deviation from the grand mean, can be partitioned as

$$Y_{ij} - \overline{Y}_{..} = Y_{ij} - \overline{Y}_{.j} + \overline{Y}_{.j} - \overline{Y}_{..} \tag{13.15}$$

for Y variate.

Therefore, the cross product sum (instead of total sum of squares) is given by

$$\sum_{j=1}^{s}\sum_{i=1}^{m}(Y_{ij} - \overline{Y}_{..})(X_{ij} - \overline{X}_{..})$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{m}[(Y_{ij} - \overline{Y}_{.j}) + (\overline{Y}_{.j} - \overline{Y}_{..})][(X_{ij} - \overline{X}_{.j}) + (\overline{X}_{.j} - \overline{X}_{..})] \tag{13.16}$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{m}[(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{.j}) + (\overline{Y}_{.j} - \overline{Y}_{..})(X_{ij} - \overline{X}_{.j})$$

$$+ (Y_{ij} - \overline{Y}_{.j})(\overline{X}_{.j} - \overline{X}_{..}) + (\overline{Y}_{.j} - \overline{Y}_{..})(\overline{X}_{.j} - \overline{X}_{..})]$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{m}[(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{.j})] + \sum_{j=1}^{s}\sum_{i=1}^{m}[(\overline{Y}_{.j} - \overline{Y}_{..})(X_{ij} - \overline{X}_{.j})]$$

$$+ \sum_{j=1}^{s}\sum_{i=1}^{m}[(Y_{ij} - \overline{Y}_{.j})(\overline{X}_{.j} - \overline{X}_{..})] + \sum_{j=1}^{s}\sum_{i=1}^{m}[(\overline{Y}_{.j} - \overline{Y}_{..})(\overline{X}_{.j} - \overline{X}_{..})]$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{m}[(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{.j})] + \sum_{j=1}^{s}(\overline{Y}_{.j} - \overline{Y}_{..})\sum_{i=1}^{m}(X_{ij} - \overline{X}_{.j})$$

$$+ \sum_{j=1}^{s}(\overline{X}_{.j} - \overline{X}_{..})\sum_{i=1}^{m}(Y_{ij} - \overline{Y}_{.j}) + \sum_{j=1}^{s}\sum_{i=1}^{m}[(\overline{Y}_{.j} - \overline{Y}_{..})(\overline{X}_{.j} - \overline{X}_{..})]$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{m}[(Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{.j})] + \sum_{j=1}^{s}\sum_{i=1}^{m}[(\overline{Y}_{.j} - \overline{Y}_{..})(\overline{X}_{.j} - \overline{X}_{..})] \tag{13.17}$$

since $\sum_{i=1}^{m}(Y_{ij} - \overline{Y}_{.j}) = \sum_{i=1}^{m}(X_{ij} - \overline{X}_{.j}) = 0$.

Case A: When regression coefficients are not the same for all groups

In the absence of the treatment effect the regression model for group j is,

$$Y_{ij} = \mu + \beta_j X_{ij} + e_{ij} \tag{13.18}$$

where Y_{ij} is the ith response (main effect) at the jth treatment in the presence of the supplementary variable

X_{ij} is the ith effect of the supplementary variable in the jth treatment

μ is the grand mean

β_j is the coefficient of regression of Y_{ij} on X_{ij} in the jth group

e_{ij} is the error associated with Y_{ij}

From (13.18)

$$\hat{e}_{ij} = Y_{ij} - (\hat{\mu} + \hat{\beta}_j X_{ij}) \tag{13.19}$$

$$\sum_{i=1}^{m} \hat{e}_{ij}^2 = \sum_{i=1}^{m} \left[Y_{ij} - (\hat{\mu} + \hat{\beta}_j X_{ij}) \right]^2 \tag{13.20}$$

The OLS estimates of the parameters are:

$$\hat{\mu} = \overline{Y}_{.j} - \hat{\beta}_j \overline{X}_{.j} \tag{13.21}$$

$$\hat{\beta}_j = \frac{\sum_{i=1}^{m} (Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{.j})}{\sum_{i=1}^{m} (X_{ij} - \overline{X}_{.j})^2} \tag{13.22}$$

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\beta} X_{ij}$$

$$= \overline{Y}_{.j} + \hat{\beta}(X_{ij} - \overline{X}_{.j}) \tag{13.23}$$

$$\sum_{i=1}^{m} \hat{e}_{ij}^2 = \sum_{i=1}^{m} \left[Y_{ij} - [\overline{Y}_{.j} + \hat{\beta}_j (X_{ij} - \overline{X}_{.j})] \right]^2$$

$$= \sum_{i=1}^{m} \left[(Y_{ij} - \overline{Y}_{.j}) - \hat{\beta}_j (X_{ij} - \overline{X}_{.j}) \right]^2$$

$$= \sum_{i=1}^{m} (Y_{ij} - \overline{Y}_{.j})^2 - 2\hat{\beta}_j \sum_{i=1}^{m} (Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{.j}) + \hat{\beta}_j^2 \sum_{i=1}^{m} (X_{ij} - \overline{X}_{.j})^2 \tag{13.24}$$

If we let $y_{ij} = (Y_{ij} - \overline{Y}_{.j})$ and $x_{ij} = (X_{ij} - \overline{X}_{.j})$, then

$$\sum_{i=1}^{m} \hat{e}_{ij}^2 = \sum_{i=1}^{m} y_{ij}^2 - 2\hat{\beta}_j \sum_{i=1}^{m} y_{ij} x_{ij} + \hat{\beta}_j^2 \sum_{i=1}^{m} x_{ij}^2 \tag{13.25}$$

From (13.22)

$$\hat{\beta}_j = \frac{\sum_{i=1}^{m} (Y_{ij} - \overline{Y}_{.j})(X_{ij} - \overline{X}_{.j})}{\sum_{i=1}^{m} (X_{ij} - \overline{X}_{.j})^2} = \frac{\sum_{i=1}^{m} x_{ij} y_{ij}}{\sum_{i=1}^{m} x_{ij}^2} \tag{13.26}$$

Therefore, the 'within treatment/group' error sum of squares is

$$\sum_{i=1}^{m} \hat{e}_{ij}^2 = \sum_{i=1}^{m} \left(Y_{ij} - \hat{Y}_{ij} \right) = \sum_{i=1}^{m} y_{ij}^2 - 2 \left(\frac{\sum_{i=1}^{m} x_{ij} y_{ij}}{\sum_{i=1}^{m} x_{ij}^2} \right) \sum_{i=1}^{m} y_{ij} x_{ij} + \left(\frac{\sum_{i=1}^{m} x_{ij} y_{ij}}{\sum_{i=1}^{m} x_{ij}^2} \right)^2 \sum_{i=1}^{m} x_{ij}^2$$

$$\sum_{i=1}^{m}\left(Y_{ij}-\hat{Y}_{ij}\right)^2=\sum_{i=1}^{m}y_{ij}^2-\frac{\left[\sum_{i=1}^{m}x_{ij}\,y_{ij}\right]^2}{\sum_{i=1}^{m}x_{ij}^2} \qquad (13.27)$$

with df $(m-2) = (m-1) - 1$

Note that two parameters (μ and β) were estimated from each group. So, the Residual Sum Squares (RSS) loses two degrees of freedom from m observations. In $\sum_{i=1}^{m}y_{ij}^2=\sum_{j=1}^{m}(Y_{ij}-\overline{Y}_{.j})^2$ estimate ($\overline{Y}_{.j}$) of the mean of the jth group ($\mu_{.j}$) was obtained from the m observations, so, the Total Sum of Squares (SSTO) loses one degree of freedom from m observations.

Table13.2: ANOVA table for each group.

Source of variation	Sum of Squares	df	Ms	F-cal
Regression	$\left(\sum_{i=1}^{m}x_{ij}\,y_{ij}\right)^2\Big/\sum_{i=1}^{m}x_{ij}^2$	1	$\left(\sum_{i=1}^{m}x_{ij}\,y_{ij}\right)^2\Big/\sum_{i=1}^{m}x_{ij}^2$	$\frac{MS(Reg)}{MSE}$
Error	$\sum_{i=1}^{m}\left(Y_{ij}-\hat{Y}_{ij}\right)^2$	m -2	$\sum_{i=1}^{m}\left(Y_{ij}-\hat{Y}_{ij}\right)^2/(m-2)$	-
Total	$\sum_{i=1}^{m}y_{ij}^2$	m -1	-	-

MS(Reg) - Mean sum of squares for regress, MSE - Error mean square

When the within treatment sum of squares are summed over all groups, the following results are obtained. From (13.27)

$$\sum_{j=1i=1}^{s}\sum^{m}\left(Y_{ij}-\hat{Y}_{ij}\right)^2=\sum_{j=1i=1}^{s}\sum^{m}y_{ij}^2-\sum_{j=1}^{s}\frac{\left(\sum_{i=1}^{m}x_{ij}\,y_{ij}\right)^2}{\sum_{i=1}^{m}x_{ij}^2}$$

$$=\sum_{j=1i=1}^{s}\sum^{m}y_{w}^2-\sum_{j=1}^{s}\frac{\left(\sum_{i=1}^{m}x_{w}\,y_{w}\right)^2}{\sum_{i=1}^{m}x_{w}^2} \qquad (13.28)$$

SSE(1) = SSTO - SSTR

with df $s(m-2) = s(m-1) - sx(1)$

(SSE - Error Sum of Squares, SSTR - Treatment Sum of Squares)

The corresponding degrees of freedom are obtained from (13.21) by summation.

Table13.3: ANOVA table when coefficients of regression are unequal

Source of variation	Sum of Squares	Df	Ms	F-cal
Regression	$\sum_{j=1}^{s}\left(\sum_{i=1}^{m} x_{ij} y_{ij}\right)^2 \Big/ \sum_{i=1}^{m} x_{ij}^2$	S	$\left[\sum_{j=1}^{s}\left(\sum_{i=1}^{m} x_{ij} y_{ij}\right)^2 \Big/ \sum_{i=1}^{m} x_{ij}^2\right] \Big/ s-1$	$\frac{MS(Reg)}{MSE}$
Error (1) S_1	$\sum_{j=1}^{s}\sum_{i=1}^{m}\left(Y_{ij} - \hat{Y}_{ij}\right)^2$	s(m-2)	$\sum_{i=1}^{m}\left(Y_{ij} - \hat{Y}_{ij}\right)^2 \Big/ s(m-2)$	-
Total	$\sum_{j=1}^{s}\sum_{i=1}^{m} y_{ij}^2$	s(m –1)	-	-

Case B: When regression coefficients are equal for all the groups

If all the groups admit the same regression coefficient, i.e. $\hat{\beta}_j = \hat{\beta}$ for all j, then the estimate of the common slope may be derived from (13.22) as

$$b_w = \frac{\sum_{j=1}^{s}\sum_{i=1}^{m} x_{ij} y_{ij}}{\sum_{j=1}^{s}\sum_{i=1}^{m} x_{ij}^2} \tag{13.29}$$

and Equation (13.18) becomes the

$$\hat{Y}_{ij} = \overline{Y}_{.j} + b_w (X_{ij} - \overline{X}_{.j}) \tag{13.30}$$

$$\sum_{i=1}^{m} \hat{e}_{ij}^2 = \sum_{i=1}^{m}\left[Y_{ij} - [\overline{Y}_{.j} + b_{wj}(X_{ij} - \overline{X}_{.j})]\right]^2 = \sum_{i=1}^{m}\left[(Y_{ij} - \overline{Y}_{.j}) - b_w (X_{ij} - \overline{X}_{.j})\right]^2$$

$$= \sum_{i=1}^{m} y_{ij}^2 - 2 b_w \sum_{i=1}^{m} y_{ij} x_{ij} + b_w^2 \sum_{i=1}^{m} x_{ij}^2 \tag{13.31}$$

Sum of squares over all groups SSE (2)

$$\sum_{j=1}^{s}\sum_{i=1}^{m} \hat{e}_{ij}^2 = \sum_{j=1}^{s}\sum_{i=1}^{m} y_{ij}^2 - 2 b_w \sum_{j=1}^{s}\sum_{i=1}^{m} y_{ij} x_{ij} + b_w^2 \sum_{j=1}^{s}\sum_{i=1}^{m} x_{ij}^2$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{m} y_{ij}^2 - 2 b_w \sum_{j=1}^{s}\sum_{i=1}^{m} y_{ij} x_{ij} + b_w^2 \sum_{j=1}^{s}\sum_{i=1}^{m} x_{ij}^2$$

$$\sum_{j=1}^{s}\sum_{i=1}^{m}\left(Y_{ij} - \hat{Y}_{ij}\right)^2 = \sum_{j=1}^{s}\sum_{i=1}^{m} y_{ij}^2 - \frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m} x_{ij} y_{ij}\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m} x_{ij}^2}$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{m} y_w^2 - \frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m} x_w y_w\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m} x_w^2} \tag{13.32}$$

with df s(m – 1) -1 = s(m – 1) - 1

Note that in $\sum_{i=1}^{m} y_{ij}^2 = \sum_{j=1}^{m} (Y_{ij} - \overline{Y}_{.j})^2$ contains estimate ($\overline{Y}_{.j}$) of the mean of the jth group ($\mu_{.j}$), obtained from the m observations. So, it loses one degree of freedom from m observations. Sum of the (m – 1) degrees of freedom over the s groups leaves $\sum_{j=1i=1}^{s} \sum^{m} y_{ij}^2$ with s(m – 1) degrees of freedom. With one degree of freedom for the second term on the right hand side (RHS), of (13.32), the RSS is left with

s(m – 1) - 1 degrees of freedom. Since Error (1) is the sum of squares of deviations from the line of best-fit, it is clear that it is smaller than sum of squares of deviations about any other line.

Table 13.5: ANOVA table when regression coefficients are equal (i.e $\hat{\beta}_j = \hat{\beta}$, for all j)

Source of variation	Sum of Squares	Df	Ms	F-cal
Regression	$\left(\sum_{j=1i=1}^{s} \sum^{m} x_{ij} y_{ij}\right) \Big/ \sum_{j=1i=1}^{s} \sum^{m} x_{ij}^2$	1	$\left[\left(\sum_{j=1i=1}^{s} \sum^{m} x_{ij} y_{ij}\right) \Big/ \sum_{j=1i=1}^{s} \sum^{m} x_{ij}^2\right] \Big/ 1$	$\frac{MS(\text{Re}\, g)}{MSE}$
Error (2) S_2	$\sum_{j=1i=1}^{s} \sum^{m} (Y_{ij} - \hat{Y}_{ij})^2$	s(m-1) -1	$\sum_{j=1}^{s} \sum_{i=1}^{m} (Y_{ij} - \hat{Y}_{ij})^2 \Big/ (s(m-1)-1$	-
Total	$\sum_{j=1i=1}^{s} \sum^{m} y_{ij}^2$	S(m –1)	-	-

Thus, it is expected that Error (2) ≥ Error (1), i.e. $S_2 \geq S_1$. Therefore, the difference,

$$S_3 = S_2 - S_1 \qquad (13.33)$$

with $[s(m-1)-1]-[s(m-2)] = [-s-1]+2s = s-1$ degrees of freedom, will be used as a measure of inequality of the regression coefficients and to test the hypothesis:

$H_0 : b_1 = b_2 = \ldots b_s$, against the alternative; $H_1 : b_j \neq b_{j'}, \forall\, j \neq j'$

The test is summarized in Table 13.6.

The test statistic is,

$$F_{cal} = \frac{S_3/(s-1)}{S_1/s(m-2)} \qquad (13.34)$$

Table 13.6: ANOVA table for Measuring inequality of regression coefficients

Source of variation	Sum of Squares	Df	Ms	F-cal
Error (1)	S_1	$s(m-2)$	$S_1/[s(m-2)]$	-
Inequality (S_3)	$S_2 - S_1$	$s-1$	$S_3/(s-1)$	$MS(S_3)/MS(S_1)$
Error (2)	S_2	$[s(m-1)-1]$	-	-

$MS(S_3)$ - Mean Square of S_3, $MS(S_1)$ - Mean Square of S_1

Under the null hypothesis, the statistic in (13.34) is zero (if S_3 is zero) and the null hypothesis is supported. The more it differs from zero, the more the null hypothesis is negated. Therefore, the null hypothesis is rejected at α level of significance, if F_{cal} exceeds the tabulated value of F-Distribution at $\nu_1 = s-1$ and $\nu_2 = s(m-2)$ degrees of freedom, or not rejected otherwise.

In line with assumption (ii), it is important to emphasize that the application of analysis of covariance assumes that the regression lines for the various treatment groups have common slope; (b_w), that is, the null hypothesis is true. So, if the null hypothesis above is not rejected, we can proceed with analysis of covariance to test the significance of the differences between treatment means. However, the emphasis will now be on the adjusted treatment means (i.e. treatment means adjusted for regression effect and not just on the treatment means).

13.2 Test of significance of the Treatment Effects.

As noted above, if the null hypothesis of common slope is not rejected, the common regression coefficient is calculated as

$$b_t = \frac{\sum_{j=1}^{s}\sum_{i=1}^{m}(Y_{ij}-\overline{Y}_{..})(X_{ij}-\overline{X}_{..})}{\sum_{j=1}^{s}\sum_{i=1}^{m}(X_{ij}-\overline{X}_{..})^2} = b_t = \frac{\sum_{j=1}^{s}\sum_{i=1}^{m}x_{ij}y_{ij}}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_{ij}^2} \qquad (13.35)$$

where, $x_{ij} = X_{ij} - \overline{X}_{..} = x_t$ and $y_{ij} = Y_{ij} - \overline{Y}_{..} = y_t$

and the fitted regression line is given by

$$\hat{Y}_{ij} = \overline{Y}_{..} + \hat{\beta}(X_{ij} - \overline{X}_{..}) \qquad (13.36)$$

The error and error sum of squares are:

$$\hat{e}_{ij} = (Y_{ij} - \hat{Y}_{ij}) = (Y_{ij} - \overline{Y}_{..}) - b_t(X_{ij} - \overline{X}_{..}) \qquad (13.37)$$

SSE (4) = $S_4 = \sum_{i=1}^{m}\hat{e}_{ij}^2 = \sum_{i=1}^{m}\left[(Y_{ij} - \overline{Y}_{..}) - b_t(X_{ij} - \overline{X}_{..})\right]^2$

$$= \sum_{i=1}^{m}\left[(y_{ij}) - b_t(x_{ij})\right]^2$$

$$= \sum_{i=1}^{m}y_{ij}^2 - 2b_t\sum_{i=1}^{m}y_{ij}x_{ij} + b_t^2\sum_{i=1}^{m}x_{ij}^2 \qquad (13.38)$$

If we substitute $b_t = \frac{\sum_{j=1}^{s}\sum_{i=1}^{m}x_{ij}y_{ij}}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_{ij}^2}$ in (13.1) into (13.4), we have

$$SSE(4) = S_4 = \sum_{j=1}^{s}\sum_{i=1}^{m} y_{ij}^2 - \frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m} x_{ij} y_{ij}\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m} x_{ij}^2}$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{m} y_t^2 - \frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m} x_t y_t\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m} x_t^2} \qquad (13.39)$$

with df $(sm-2) = (sm-1) - 1$

In the absence of treatment effect, SSE (4) should be equal to SSE (2) because both are based on the assumption of equal slope for all treatment groups. Both are error sums of squares adjusted for regression. The difference, $S_5 = S_4 - S_2$, between the two (S_4 with sm – 2 degrees of freedom and S_2 with s(m - 1) -1 degrees of freedom) is a measure of the presence, or otherwise, of treatment effect. The corresponding degrees of freedom of the difference is [(sm – 2) - (s (m - 1) -1)] = s – 1.

Table 13.7: ANOVA table for Measuring the presence or otherwise of treatment effect.

Source of variation	Sum of Squares	df	Ms	F-cal
Error (2)	S_2	$[s(m-1)-1]$	$S_2/[s(m-1)-1]$	-
Treatment (S_5)	$S_4 - S_2$	$s-1$	$S_5/(s-1)$	$MS(S_5)/MS(S_2)$
Error (3)	S_4	$sm-2$	-	-

$MS(S_5)$ - Mean Square of S_5, $MS(S_2)$ - Mean Square of S_2

The test statistic for the presence of treatment effect is

$$F_{cal} = \frac{S_5/(s-1)}{S_2/[s(m-1)-1]} = \frac{MS(S_5)}{MS(S_2)} \qquad (13.40)$$

The null hypothesis is rejected at α level of significance if $F_{cal} > F_{tab}$ or not rejected otherwise, where F_{tab}, is the tabulated value of the F-Distribution with $\nu_1 = s-1$ and $\nu_2 = [s(m-1)-1]$ degrees of freedom. From (13.34) and (13.32), Equation (11.41) can be written as

$S_5 = S_4 - S_2$

$$= \left[\sum_{j=1}^{s}\sum_{i=1}^{m} y_t^2 - \frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m} x_t y_t\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m} x_t^2}\right] - \left[\sum_{j=1}^{s}\sum_{i=1}^{m} y_w^2 - \frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m} x_w y_w\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m} x_w^2}\right] \qquad (13.41)$$

where $y_t = Y_{ij} - \overline{Y}_{..}$, $x_t = X_{ij} - \overline{X}_{..}$, $y_w = Y_{ij} - \overline{Y}_{.j}$, $x_w = X_{ij} - \overline{X}_{.w}$

The subscript t stands for 'Total sum of squares', w for 'Within treatment SS' and b for 'Between treatments SS'. From (13.4), the cross product sum is given by

$$\sum_{j=1}^{s}\sum_{i=1}^{m}(Y_{ij}-\overline{Y}_{..})(X_{ij}-\overline{X}_{..})$$

$$=\sum_{j=1}^{s}\sum_{i=1}^{m}[(Y_{ij}-\overline{Y}_{.j})(X_{ij}-\overline{X}_{.j})]+m\sum_{j=1}^{s}[(\overline{Y}_{.j}-\overline{Y}_{..})(\overline{X}_{.j}-\overline{X}_{..})]$$

$$\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t=\sum_{j=1}^{s}\sum_{i=1}^{m}x_w y_w+m\sum_{j=1}^{s}x_b y_b$$

Therefore,

$$\sum_{j=1}^{s}\sum_{i=1}^{m}x_w y_w=\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t-m\sum_{j=1}^{s}x_b y_b$$

and

$$\frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m}x_w y_w\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_w^2}=\frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t-m\sum_{j=1}^{s}x_b y_b\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_w^2} \qquad (13.42)$$

Hence, (13.27) may be re-written as

$$S_5=\left[\sum_{j=1}^{s}\sum_{i=1}^{m}y_t^2-\frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_t^2}\right]-\left[\sum_{j=1}^{s}\sum_{i=1}^{m}y_w^2-\frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t-m\sum_{j=1}^{s}x_b y_b\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_w^2}\right]$$

$$=\left[\sum_{j=1}^{s}\sum_{i=1}^{m}y_t^2-\sum_{j=1}^{s}\sum_{i=1}^{m}y_w^2\right]-\left[\frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_t^2}-\frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t-m\sum_{j=1}^{s}x_b y_b\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_w^2}\right]$$

$$=\sum_{j=1}^{s}\sum_{i=1}^{m}y_b^2-\left[\frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_t^2}-\frac{\left(\sum_{j=1}^{s}\sum_{i=1}^{m}x_w y_w\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{m}x_w^2}\right] \qquad (13.43)$$

where,

$$\sum_{j=1}^{s}\sum_{i=1}^{m}y_b^2=\sum_{j=1}^{s}\sum_{i=1}^{m}y_t^2-\sum_{j=1}^{s}\sum_{i=1}^{m}y_w^2 \qquad (13.44)$$

$$\sum_{j=1}^{s}\sum_{i=1}^{m}x_w y_w=\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t-m\sum_{j=1}^{s}x_b y_b \qquad (13.45)$$

Observe that from $\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t=\sum_{j=1}^{s}\sum_{i=1}^{m}x_w y_w+m\sum_{j=1}^{s}x_b y_b$, with $x_b=\overline{X}_{.j}-\overline{X}_{..}$, it is clear that if $\overline{X}_{.j}$ is identical, then $\sum_{j-1}^{s}x_b^2=\sum_{j-1}^{s}(\overline{X}_{.j}-\overline{X}_{..})^2=0$. So also $\sum_{j=1}^{s}x_b y_b=0$ and hence, from (11.31),

$$\sum_{j=1}^{s}\sum_{i=1}^{m}x_w y_w=\sum_{j=1}^{s}\sum_{i=1}^{m}x_t y_t$$

13.3 Test of significance of the adjusted treatment means.

When the null hypothesis on the absence of treatment effect is rejected, the next step is to determine the level(s) of the treatment that are actually different from others (ie multiple comparisons). The commonest approach to multiple comparisons is the pair wise comparison.

Let the adjusted treatment mean be

$$\bar{Y}'_{.j} = \bar{Y}_{.j} - b_w(\bar{X}_{.j} - \bar{X}_{..}) \quad (13.46)$$

$$\bar{Y}'_{.j'} = \bar{Y}_{.j'} - b_w(\bar{X}_{.j'} - \bar{X}_{..})$$

Then,

$$\bar{Y}'_{.j} - \bar{Y}'_{.j'} = (\bar{Y}_{.j} - \bar{Y}_{.j'}) - b_w(\bar{X}_{.j} - \bar{X}_{.j'})$$

Recall

$$Y_{ij} = \mu + \beta X_{ij} + e_{ij}, \quad \bar{Y}_{.j} = \mu + \beta\bar{X}_{.j} + \bar{e}_{.j}, \quad \bar{Y}_{.j'} = \mu + \beta\bar{X}_{.j'} + \bar{e}_{.j'}$$

Hence,

$$\bar{Y}_{.j} - \bar{Y}_{.j'} = (\mu + \beta\bar{X}_{.j} + \bar{e}_{.j}) - (\mu + \beta\bar{X}_{.j'} + \bar{e}_{.j'})$$

$$= \beta(\bar{X}_{.j} - \bar{X}_{.j'}) + (\bar{e}_{.j} - \bar{e}_{.j'})$$

Therefore, Equation (11.45) becomes

$$\bar{Y}'_{.j} - \bar{Y}'_{.j'} = (\bar{Y}_{.j} - \bar{Y}_{.j'}) - b_w(\bar{X}_{.j} - \bar{X}_{.j'})$$

$$= \beta(\bar{X}_{.j} - \bar{X}_{.j'}) - b_w(\bar{X}_{.j} - \bar{X}_{.j'}) + (\bar{e}_{.j} - \bar{e}_{.j'})$$

$$= -(b_w - \beta)(\bar{X}_{.j} - \bar{X}_{.j'}) + (\bar{e}_{.j} - \bar{e}_{.j'}) \quad (13.47)$$

$$E(\bar{Y}'_{.j} - \bar{Y}'_{.j'}) = -(\bar{X}_{.j} - \bar{X}_{.j'})E[(b_w - \beta)] + E[(\bar{e}_{.j} - \bar{e}_{.j'})]$$

$$= -(\bar{X}_{.j} - \bar{X}_{.j'})(\beta - \beta) + [(0 - 0)] = 0$$

$$Var(\bar{Y}'_{.j} - \bar{Y}'_{.j'}) = E\left[(\bar{Y}'_{.j} - \bar{Y}'_{.j'})\right]^2 = E\left[-(\bar{X}_{.j} - \bar{X}_{.j'})(b_w - \beta) + (\bar{e}_{.j} - \bar{e}_{.j'})\right]^2$$

$$= E\left[(\bar{X}_{.j} - \bar{X}_{.j'})^2(b_w - \beta)^2 - 2(\bar{X}_{.j} - \bar{X}_{.j'})(b_w - \beta)(\bar{e}_{.j} - \bar{e}_{.j'}) + (\bar{e}_{.j} - \bar{e}_{.j'})^2\right]$$

$$= E\left[(\bar{X}_{.j} - \bar{X}_{.j'})^2(b_w - \beta)^2 - 2(\bar{X}_{.j} - \bar{X}_{.j'})(b_w - \beta)(\bar{e}_{.j} - \bar{e}_{.j'}) + (\bar{e}_{.j} - \bar{e}_{.j'})^2\right]$$

$$= E\left[(\bar{X}_{.j} - \bar{X}_{.j'})^2(b_w - \beta)^2 - 2(\bar{X}_{.j} - \bar{X}_{.j'})(b_w - \beta)(\bar{e}_{.j} - \bar{e}_{.j'}) + (\bar{e}_{.j} - \bar{e}_{.j'})^2\right]$$

$$= E\left\{(\bar{X}_{.j} - \bar{X}_{.j'})^2\left[\frac{\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})}{\sum_{i=1}^{m} x_{ij}^2}\right]^2 + (\bar{e}_{.j} - \bar{e}_{.j'})^2\right.$$

$$- 2(\bar{X}_{.j} - \bar{X}_{.j'}) \left[\frac{\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})}{\sum_{i=1}^{m} x_{ij}^2} \right] (\bar{e}_{.j} - \bar{e}_{.j'}) \Bigg\}$$

$$= \left[\frac{(\bar{X}_{.j} - \bar{X}_{.j'})^2}{\left(\sum_{i=1}^{m} x_{ij}^2 \right)^2} E\left[\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j}) \right]^2 + E[(\bar{e}_{.j} - \bar{e}_{.j'})^2] \right]$$

$$- 2 \frac{(\bar{X}_{.j} - \bar{X}_{.j'})}{\sum_{i=1}^{m} x_{ij}^2} \sum_{i=1}^{m} x_{ij} E[(e_{ij} - \bar{e}_{.j})(\bar{e}_{.j} - \bar{e}_{.j'})]$$

$$= \frac{(\bar{X}_{.j} - \bar{X}_{.j'})^2}{\left(\sum_{i=1}^{m} x_{ij}^2 \right)^2} E\left[\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j}) \right]^2 + E[(\bar{e}_{.j} - \bar{e}_{.j'})^2 + 0$$

$$= \frac{(\bar{X}_{.j} - \bar{X}_{.j'})^2}{\left(\sum_{i=1}^{m} x_{ij}^2 \right)^2} E\left[\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j}) \right]^2 + E[(\bar{e}_{.j} - \bar{e}_{.j'})^2$$

$$= \frac{(\bar{X}_{.j} - \bar{X}_{.j'})^2}{\left(\sum_{i=1}^{m} x_{ij}^2 \right)^2} \frac{(m-1)\sigma^2}{m} \sum_{i=1}^{m} x_{ij}^2 + \frac{\sigma^2}{m_j} + \frac{\sigma^2}{m_{j'}} = \frac{\sigma^2}{m_j} + \frac{\sigma^2}{m_{j'}} + \frac{(\bar{X}_{.j} - \bar{X}_{.j'})^2}{\sum_{i=1}^{m} x_{ij}^2} \frac{(m-1)\sigma^2}{m}$$

$$\sigma^2(\bar{Y}_{j}' - \bar{Y}_{j'}') = \sigma^2 \left[\frac{1}{m_j} + \frac{1}{m_{j'}} + \frac{(m-1)}{m} \frac{(\bar{X}_{.j} - \bar{X}_{.j'})^2}{\sum_{i=1}^{m} x_{ij}^2} \right] \tag{13.48}$$

Therefore, the standard error of $\bar{Y}_{.j} - \bar{Y}_{.j'}$ is given by

$$\sigma(\bar{Y}_{j}' - \bar{Y}_{j'}') = \sqrt{\sigma^2 \left[\frac{1}{m_j} + \frac{1}{m_{j'}} + \frac{(m-1)}{m} \frac{(\bar{X}_{.j} - \bar{X}_{.j'})^2}{\sum_{i=1}^{m} x_{ij}^2} \right]} \tag{13.49}$$

where $\hat{\sigma}^2 = S_2$. For different pairs of $(\bar{Y}_{.j}', \bar{Y}_{.j'}')$, $\hat{\sigma}^2(\bar{Y}'_{j} - \bar{Y}_{j'}')$ takes different values. Thus, Finney (1946) has suggested the use of

$$\hat{\sigma}^2(\bar{Y}'_{j} - \bar{Y}_{j'}') = \frac{S_2}{df(S_2)} \left(1 + \frac{\sum_{j=1}^{s} \sum_{i=1}^{m} x_b^2}{(m-1) \sum_{j=1}^{s} \sum_{i=1}^{m} x_w^2} \right) \tag{13.50}$$

with S_2 degrees of freedom

if $m_j = m_{j'}$

$$\sigma(\bar{Y}_{j}' - \bar{Y}_{j'}') \approx \sqrt{\frac{S_2}{df(S_2)}\left[\frac{1}{m} + \frac{1}{m} + \frac{(\bar{X}_{.j} - \bar{X}_{.j'})^2}{\sum_{i=1}^{m} x_{ij}^2}\right]} \qquad (13.51)$$

To show $b_w - \beta = \dfrac{\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})}{\sum_{i=1}^{m} x_{ij}^2}$, see Appendix A

It can also be shown (see Appendix B) that

(a) $E\left[\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})\right]^2 = \dfrac{(m-1)\sigma^2}{m}\sum_{i=1}^{m} x_{ij}^2$,

(b) $E[(\bar{e}_{.j} - \bar{e}_{.j'})^2] = \dfrac{\sigma^2}{m_j} + \dfrac{\sigma^2}{m_{j'}}$ if $m_j = m_{j'} = m$

(c) $E[(e_{ij} - \bar{e}_{.j})(\bar{e}_{.j} - \bar{e}_{.j'})] = 0$,

13.4 Empirical Examples

11.1 The data in Table 11.8 refers to the supplementary ($X_{ij.}$) and the principal (Y_{ij}) measures obtained in a randomized group design with m = 4 subjects assigned to each of three treatments (Edwards p 391).

Solution

From Table 13.8, s (number of groups) = 3 and m (no of observations per group) = 4. Other computations are given in Table 13.8.

Table 13.8: Hypothetical example

I	Treatment		
	1 $X_{ij.}$, Y_{ij}	2 $X_{ij.}$, Y_{ij}	3 $X_{ij.}$, Y_{ij}
1	3, 6	2, 11	2, 20
2	9, 6	7, 14	9, 21
3	16, 8	13, 18	14, 25
4	19, 13	19, 18	20, 21

Table 13.9: Solution to problem in Table 13.8

	Treatment			Sum
	1	2	3	
$\sum_{i=1}^{4} x_w$	0	0	0	0
$\sum_{i=1}^{4} x_w^2$	274.8	292.8	285.2	852.80
$\sum_{i=1}^{4} y_w$	0	0	0	0
$\sum_{i=1}^{4} y_w^2$	44.0	52.8	56.8	153.60
$\sum_{i=1}^{4} x_w y_w$	96.0	119.2	88.4	303.60
$\left(\sum_{i=1}^{4} x_w y_w\right)^2 \Big/ \sum_{i=1}^{4} x_w^2$	33.5371	48.5268	27.4003	109.46
$\sum_{t=1}^{12} x_t^2$				857.73
$\sum_{t=1}^{12} y_t^2$				657.73
$\sum_{t=1}^{12} x_t y_t$				281.933
$\bar{Y}_j'$	8.7627	16.4611	23.1763	

$$S_1 = \sum_{j=1}^{s}\sum_{i=1}^{m} y_w^2 - \sum_{j=1}^{s} \frac{\left(\sum_{i=1}^{m} x_w y_w\right)^2}{\sum_{i=1}^{m} x_w^2} = 153.6 - 109\ 46 = 44.14 \text{ [with } s(m-2) = 6 \text{ df}$$

$$S_2 = \sum_{j=1i=1}^{3}\sum^{4} y_w^2 - \frac{\left(\sum_{j=1i=1}^{3}\sum^{4} x_w y_w\right)^2}{\sum_{j=1i=1}^{3}\sum^{4} x_w^2} = 153.60 - \frac{(303.60)^2}{852.80} = 153.60 - 108\ 08$$

$= 45.5173$ (with sm- s -1= 8 df)

$S_3 = S_2 - S_1 = 1.3773$ (with 8 – 6 = 2 df)

$$F_{cal} = \frac{S_3/2}{S_1/6} = \frac{0.6886}{7.3567} = 0.0932 < F_{0.05}^{2,6} = 5.14$$

Thus, H_0 is not rejected, indicating that slope is the same for all groups. So we proceed with analysis of covariance.

$$S_4 = \sum_{j=1}^{3}\sum_{i=1}^{4} y_t^2 - \frac{\left(\sum_{j=1}^{3}\sum_{i=1}^{4} x_t\, y_t\right)^2}{\sum_{j=1}^{3}\sum_{i=1}^{4} x_t^2} = 657.73 - \frac{(281.933)^2}{857.73} = 657.73 - 92.6704$$

$= 564.3295$ (with sm – 2 = 10 df)

$S_5 = S_4 - S_2$ = 518.8123 (with 10 – 8 = 2 df)

$$F_{cal} = \frac{S_5/2}{S_2/8} = \frac{259.4061}{5.6897} = 45.5925 > F_{0.01}^{2,8} = 8.65$$

Thus, H_0 is rejected, indicating that treatment effect is not the same for all groups.

To determine the treatment mean that is significantly different from others, we proceed as in Equations (4.5). First, we determine the standard error of $\overline{Y}'_j - \overline{Y}_{j'}'$

$$\sum_{j=1}^{s}\sum_{i=1}^{m}(X_{ij} - \overline{X}_{..})^2 = \sum_{j=1}^{s}\sum_{i=1}^{m}(X_{ij} - \overline{X}_{.j})^2 + \sum_{j=1}^{s}\sum_{i=1}^{m}(\overline{X}_{.j} - \overline{X}_{..})^2$$

$$\sum_{j=1}^{s}\sum_{i=1}^{m} x_t^2 = \sum_{j=1}^{s}\sum_{i=1}^{m} x_w^2 + \sum_{j=1}^{s}\sum_{i=1}^{m} x_b \quad \Rightarrow \sum_{j=1}^{s}\sum_{i=1}^{m} x_b = \sum_{j=1}^{s}\sum_{i=1}^{m} x_t^2 - \sum_{j=1}^{s}\sum_{i=1}^{m} x_w^2 = 857.73 - 852.80 = 4.93$$

$$\hat{\sigma}^2(\overline{Y}'_j - \overline{Y}_{j'}') = \frac{S_2}{df}\left(1 + \frac{\sum_{j=1}^{s}\sum_{i=1}^{m} x_b^2}{(m-1)\sum_{j=1}^{s}\sum_{i=1}^{m} x_w^2}\right) = \frac{45.5173}{8} x \left(1 + \frac{4.93}{(4-1) x 852.80}\right) = 5.7006$$

$$\hat{\sigma}(\overline{Y}'_j - \overline{Y}_{j'}') = \sqrt{5.7006} = 2.3876,,$$

With $t_{cal} = \frac{\overline{Y}'_{.j} - \overline{Y}'_{.j'}}{\hat{\sigma}(\overline{Y}'_j - \overline{Y}_{j'}')}$, the 99% confidence interval for $\overline{Y}'_j - \overline{Y}_{j'}'$ is given by

$$[-t^8_{\frac{\alpha}{2}=0.005} x\, \hat{\sigma}(\overline{Y}'_j - \overline{Y}_{j'}'),\ t^8_{\frac{\alpha}{2}=0.005} x\, \hat{\sigma}(\overline{Y}'_j - \overline{Y}_{j'}')],$$

where $t^8_{\frac{\alpha}{2}=0.005} = 3.355$ and $t^8_{\frac{\alpha}{2}=0.005} x\, \hat{\sigma}(\overline{Y}'_j - \overline{Y}_{j'}') = 3.355 x 2.3876 = 8.0104$

So, a pair of adjusted treatment means $(\overline{Y}'_j, \overline{Y}'_{j'})$ is considered significantly different if $\left|\overline{Y}'_j - \overline{Y}'_{j'}\right| > 8.0104$ or not significantly different otherwise.

where $b_w = \frac{\sum_{j=1}^{3}(\sum_{i=1}^{4} x_w y_w)}{\sum_{j=1}^{3}(\sum_{i=1}^{4} x_w^2)} = \frac{\sum_{j=1}^{3}(\sum_{i=1}^{4} x_w y_w)}{\sum_{j=1}^{3}(\sum_{i=1}^{4} x_w^2)} = \frac{303.6}{852.80} = 0.3560$

and $\overline{Y}'_j = \overline{Y}_j - b_w(\overline{X}_{.j} - \overline{X}_{..})$

For the example, the values of the adjusted treatment means ($\overline{Y}'_{.j}$) and the differences between pairs of adjusted treatment means $(\overline{Y}'_j, \overline{Y}'_{j'})$ are given in Table 5.1c

Table 13.10: Differences between pairs of adjusted treatment means

	$\overline{Y}_1'$	$\overline{Y}_2'$	$\overline{Y}_3'$
	8.7627	16.4611	23.1763
$\overline{Y}'_1$	-	7.6989	14.4141
$\overline{Y}'_2$	-	-	6.7152

When the differences between pairs of adjusted treatment means $(\overline{Y}'_j, \overline{Y}'_{j'})$ are compared with 8.0104, only the treatment mean of the third group is significantly greater than that of the first group at $\alpha = 0.01$ level of significance.

13.2 The data in Table 11.11 refers to the amount of intake (supplementary measure; X_{ij}) and weight gain (main or principal measure Y_{ij}) of animals fed with different Rations.

Complete the analysis of covariance. (Steel and Torrie p 410)

Table 13.11: Amount of intake (X_{ij}) and weight gain (Y_{ij}) by different Rations

I	Ration			
	000	100	010	001
	$X_{ij.}$, Y_{ij}	$X_{ij.}$, Y_{ij}	$X_{ij.}$, Y_{ij}	$X_{ij.}$, Y_{ij}
1	209.3, 11.2	252.4, 26.1	241.5, 13.2	259.1, 24.4
2	201.1, 18.8	287.5, 31.0	286.6, 27.9	255.7, 20.8
3	286.9, 27.2	301.1, 30.6	246.8, 15.4	273.2, 24.0
4	274.6, 20.1	276.3, 24.4	270.7, 29.9	253.0, 20.8

Example 13.3 in Table 13.12 refers to the effects of Breed and age (in weeks) on the Heart Girt Circumference (in cm) of Hogs (Oyeka p 247). Complete the analysis of covariance

Table 13.12: Heart Girt Circumference (Y_{ij}) of Hogs by Breed and Age ($X_{ij.}$ weeks)

I	Breed		
	Large White	Local	Crossbreed
	$X_{ij.}$, Y_{ij}	$X_{ij.}$, Y_{ij}	$X_{ij.}$, Y_{ij}
1	0, 34.	0, 26.6.	0, 30.8.
2	24, 73.3	24, 53.8	24, 64.7
3	48, 82.2	48, 71.0	48, 75.3
4	72, 96.5	72, 79.5	72, 84.3

Table 13.14 contains the Quarterly production (Y_{ij}) in millions of Barrels between 1975 and 1982

Year ($X_{ij.}$)	Quarter			
	I	II	III	IV
1975	36.14	44.60	44.15	35.72
1976	36.19	44.63	46.95	36.90
1977	39.66	49.72	44.49	36.54
1978	41.44	49.07	48.98	39.59
1979	44.29	50.09	48.42	41.39
1980	46.11	53.44	53.00	42.52
1981	44.61	55.18	52.24	41.66
1982	47.84	54.27	52.31	41.83

Appendix A

To show that $(b_w - \beta) = \dfrac{\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})}{\sum_{i=1}^{m} x_{ij}^2}$

Recall,

$$b_w = \frac{\sum_{i=1}^{m}(Y_{ij} - \bar{Y}_{.j})(X_{ij} - \bar{X}_{.j})}{\sum_{i=1}^{m}(X_{ij} - \bar{X}_{.j})^2} = \frac{\sum_{i=1}^{m} X_{ij}(Y_{ij} - \bar{Y}_{.j})}{\sum_{i=1}^{m} x_{ij}^2} - \frac{\bar{X}_{.j}\sum_{i=1}^{m}(Y_{ij} - \bar{Y}_{.j})}{\sum_{i=1}^{m} x_{ij}^2}$$

$$= \frac{\sum_{i=1}^{m} X_{ij}(Y_{ij} - \bar{Y}_{.j})}{\sum_{i=1}^{m} x_{ij}^2} - 0 = \frac{\sum_{i=1}^{m} X_{ij}\left[\mu + \beta X_{ij} + e_{ij} - (\mu + \beta \bar{X}_{.j} + \bar{e}_{.j})\right]}{\sum_{i=1}^{m} x_{ij}^2}$$

$$= \frac{\sum_{i=1}^{m} X_{ij}\left[\beta(X_{ij} - \bar{X}_{.j}) + (e_{ij} - \bar{e}_{.j})\right]}{\sum_{i=1}^{m} x_{ij}^2}$$

$$= \frac{\sum_{i=1}^{m}[(X_{ij} - \bar{X}_{.j}) + \bar{X}_{.j}]\left[\beta(X_{ij} - \bar{X}_{.j}) + (e_{ij} - \bar{e}_{.j})\right]}{\sum_{i=1}^{m} x_{ij}^2} = \frac{\sum_{i=1}^{m}[x_{ij} + \bar{X}_{.j}]\left[\beta x_{ij} + (e_{ij} - \bar{e}_{.j})\right]}{\sum_{i=1}^{m} x_{ij}^2}$$

$$= \frac{\sum_{i=1}^{m}\left[\beta x_{ij}^2 + x_{ij}(e_{ij} - \bar{e}_{.j})\right] + \bar{X}_{.j}\sum_{i=1}^{m}\left[\beta x_{ij} + (e_{ij} - \bar{e}_{.j})\right]}{\sum_{i=1}^{m} x_{ij}^2}$$

$$= \frac{\sum_{i=1}^{m}\left[\beta x_{ij}^2 + x_{ij}(e_{ij} - \bar{e}_{.j})\right] + 0}{\sum_{i=1}^{m} x_{ij}^2}$$

Therefore,

$$b_w = \frac{\sum_{i=1}^{m}\left[\beta x_{ij}^2 + (e_{ij} - \bar{e}_{.j})x_{ij})\right]}{\sum_{i=1}^{m} x_{ij}^2}$$

$$= \frac{\beta\sum_{i=1}^{m} x_{ij}^2 + \sum_{i=1}^{m}(e_{ij} - \bar{e}_{.j})x_{ij})}{\sum_{i=1}^{m} x_{ij}^2} = \beta + \frac{\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})}{\sum_{i=1}^{m} x_{ij}^2}$$

$$E(b_w) = \beta + \frac{\sum_{i=1}^{m} x_{ij} E(e_{ij} - \bar{e}_{.j})}{\sum_{i=1}^{m} x_{ij}^2}$$

Therefore,

$$(b_w - \beta) = \beta + \frac{\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})}{\sum_{i=1}^{m} x_{ij}^2} - \beta \qquad = \frac{\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})}{\sum_{i=1}^{m} x_{ij}^2}$$

Appendix B

To show that (a) $E\left[\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})\right]^2 == \frac{(m-1)\sigma^2}{m}\sum_{i=1}^{m} x_{ij}^2$,

(b) $E[(\bar{e}_{.j} - \bar{e}_{.j'})^2] = \frac{\sigma^2}{m_j} + \frac{\sigma^2}{m_{j'}} = \frac{2\sigma^2}{m}$, if $m_j = m_{j'} = m$ and

(c) $\sum_{i=1}^{m} E[(e_{ij} - \bar{e}_{.j})(\bar{e}_{.j} - \bar{e}_{.j'})] = 0$,

Proof

(a) $E\left[\sum_{i=1}^{m} x_{ij}(e_{ij} - \bar{e}_{.j})\right]^2 = E\left[\sum_{i=1}^{m} x_{ij}^2(e_{ij} - \bar{e}_{.j})^2 + \sum_{i\neq i'i=1}^{m}\sum^{m} x_{ij}(e_{ij} - \bar{e}_{.j})x_{i'j}(e_{i'j} - \bar{e}_{.j})\right]$

$= \left[\sum_{i=1}^{m} x_{ij}^2 E[(e_{ij} - \bar{e}_{.j})^2] + \sum_{i\neq i'i=1}^{m}\sum^{m} x_{ij}x_{i'j}E[(e_{ij} - \bar{e}_{.j})(e_{i'j} - \bar{e}_{.j})]\right]$

$= \left[\sum_{i=1}^{m} x_{ij}^2 E[(e_{ij})^2 - 2\bar{e}_{.j}e_{ij} + \bar{e}_{.j}^2] + 0\right] = \left[\sum_{i=1}^{m} x_{ij}^2 E[e_{ij}^2 - \frac{2e_{ij}\sum_{i=1}^{m} e_{ij}}{m} + \bar{e}_{.j}^2]\right]$

$\sum_{i=1}^{m} x_{ij}^2[\sigma^2 - \frac{2\sigma^2}{m} + \frac{\sigma^2}{m}] = \sum_{i=1}^{m} x_{ij}^2\left(\sigma^2 - \frac{\sigma^2}{m}\right) = \frac{(m-1)\sigma^2}{m}\sum_{i=1}^{m} x_{ij}^2$

$E[(e_{ij} - \bar{e}_{.j})(e_{i'j} - \bar{e}_{.j})] = 0$

(b) $E[(\bar{e}_{.j} - \bar{e}_{.j'})^2] = \frac{\sigma^2}{m_j} + \frac{\sigma^2}{m_{j'}} = \frac{2\sigma^2}{m}$, if $m_j = m_{j'} = m$

$E[(\bar{e}_{.j} - \bar{e}_{.j'})^2] = [E(\bar{e}_{.j}^2) - 2E(\bar{e}_{.j}\bar{e}_{.j'}) + E(\bar{e}_{.j'}^2)]$

$= \frac{\sigma^2}{m_j} + \frac{\sigma^2}{m_j}, \qquad E(\bar{e}_{.j}\bar{e}_{.j'}) = 0$

$= \frac{2\sigma^2}{m}$, if $m_j = m_{j'} = m$

(c) $\sum_{i=1}^{m} x_{ij} E[(e_{ij} - \bar{e}_{.j})(\bar{e}_{.j} - \bar{e}_{.j'})] = 0$,

$\sum_{i=1}^{m} x_{ij} E[(e_{ij} - \bar{e}_{.j})(\bar{e}_{.j} - \bar{e}_{.j'})]$

$= \sum_{i=1}^{m} x_{ij} E[(e_{ij}\bar{e}_{.j} - \bar{e}_{.j}^2) - \bar{e}_{.j'}(e_{ij} - \bar{e}_{.j})]$

$= \sum_{i=1}^{m} x_{ij}[(\frac{\sigma^2}{m} - \frac{\sigma^2}{m}) - E(\bar{e}_{.j'}e_{ij}) + E(\bar{e}_{.j'}\bar{e}_{.j})]$

$= \sum_{i=1}^{m} x_{ij}[\frac{\sigma^2}{m} - \frac{\sigma^2}{m}) - 0 + 0] = 0$

Bibliography

Afonja, B.M. (1975). Introductory Statistics: A Learner's Motivated Approach. *Evans Brothers (Nigeria Publishers), Ibadan*.

Cochran, W. G. (1977). Sampling Techniques. *Third Edition, John Wiley and Sons, New York.*

Frederic, C. Salvador G. and Miguel G. (2012). *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Ltd. DOI: 10.1002/9780470027318.a5610.pub2

Harry F. and Sleoeri, C.A. (1994). *Statistics Concepts and Applications*, Cambridge University Press, Cambridge.

Johnson, R., & Kuby, P. (2007). *Elementary statistics (10th ed.)*. Belmont, CA: Thomson Brooks/Cole.

Larson, R. and Farber, B (2014). Elementary Statistics: Picturing the World, 6th ed. Indianapolis: Pearson Higher Education.

Okafor, F.C. (2002). Sample Survey Theory with Applications. Afro-Orbis Publications Ltd., Nsukka, Nigeria.

S.C Gupta (1984). Fundamentals of statistics: Himalaya publishing house,6th revised & enlarged edition

Stover, C. "Contingency Table." From MathWorld--A Wolfram Web Resource, created by Eric W. Weisstein. http://mathworld.wolfram.com/ContingencyTable.html

Stroock, Daniel (2011). Probability Theory. New York: Cambridge University Press.

Sturges, H. (1926). The choice of a class interval. Journal of the American Statistical Association 21(153), 65-66.

Triola, M. F. (2011). Elementary Statistics, 11th ed. Boston: Addison-Wesley.

Udom, A.U.(2011). Essentials of Statistics, Louis Chumez Printing Enterprises (Nig).

Van Zwet, WR (1979)."Mean, median, mode II". Statistica Neerlandica. 33 (1), 1–5.

ANSWERS TO EXERCISE ONE

1. Population, Sample respectively
2. Secondary data
3. Randomization
4. Descriptive and Inferential
5. Qualitative
6. Pilot Study
7. Variable
8. True
9. Nominal
10.
    i. Categorical
    ii. Categorical
    iii. Quantitative
    iv. Quantitative
    v. Categorical
    vi. Quantitative
11.
    i. Discrete
    ii. Continuous
    iii. Continuous
    iv. Continuous
    v. Discrete
    vi. Discrete
12.
    i. Categorical
    ii. Continuous
    iii. Discrete
    iv. Quantitative
    v. Categorical
    vi. Continuous
13. A
14. D
15. A
16. B
17. C
18. B
19. C
20. A
21. A
22. B
23. C
24. D
25. B
26. B
27. B
28. C
29. A

30. ANSWERS TO EXERCISE TWO

1.

Hours	Frequency	Cumulative Frequency
2	3	3
3	5	8
4	3	11
5	6	17
6	2	19
7	1	20
Total	20	

2.

No. of courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	**0.6**
2	15	**0.3**	**0.9**
3	**5**	**0.1**	**1.0**

3. Percent of students that take exactly two courses = 0.3 x 100 = 30 %
4. Percent of students that take one or two courses = (0.6 + 0.3) x 100 = 90%
5.

No. of flossing per week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.4500	**0.4500**
1	18	**0.3000**	**0.7500**
3	**11**	**0.1833**	0.9333
6	3	0.0500	**0.9833**
7	1	0.0167	**1.0000**

6. 5.00%
7. 93.33%
8.

Classes	Frequency	Class mark	Class boundary
2 -5	10	3.5	1.5 – 5.5
6 – 9	3	7.5	5.5 – 9.5
10 - 13	5	11.5	9.5 – 13.5
14 - 17	4	15.5	13.5 – 17.5
18 - 21	6	19.5	17.5 – 21.5
22 - 25	2	23.5	21.5 – 25.5
Total	30		

9.

Score	Frequency	Relative Frequency	Cumulative Frequency
10	1	0.056	1
11	1	0.056	2
12	1	0.056	3
13	4	0.222	7
14	6	0.333	13
15	4	0.222	17
17	1	0.056	18
Total	18	1	

Relative frequency of 12 is 0.056

10. 13
11. 14

12 (a) 23
(b) 6
(c) 21.7%
(d) 52.1%

13. n = 256
14. (D)
15. (C)
16. (C)
17. 41
18. B
19. A
20. D

Serial Number	Distance (Km)	Frequency	cf	rf	%cf
1	10-14	3	3	0.065	6.522
2	15-19	12	15	0.261	32.609
3	20-24	17	32	0.370	69.565
4	25-29	9	41	0.196	89.130
5	30-34	4	45	0.087	97.826
6	35-39	1	46	0.022	100.000

21. 32
22. 0.022
23. 89.13%
24. (a) 45
(b) 5
25. (a) 40 - 44
(b) 32, 37, 42, 47, 52
26.

Classes	130-135	135-140	140-145	145-150	150-155	155-160	160-165	165-170
Frequency	3	4	4	5	1	1	1	1

The range of heights of the boys is 34 cm

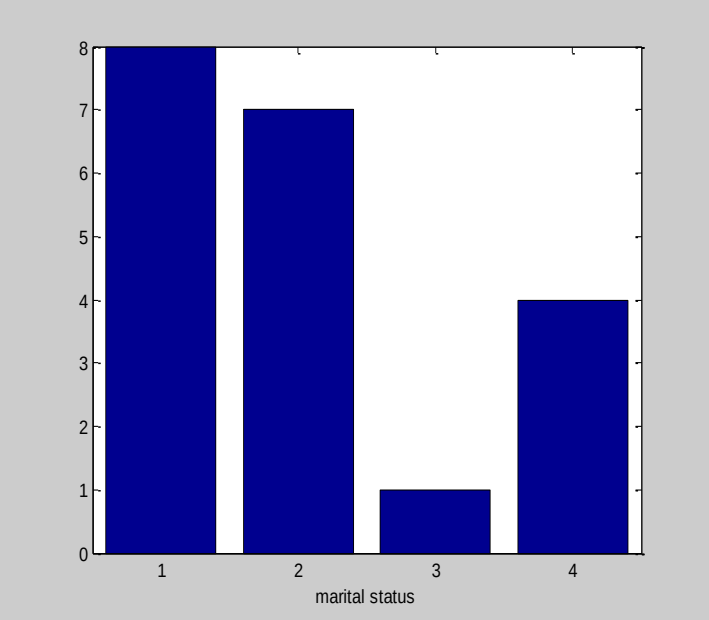
27. Solution to Question Twenty-seven

Classes	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79
Frequency	2	2	3	3	5	3	6	1	4	1

28. 44 gm
29. 10
30. 52
31. 60-64
32. 7
33. 0.1
34. 33.33%

ANSWERS TO EXERCISE THREE

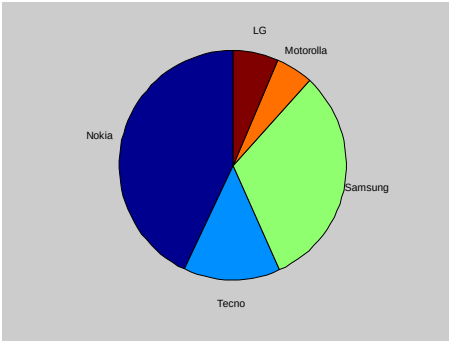
1.



1 - S	8
2 - M	7
3 - W	1
4 - D	4

2. In a component bar chart, the height of each bar is proportional to the sum of frequencies of each item in a category. While in a multiple bar chart, two or more bars lie side by side to each other and the height of each bar is proportional to the frequency of each item in a category.
3. (i) The bars have equal width (ii) There are also equal gaps between the bars.
   (iii) The height of each bar is proportional to the frequency of the item.
4. (i) bar chart (ii) pie chart
5. Frequency
6. (a)

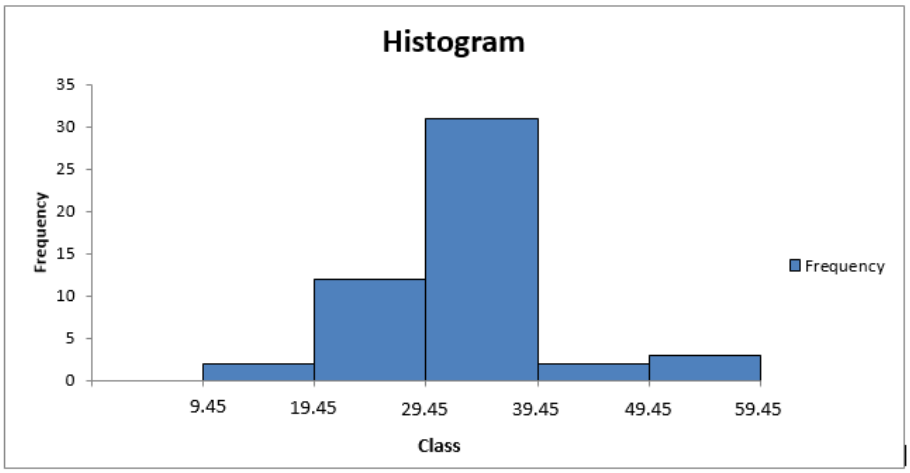
Phone type	frequency	Relative frequency	Sector angle
Nokia	800	800/1860	154.84
Tecno	255	255/1860	49.35
Samsung	586	586/1860	113.42
Motorolla	99	99/1860	19.16
LG	120	120/1860	23.23
Total	1860	1.00	360



(b) 31.51%

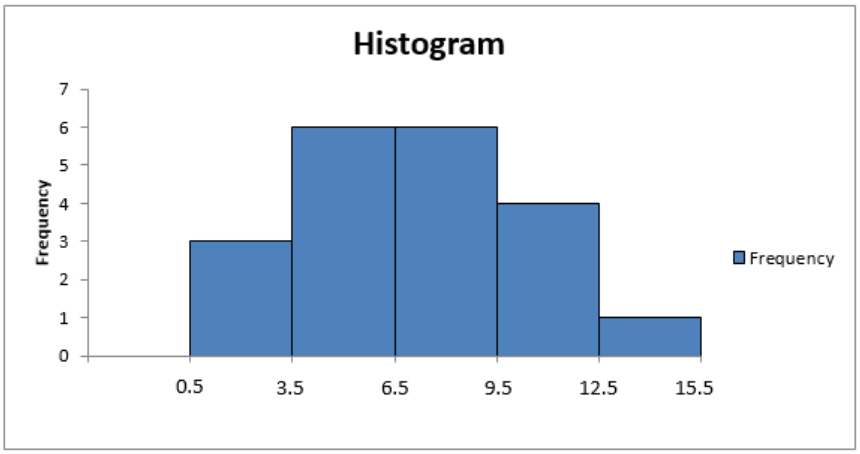
7. Histogram
8. The class boundaries
9. (a) 4 (b) 10 (c) 15

10.



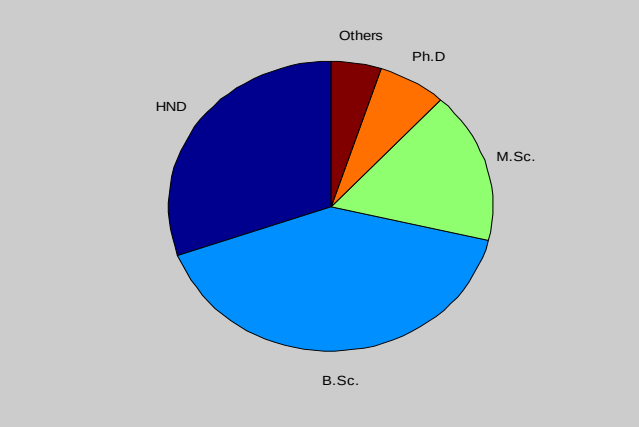
11.

Class	Frequency	Class Boundary
1 – 3	3	0.5 – 3.5
4 – 6	6	3.5 – 6.5
7 – 9	6	6.5 – 9.5
10 - 12	4	9.5 – 12.5
13 - 15	1	12.5 – 15.5



12. Class mark
13. (i) histogram (ii) frequency polygon (iii) stem and leaf plot
(iv) cumulative frequency polygon
14.

Qualification	frequency	Relative frequency	Sector angle
Others	6	5/118	18.30
HND	36	36/118	109.83
B,Sc.	48	48/118	146.44
M.Sc.	20	20/118	61.02
Ph.D	8	8/118	24.41
Total	118	1.00	360



15. In a stem-and-leaf plot, there is no loss of information on individual observations.
16. stem and leaf
17. (a) 1.1 1.3 1.6 2.7 3.5 4.1 4.1 4.4 4.5 4.7 4.7 4.8
    (b) 4.8
18.

0	81 91
1	60 72
2	35 51
3	25 56 70
4	74

0|81 represent 0.81

19.

0	9
1	0 2 5 5 7
2	5 8 8
3	0 0 2

0|9 represent 9

20.

1	7
2	0 2 5
3	5
4	0
5	0
6	2
7	6
8	0
9	4
10	0

2|2 represent 22

21. An Ogive can be used to locate the quartiles and percentiles.

22.

3	9
4	2 5 7
5	0 2 7
6	0 0 3 4
7	0 3 8
8	
9	1
10	0

6|0 represent 60

23.

3	1 3 3 4 5 6 7 8 8 8 9
4	1 5 5

24. 3 4 6 9 10 11 12 15 15 18 21 25 26 27 30

25. 30

26.

Histogram	Bar Chart
1. There is no gap between the bars	There is equal gap between bars
2. Used for quantitative data only	Used for both quantitative and qualitative data

27. B

28. In frequency histogram, we plot frequency versus class boundaries while in relative frequency histogram, relative frequency is plotted on class boundaries.

ANSWERS TO EXERCISE FOUR

1. 81.6
   Median = **81**; Mode = **99**
2. P_{90} = **99**
   P_{80} = **96**
3. (a) Mean = **74;** Mode (most frequent) = **66**
   Median = **72**
   (b) Q1 = **66**
   Q2 = **72**
   Q3 = **81**
4. (c) Variance
5. (c) An extreme value is likely to have a greater effect on the median than the mean
6. (a) Mode (b) Median and Mode (c) Mean, Median and Mode
7. WM = **205.648**
8. GM = **19.681**
9. WM = **3.7**
10. Therefore the median = **4**
11. Median = **40.458**
12. $Q_1 = \frac{1*399}{4} = 99.75,$
    Q_1 = **32.146;** Q_3 = **53.677**
13. D_2 = = **30.447;**
    P_{50} = = **40.458**
14.

Expenditure	No of families	Class Boundaries	Cumulative frequency
0-19	14	0 – 19.5	14
20-39	x	19.5 – 39.5	14 + x
40-59	27	39.5 – 59.5	41 + x
60-79	y	59.5 – 79.5	41 + x + y
80-99	15	79.5 – 99.5	56 + x + y

    x = **22;** y = **22**
15. Mode = **10**
16. $\bar{x} = A + \frac{\sum_{i=1}^{k} f_i d_i}{n}$ = = 31.11
17. The temperature on day 10 is **20**
18. Mean = **29.18;** Mode = **23**
19. **x = 9**
    Mode: = **5**.
    Median: = **5**
20. (a) Given n numbers $x_1, x_2, \ldots\ldots x_n$ having deviations from any number A given respectively by
    $d_1 = x_1 - A, d_2 = x_2 - A, \ldots\ldots d_n = x_n - A$
    $d_j = x_j - A$ and $x_j = A + d_j$ then

$$\bar{x} = \frac{\sum x_j}{n} = \frac{\sum (A + d_j)}{n} = \frac{\sum A + \sum d_j}{n} = \frac{nA}{n} + \frac{\sum d_j}{n} = A + \frac{\sum d_j}{n}$$

(b)(i) $\bar{x} = A + \frac{\sum d}{n} = 9 + \frac{3}{8} = 9.375$

(ii) $\bar{x} = A + \frac{\sum d}{n} = 20 + \frac{(-85)}{8} = 9.375$

21. (a) $\bar{x} = \frac{\sum fx}{\sum f} = \frac{15(1.62) + 20(1.48) + 10(1.53) + 18(1.40)}{15 + 20 + 10 + 18} = 1.50m$

(b) HM = $\frac{7}{\frac{140 + 84 + 70 + 70 + 60 + 42 + 35}{420}} = \frac{2940}{501} = 5.87$

22 (a) GM = 6.43
Arithmetic mean = 7

23. (a) total time = $\frac{2ykm}{0.05yhrs} = 40km/h$

24. 110.9km

25. N79.77

26. Mode = N77.50,

27. Median = N79.06

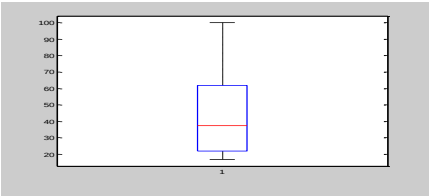
28. (a) 50.03; 50.04, (b) 50.03, 50.04 (c) 49.90; 50.10

29. (a) N7,875.54 (b)N7,683.57 (c)N7,055, N8,560.56

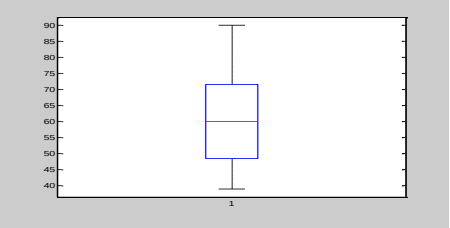
30. Mean ($\bar{x}$) = $\frac{\sum fx}{\sum f} = \frac{99}{50} = 1.98$

Median is 2.
Mode is 0

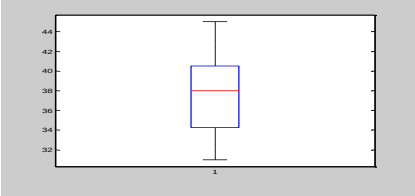
31.



32.



33.



ANSWERS TO EXERCISE FIVE

1.
   i. Standard deviation = 5.9074
   ii. Coefficient of variation = 21.9%
   iii. Interpretation: 1% change in the mean leads to a 21.9% change in the standard deviation.
2. 31.22
3. 5.9074
4. 5.9074
5. 10.38
6. 10.38
7. 10.38
8. 5.9
9. 10.17
10. 4.34
11. 9.03
12. 7.88

    INTERPRETATION: The length of the interval that contains 50% of the data is 7.88
13. **18.8**
    INTERPRETATION: The length of the interval that contains 50% of the data is 18.8
14. For the data in Question 1, the coefficient of variation is 21.9%
    For the data in Question 2, the coefficient of variation is 14.9%
    Therefore, the data in Question 1 is more variable than that of Question 2.
15. For Location A, Z = 0.56; for Location B, Z = -1.30 and for Location C, Z = -1.11
    Therefore, the coolest day was experienced by Location B.
16. 26.16 (by using equation 5.12)
17. 26.16 (by using equation 5.8).
18. Oil Palm = 114.29% and Rubber = 65.22%. Thus, Oil Palm land use is more variable than Rubber land use (by using the coefficient of variation).
19. 16.
20. (82 – 77 = 5) and {(82-77)/2 =2.5} respectively.
21. 15.09.
22. 188,458.65.
23. 64.9%.
24. The salesman's profit is relatively higher in December using medium A since his Z-score for that month is 17.86 against his Z-score of -19.87 using medium B.
25. Variance = 49.71 and Standard Deviation = 7.05.
26. Variance = 50.54 and Standard Deviation = 7.11.
27. 5.23.

28. Variance = 50.54 and Standard Deviation = 7.11.
29. Variance = 50.54 and Standard Deviation = 7.11.

ANSWERS TO EXERCISE SIX

1. (i). The distribution of x is

χ	0	1	2	3	Total
outcome	1	3	3	1	8
$P(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

Where P(x) is the respective probabilities

(i) 45

(ii). P(x = 0) = $\frac{1}{8}$; P(x = 1) = $\frac{3}{8}$; P(x = 2) = $\frac{3}{8}$ and P(x = 3) = $\frac{1}{8}$

(iii). P(x < 2) = P(x = 0 or 1)

= P(x = 0) + P(x = 1)

= $\frac{1}{8} + \frac{3}{8} = \frac{1}{2}$

(iv). P(x > 2) = $\frac{1}{8}$

2. (i) $(P \cup M)^c = 45$

(ii). n(P)= 133 – 95 = 38 students

(iii). n(M) = 117 – 95 = 22 students

(iv). P[n $(P \cup M)^c$] = $\frac{45}{200} = \frac{9}{40}$

3. (a). P(a person picked at random from the gathering wears glasses)

= $\frac{n(E_1)+n(E_2)}{n(S)} = \frac{24+21}{100} = \frac{9}{20}$

(b). Pr(a man picked at random wears glasses) = $\frac{n(E_1)}{n(M)} = \frac{24}{60} = \frac{2}{5}$

(c). Pr(a woman picked at random does not wear glasses) = $1 - \frac{n(E_1)}{n(M)} \Rightarrow 1 - \frac{21}{40} = \frac{19}{40}$

(d). $P(E_1 \cup E_2{}^c) = P(E_1) + P(E_2{}^c)$

$= \frac{2}{5} + \frac{19}{40} = \frac{7}{8}$

4.

$$P(A_2 / B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{P(A_2).P(B/A_2)}{P(B)}$$

$$\therefore P(A_2 / B) = \frac{\frac{1}{3} \times \frac{3}{10}}{\frac{7}{12}} = \frac{1}{10} \times \frac{12}{7} = \frac{6}{35}$$

5. Ans. 1/3

6. Ans. 1/3
7. Ans. 1/7
8. Ans. 1/25
9. Ans. 7/10
10. Ans. 5/18
11. Let n(s) = 24 be the sample space. Let p be the probability that a dishwasher is in good condition.

$P = \frac{24-4}{24} = \frac{20}{24} = \frac{5}{6}$

The number of trial n = 4

$P(X = x) = \sum_{i=1}^{n} \binom{n}{x} p^x q^{n-x}$

$P(X = 4) = \sum_{x=4}^{4} \binom{4}{4} \left(\frac{5}{6}\right)^4 \left(\frac{1}{6}\right)^{4-4} = \left(\frac{5}{6}\right)^4 = \frac{625}{1296} = 0.48$

12. (i) $P(X = 3) = \binom{5}{3} \left(\frac{1}{2}\right)^5 = \frac{5}{16}$

(ii) $P(X \geq 3) = \binom{5}{3} \left(\frac{1}{2}\right)^5 + \binom{5}{4} \left(\frac{1}{2}\right)^5 + \binom{5}{5} \left(\frac{1}{2}\right)^5 = \frac{1}{2}$

13. $P(X = 2) = \binom{3}{2} (0.6)^2 (0.4) = 0.432$

14. (a). P(winning a game) = P(X=3 or 4 or 5)

$= \binom{5}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 + \binom{5}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right) + \left(\frac{1}{6}\right)^5 = 0.0355$

(b). $2.90 per game on the average.

15. a (i). $P(X = x) = \sum_{x=0}^{8} \binom{8}{x} (0.4)^x (0.6)^{8-x}$

$P(X = 3) = \binom{8}{3} (0.4)^3 (0.6)^5 = 0.279$

(ii). $P(X \geq 2) = 1 - P(X < 2)$

$= 1 - P(X = 0 \text{ or } X = 1)$

$= 1 - [(0.6)^8 + 8(0.4)(0.6)^7] = 0.894$

b(i). $E(X) = np = mean$

$= 8(0.4) = 3.2$

(ii). $Var(X) = npq$

$= 8(0.4)(0.6) = 1.92$

16. 245/15625 or 0.01536

17. 0.26

18. 0.04

19. 0.23

20. (a) 0
(b) 1
21. 0.5768
22. 0.35
23. 0.92
24. (a) 0.4868 (b) 0.3505 (c) 0.1262 (d) 0.0303
25. (a) 3 (b) 0.22404
26. (a) 0.1755 (b) 0.4972
27. (a) 0.7922 (b) (i) 0.001670 (ii) 0.001776
28. (a) 0.50113 (b) 0.9744
29. (a) 0.8009 (b) 188.2115 = 188
30. (a) x = 0, 1, 2, . . . (b). $2.6486e^{-16}$
31. 52.06kg
32. 1.736
33. 0.781
34. (i) 0.214 (ii) 0.5899 (iii) 0.0918
35. 0.0668
36. 0.6247
37. 16.5
38. (i). 0.925 (ii). 0.9772
39. (i). 0.8413 (ii). 0.628

40. Solution

Let $P(B_i)$ be the probability of selecting box i; i = 1, 2, 3, 4.
Let P(D/B) be the probability that the bead selected from box i is defective.
Number of defective beads in

Box1 $= \frac{4}{100} \times 400 = 16$

Box2 $= \frac{6}{100} \times 120 = \frac{36}{5}$

Box3 = Box4 $= \frac{5}{100} \times 500 = 25$

$$\therefore P(B_1) = P(B_2) = P(B_3) = P(B_4) = \frac{1}{4}$$

$$\therefore P(D/B_1) = \frac{16}{400} = \frac{1}{25}$$

$$\therefore P(D/B_2) = \frac{\left(\frac{36}{5}\right)}{120} = \frac{3}{50}$$

$$\therefore P(D/B_3) = \frac{25}{500} = \frac{1}{20}$$

$$\therefore P(D/B_4) = \frac{25}{500} = \frac{1}{20}$$

Let P(D) denote the probability of choosing a defective bead.

$$P(D) = P(B_1)P(D/B_1) + P(B_2)P(D/B_2) + P(B_3)P(D/B_3) + P(B_4)P(D/B_4)$$

$$\therefore P(D) = \frac{1}{4} \times \frac{1}{25} + \frac{1}{4} \times \frac{3}{50} + \frac{1}{4} \times \frac{1}{20} + \frac{1}{4} \times \frac{1}{20}$$

$$\therefore P(D) = \frac{1}{4}\left(\frac{1}{25} + \frac{3}{50} + \frac{1}{20} + \frac{1}{20}\right) = \frac{1}{20}$$

ANSWERS TO EXERCISE SEVEN

1) In statistics, **estimation** (or inference) refers to the process by which one makes inferences (e.g. draws conclusions) about a population, based on information obtained from a sample.

2) (i) The point estimate of the sample mean $\overline{X}$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

(ii) A **point estimate** of a population parameter is a single value of a statistic (e.g. the average height). This in general changes with the selected sample while An **interval estimate** is generally associated with a confidence level. Suppose we collected many different samples (with the same sampling strategy) and computed confidence intervals for each of them. Some of the confidence intervals would include the population parameter, others would not. A 95% confidence level means that 95% of the intervals contain the population parameter

3) (i) A **population parameter** is a quantity usually estimated from the sample realization of an estimator while An **estimate** is the particular value of an estimator that is obtained by a particular sample of data and used to indicate the value of a parameter.
(ii) An **estimator** is any quantity calculated from the sample data which is used to give information about an unknown quantity in the population (the estimand).

4) The 95% confidence interval for

$$\mu = 52.4 \pm 2.262\left(\frac{470.9333}{\sqrt{10}}\right) = 52.4 \pm 2.262(148.9222) = 52.4 \pm 336.862 = \left[-284.462, \quad ,389.262\right]$$

5) The 95% confidence interval for

$$\mu = 1.5 \pm 3.25\left(\frac{0.6133}{\sqrt{10}}\right) = 1.5 \pm 3.25(0.193953) = 1.5 \pm 0.630347 = \left[0.869653, \quad ,2.130347\right]$$

6) The 95% confidence interval is:

$$16 \pm 1.96\left(\frac{3}{\sqrt{100}}\right) = 16 \pm 0.588 = \left[15.412, \quad 16.588\right]$$

7) We need a sample size of $n \geq \left(\frac{(1.96)(3)}{0.5}\right)^2 = 138.3$

That is we need a sample size of at least 139.

8) The mean assembly time for a worker is estimated to be between 15.6 min and 16.8 min, with 92% confidence.
b) 636 workers are need in the study to achieve the desired precision of inference

9) The desired confidence interval is

$$37.7 \pm (1.645)\frac{9.2}{\sqrt{100}} = 37.7 \pm 1.5 = [36.2, 39.2].$$

10) The mean weight is estimated to be between 0.94 to 1.04 grams, with 95% confidence.

11) The critical point in this case is $t_{n-1,\alpha/2}$

12) The mean installation time is estimated to be between 40.8 min and 43.2 min, with 95% confidence

13) The mean installation time is estimated to be between 42.361 secs and 47.639 secs, with 95% confidence

14) A **point estimate** of a population parameter is a single value of a statistic (e.g. the average height). This in general changes with the selected sample while An **interval estimate** is generally associated with a confidence level. Suppose we collected many different samples (with the same sampling strategy) and computed confidence intervals for each of them. Some of the confidence intervals would include the population parameter, others would not.

15) A 95% confidence level means that 95% of the intervals contain the population parameter.

16) Depending on the assumptions, n should be chosen such that

$$n \geq \left(\frac{Z_{\alpha/2}\sigma}{w}\right)^2$$

or

$$n \geq \left(\frac{Z_{\alpha/2}S}{w}\right)^2$$

or

$$n \geq \left(\frac{t_{(n-1)\alpha/2}S}{w}\right)^2$$

17) A minimum of 107 homes must be sampled.

18) The 99% confidence interval for the mean breaking strength of all the iron rods produced by the steel complex is (34.3058, 34.6942).

19) $1-\alpha$ is called confidence coefficient or degree of confidence.

20) The interval is called the $100(1-\alpha)\%$ confidence interval.

21) (i) Two conditions includes: Small sample size (n < 30) and unknown variance.
(ii) $t_{(n_1+n_2-2),\frac{\alpha}{2}} = t_{(9+16-2)0.05/2} = t_{(23),0.025} = 2.069$ {please refer to t-distribution table}

22) (i) Our objective is to design an interval estimator such that (a) the confidence level sufficiently high and (b) the margin of error sufficiently small

(ii) An interval estimator is usually designed in the following way:
(a) take an unbiased point estimator, and
(b) define an interval of reasonable width around it.

ANSWERS TO EXERCISE EIGHT

1. C
2. A
3. D
4. B
5. A
6. D
7. D
8. C
9. B
10. B
11. A
12. B
13. C
14. A
15. B
16. D
17 A
18. A
20. A
21. C
22. B
23. B
24. A
25. B
26. Test the hypotheses for the problems in questions 25 and 26.

ANSWERS TO EXERCISE NINE

(1) Calculation of rank correlation coefficient

X	Y	R_X	R_Y	$D=(R_X-R_Y)$	D^2
50	110	2	1.5	0.5	0.25
55	110	4.50	1.50	3.00	9.00
65	115	7.50	4.00	3.50	12.25
50	125	2.00	7.00	-5.00	25.00
55	140	4.50	9.00	-4.50	20.25
60	115	6.00	4.00	2.00	4.00
50	130	2.00	8.00	-6.00	36.00
65	120	7.50	6.00	1.50	2.25
70	115	9.00	4.00	5.00	25.00
75	160	10.00	10.00	0.00	0.00

$$\sum D^2 = 134.00$$

$$r = 0.155$$

(2) Calculation of rank correlation coefficient

LIPSTICKS	CHIOMA (R_X)	NKECHI (R_Y)	$D=(R_X-R_Y)$	D^2
A	2	1	1	1
B	1	3	-2	4
C	4	2	2	4
D	3	4	-1	1
E	5	5	0	0
F	7	6	1	1
G	6	7	-1	1

$$\sum D^2 = 12$$

$$r = 1 - \frac{6 \times 12}{7^3 - 7} = 0.786$$

(3) Calculation of rank correlation coefficient

Year	Debenture Price	Share Price	R_X	R_Y	$D=(R_X-R_Y)$	D^2
1	97.8	73.2	3	1	2	4
2	99.2	85.8	7	6	1	1
3	98.8	78.9	6	4	2	4
4	98.3	75.8	4	2	2	4
5	98.4	77.2	5	3	2	4
6	96.7	87.2	1	7	-6	36
7	97.1	83.8	2	5	-3	9

$$\sum D^2 = 62$$

$$r = 1 - \frac{6 \times 62}{7^3 - 7} = -0.107$$

(4) Calculation of rank correlation coefficient

PRICE OF TEA	PRICE OF COFFEE	R_X	R_Y	$D = (R_X - R_Y)$	D^2
75	120	4	4	0	0
88	134	7	5	2	4
95	150	8	8	0	0
70	115	3	3	0	0
60	110	2	2	0	0
80	140	5	6	-1	1
81	142	6	7	-1	1
50	100	1	1	0	0
					$\sum D^2 = 6$

$$r = 1 - \frac{36}{512 - 8} = 0.929$$

(5) Calculation of rank correlation coefficient

Laboratory (R_X)	Lecture (R_Y)	$D = (R_X - R_Y)$	D^2
8	9	-1	1
3	5	-2	4
9	10	-1	1
2	1	1	1
7	8	-1	1
10	7	3	9
4	3	1	1
6	4	2	4
1	2	-1	1
5	6	-1	1
			$\sum D^2 = 24$

$$r = 1 - \frac{6 \times 24}{10^3 - 10} = 0.8545$$

Which indicates that there is a marked relationship between achievements in laboratory and lecture

(6) Let X and Y denote the average prices of stocks and bonds, using deviation from the mean method we have

Year	Average Price of Stocks (X)	Average price of bonds (Y)	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
2000	35.22	102.43	-10.04	4.91	100.80	24.11	-49.30
2001	39.87	100.93	-5.39	3.41	29.05	11.63	-18.38
2002	41.85	97.43	-3.41	-0.09	11.63	0.01	0.31
2003	43.23	97.81	-2.03	0.29	4.12	0.08	-0.59
2004	40.06	98.32	-5.20	0.80	27.04	0.64	-4.16

2005	53.29	100.07	8.03	2.55	64.48	6.50	20.48
2006	54.14	97.08	8.88	-0.44	78.85	0.19	-3.91
2007	49.12	91.59	3.86	-5.93	14.90	35.16	-22.89
2008	40.71	94.85	-4.55	-2.67	20.70	7.13	12.15
2009	55.15	94.65	9.89	-2.87	97.81	8.24	-28.38

$\sum X = 452.64, \quad \sum Y = 975.16, \quad \sum x^2 = 449.38, \quad \sum y^2 = 93.69, \quad \sum xy = -94.67$

$\bar{X} = 45.26, \quad \bar{Y} = 97.52$

$$r = \frac{\sum xy}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}} = \frac{-94.67}{\sqrt{(449.38)(93.69)}} = -0.4614$$

We conclude that there is some negative correlation between stock and bond prices, although this relationship is not marked.

(7) Calculate the correlation coefficient

GRADE ON FIRST QUIZ (X)	GRADE ON SECOND QUIZ (Y)	X^2	Y^2	XY
6	8	36	64	48
5	7	25	49	35
8	7	64	49	56
8	10	64	100	80
7	5	49	25	35
6	8	36	64	48
10	10	100	100	100
4	6	16	36	24
9	8	81	64	72
7	6	49	36	42

$\sum X = 70, \quad \sum Y = 75, \quad \sum X^2 = 520, \quad \sum Y^2 = 587, \quad \sum XY = 540$

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} = 0.5533$$

(8)

AGE (X)	BLOOD PRESSURE (Y)	X^2	Y^2	XY
56	147	3136	21609	8232
42	125	1764	15625	5250
72	160	5184	25600	11520
36	118	1296	13924	4248
63	149	3969	22201	9387
47	128	2209	16384	6016
55	150	3025	22500	8250
49	145	2401	21025	7105
38	115	1444	13225	4370
42	140	1764	19600	5880
68	152	4624	23104	10336
60	155	3600	24025	9300

$\sum X = 628, \quad \sum Y = 1684, \quad \sum X^2 = 34416, \quad \sum Y^2 = 238822, \quad \sum XY = 89894$

iv. $r = 0.8961$

v. The least square regression line of Y on X is

$$Y = 80.78 + 1.138X$$

vi. 132

(9) Based on the table in question 7 above,

ii. $Y = 4.0 + 0.50X$

iii. $X = 2.408 + 0.612Y$

(10) Calculation of rank correlation coefficient

R_A	R_B	$D = (R_A - R_B)$	D^2
1	10	-9	81
2	7	-5	25
3	2	1	1
4	6	-2	4
5	4	1	1
6	8	-2	4
7	3	4	16
10	1	9	81
9	11	-2	4
10	15	-5	25
11	9	2	4
12	5	7	49
13	14	-1	1
14	12	2	4
15	13	2	4
			$\sum D^2 = 304$

$$r = 1 - \frac{6 \times 304}{15^3 - 15} = 0.457$$

(11) Calculate the regression line

X	Y	X^2	Y^2	XY
0.5	10	0.25	100	5
1.0	12	1.00	144	12
1.5	14	2.25	196	21
2.0	16	4.00	256	32
2.5	17	6.25	289	42.5
3.0	18	9.00	324	54
3.5	21	12.25	441	73.5
14	108	35	1750	240

$$\sum X = 14, \quad \sum Y = 108, \quad \sum X^2 = 35, \quad \sum Y^2 = 1750, \quad \sum XY = 240$$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{7 \times 240 - 108 \times 14}{7 \times 35 - 14^2} = 3.4$$

$$a = \bar{Y} - b\bar{X} = 8.6$$

Regression line of Y on X is $Y = 8.6 + 3.4X$

(12)

X	Y	X^2	Y^2	XY
1.2	4.5	1.44	20.25	5.4
1.8	5.9	3.24	34.81	10.62
3.1	7	9.61	49	21.7
4.9	7.8	24.01	60.84	38.22
5.7	7.2	32.49	51.84	41.04
7.1	6.8	50.41	46.24	48.28
8.6	4.5	73.96	20.25	38.7

9.8	2.7	96.04	7.29	26.46
42.2	46.4	291.2	290.52	230.42

$\sum X = 42.2, \quad \sum Y = 46.4, \quad \sum X^2 = 291.2, \quad \sum Y^2 = 290.52, \quad \sum XY = 230.42$

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} = -0.3743$$

(13)

Candidate	JUDGE 1 (R_A)	JUDGE 2 (R_B)	$D = (R_A - R_B)$	D^2
A	5	4	1	1
B	2	5	-3	9
C	8	7	1	1
D	1	3	-2	4
E	4	2	2	4
F	6	8	-2	4
G	3	1	2	4
H	7	6	1	1
				$\sum D^2 = 28$

$$r = \frac{2}{3}$$

(14) Calculate the correlation coefficient

STUDENT	X	Y	X^2	Y^2	XY
1	97	89	9409	7921	8633
2	68	57	4624	3249	3876
3	85	87	7225	7569	7395
4	74	76	5476	5776	5624
5	92	97	8464	9409	8924
6	92	79	8464	6241	7268
7	100	91	10000	8281	9100
8	63	50	3969	2500	3150
9	85	85	7225	7225	7225
10	87	84	7569	7056	7308
11	81	91	6561	8281	7371
12	93	91	8649	8281	8463
13	77	75	5929	5625	5775
14	82	77	6724	5929	6314

$\sum X = 1176, \quad \sum Y = 1129, \quad \sum X^2 = 100288, \quad \sum Y^2 = 93343, \quad \sum XY = 96426$

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} = 0.86$$

(15) Calculate the correlation coefficient

STUDENT	X	Y	X^2	Y^2	XY
1	70	2.5	4900	6.25	175
2	90	4	8100	16	360
3	75	3.5	5625	12.25	262.5
4	85	3	7225	9	255
5	80	3	6400	9	240
6	70	2	4900	4	140
7	90	3	8100	9	270

$\sum X = 560, \quad \sum Y = 21, \quad \sum X^2 = 45250, \quad \sum Y^2 = 65.5, \quad \sum XY = 1702.5$

$$r=\frac{n\sum XY-\left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2-\left(\sum X\right)^2\right]\left[n\sum Y^2-\left(\sum Y\right)^2\right]}}=0.67$$

ii. $b=\frac{n\sum XY-\sum X\sum Y}{n\sum X^2-\left(\sum X\right)^2}=0.05$

$a=\bar{Y}-b\bar{X}=-1.0$

Regression line of Y on X is $Y=-1.0+0.05X$

(16) Calculate the correlation coefficient

X	Y	X^2	Y^2	XY
5	25	25	625	125
3	20	9	400	60
4	21	16	441	84
10	35	100	1225	350
15	38	225	1444	570

$\sum X=37,\quad \sum Y=139,\quad \sum X^2=375,\quad \sum Y^2=4135,\quad \sum XY=1189$

$$r=\frac{n\sum XY-\left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2-\left(\sum X\right)^2\right]\left[n\sum Y^2-\left(\sum Y\right)^2\right]}}=0.97$$

(17) Calculate the correlation coefficient

X	Y	X^2	Y^2	XY
1	95	1	9025	95
0	90	0	8100	0
2	90	4	8100	180
6	55	36	3025	330
4	70	16	4900	280
3	80	9	6400	240
3	85	9	7225	255

$\sum X=19,\quad \sum Y=565,\quad \sum X^2=75,\quad \sum Y^2=46775,\quad \sum XY=1380$

$$r=\frac{n\sum XY-\left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2-\left(\sum X\right)^2\right]\left[n\sum Y^2-\left(\sum Y\right)^2\right]}}=-0.93$$

(18) Calculate the correlation coefficient

X	Y	X^2	Y^2	XY
68	90	4624	8100	6120
72	85	5184	7225	6120
65	88	4225	7744	5720
70	100	4900	10000	7000
62	105	3844	11025	6510
75	98	5625	9604	7350
78	70	6084	4900	5460
64	65	4096	4225	4160
68	72	4624	5184	4896

$\sum X=622,\quad \sum Y=773,\quad \sum X^2=43206,\quad \sum Y^2=68007,\quad \sum XY=53336$

$$r=\frac{n\sum XY-\left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2-\left(\sum X\right)^2\right]\left[n\sum Y^2-\left(\sum Y\right)^2\right]}}=-0.15$$

(19) Calculate the regression line

X	Y	X^2	Y^2	XY
14	2	196	4	28
20	3	400	9	60
32	5	1024	25	160
42	7	1764	49	294
44	8	1936	64	352

$\sum X=152,\quad \sum Y=25,\quad \sum X^2=5320,\quad \sum Y^2=151,\quad \sum XY=894$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = 0.19$$

$$a = \bar{Y} - b\bar{X} = -0.83$$

Regression line of Y on X is $Y = -0.83 + 0.19X$

To get the approximate weight of a six year old child, then we get regression line of X on Y which is

$$X = 4.63 + 5.15Y$$

Therefore, the approximate weight when Y = 6 is 35.55

(20) Calculate the correlation coefficient

X	Y	X^2	Y^2	XY
8	15	64	225	120
7	19	49	361	133
6	25	36	625	150
4	23	16	529	92
2	34	4	1156	68
1	40	1	1600	40

$\sum X = 28, \quad \sum Y = 156, \quad \sum X^2 = 170, \quad \sum Y^2 = 4496, \quad \sum XY = 603$

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} = -0.95$$

ii. $X = 12.05 + -0.28Y$

if Y = 2, then X = 11.48

iii $Y = 40.83 - 3.18X$

if X = 5, then Y = 24.94

(21) Calculate the regression line

X	Y	X^2	Y^2	XY
6	6.5	36	42.25	39
4	4.5	16	20.25	18
8	7	64	49	56
5	5	25	25	25
3.5	4	12.25	16	14

$\sum X = 26.5, \quad \sum Y = 27, \quad \sum X^2 = 153.25, \quad \sum Y^2 = 152.5, \quad \sum XY = 152$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = 0.70$$

$$a = \bar{Y} - b\bar{X} = 1.71$$

Regression line of Y on X is $Y = 1.71 + 0.70X$

When X = 7.5, then Y = 6.93

(22) Calculate the correlation coefficient

X	Y	X^2	Y^2	XY
186	85	34596	7225	15810
189	85	35721	7225	16065
190	86	36100	7396	16340
192	90	36864	8100	17280
193	87	37249	7569	16791
193	91	37249	8281	17563
198	93	39204	8649	18414
201	103	40401	10609	20703
203	100	41209	10000	20300
205	101	42025	10201	20705

$\sum X = 1950, \quad \sum Y = 921, \quad \sum X^2 = 380618, \quad \sum Y^2 = 85255, \quad \sum XY = 179971$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = 1.02$$

$a = \bar{Y} - b\bar{X} = -107.14$

Regression line of Y on X is $Y = -107.14 + 1.02X$

When X = 208, then Y = 105.38

$$r = \frac{n\sum XY}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} = 0.94$$

(23) Calculate the correlation coefficient

X	Y	X^2	Y^2	XY
80	300	6400	90000	24000
79	302	6241	91204	23858
83	315	6889	99225	26145
84	330	7056	108900	27720
78	300	6084	90000	23400
60	250	3600	62500	15000
82	300	6724	90000	24600
85	340	7225	115600	28900
79	315	6241	99225	24885
84	330	7056	108900	27720
80	310	6400	96100	24800
62	240	3844	57600	14880

$\sum X = 936, \quad \sum Y = 3632, \quad \sum X^2 = 73760, \quad \sum Y^2 = 1109254, \quad \sum XY = 285908$

$$b = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - \left(\sum X\right)^2} = 3.47$$

$a = \bar{Y} - b\bar{X} = 31.74$

Regression line of Y on X is $Y = 31.74 + 3.47X$

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} = 0.95$$

(24) Calculate the correlation coefficient

X	Y	X^2	Y^2	XY
25	42	625	1764	1050
42	72	1764	5184	3024
33	50	1089	2500	1650
54	90	2916	8100	4860
29	45	841	2025	1305
36	48	1296	2304	1728

$\sum X = 219, \quad \sum Y = 347, \quad \sum X^2 = 8531, \quad \sum Y^2 = 21877, \quad \sum XY = 13617$

$$b = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - \left(\sum X\right)^2} = 1.77$$

$a = \bar{Y} - b\bar{X} = -6.78$

Regression line of Y on X is $Y = -6.78 + 1.77X$

When X = 47, then Y = 76.42

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} = 0.96$$

(25) Calculate the correlation coefficient

X	Y	X^2	Y^2	XY
3	4	9	16	12
4	3	16	9	12
5	2	25	4	10
6	1	36	1	6

$\sum X = 18, \quad \sum Y = 10, \quad \sum X^2 = 86, \quad \sum Y^2 = 30, \quad \sum XY = 40$

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}} = -1.00$$

ANSWERS TO EXERCISE TEN

1. 8.5
2. 5.75
3. 4.5
4. 558
5. 150
6. 7.35
7. The test used to compare three or more means is F test
8. Comparing two means at a time ignores all other means.
   The probability of type I error is larger than α when multiple t tests are used. The more sample means, the more t tests are used.
9. The populations from which the samples were obtained must be normal or approximately normally distributed.
   The samples must be independent of each other.
   The variance of the populations must be equal.
10. The F test formula for comparing three or more means is a ratio of the mean square treatment to mean square error and is given as:
    $$F = \frac{MST}{MSE}.$$
11. The hypotheses used in the ANOVA test are
    $$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_p$$
    H_1 : at least two of the means are not equal
12. The means of diet A =5, B = 10.17, and C = 4.
13. 3.33
14. 8.17
15. 7
16. 2
17. 11
18. 7.07
19. 29.18
20. 6.53
21. To answer the question of whether the researcher can conclude that there is a difference in the diets? We first determine the critical value for the test, i.e., $F_{0.05,2,11} = 8.91$. Since 29.18 > 8.91, the researcher can conclude that there is a difference in the diets.
22. $H_0 : \mu_{1996} = \mu_{1997} = \mu_{1998}$
    H_1 : At least one mean is different from the others.(Claim)
23. 6.586282 X 10^{10}
24. 14
25. 5

26. 30
27. 37.7
28. 4.62
29. 4.62
30. $SSC = p\sum_{j=1}^{q} \overline{X}_{.j}^{2} - \frac{G^2}{n}$

ANSWERS TO EXERCISE ELEVEN

(1) χ^2_{cal} = 36.4524, the age actually has effect on the number of accidents.

(2) χ^2_{cal} = 252.89, there is an association between cigarette smoking and lung cancer.

(3) χ^2_{cal} = 2.38. Recovery is independent of the drug.

(4) χ^2_{cal} = 1474.07, HIV is dependent of exposure to sex.

(5) χ^2_{cal} = 3.0

(6) χ^2_{cal} = 0.9569, the sale of local drink does not depend on whether or an individual drinks.

(7) χ^2_{cal} = 15.2387

(8) χ^2_{cal} = 2.38.

(9) The hypothesis cannot be rejected.

(10) The hypothesis cannot be rejected.

(11) χ^2_{cal} = 2.57

(12) χ^2_{cal} = 8.88

(13) χ^2_{cal} = 2.66

(14) χ^2_{cal} = 2.097

(15) χ^2_{cal} = 57.51

(16) -

(17) χ^2_{cal} = 0.666

(18) χ^2_{cal} = 17.78

(19) χ^2_{cal} = 4.63

(20) χ^2_{cal} = 2.67

(21) χ^2_{cal} = 8.224

(22) χ^2_{cal} = 20.59

(23) χ^2_{cal} = 16.402

(24) χ^2_{cal} = 6.73

(25) χ^2_{cal} = 8.23

(26) In the population the distribution of preferred factors for men has the same proportions as the distribution for women.

(27) We do not reject the hypothesis.

(28) χ^2_{cal} = 2.38

(29) C

(30) B

ANSWERS TO EXERCISE TWELVE

(1) GDP (in x 10^{10} ₦) and PCI (in ₦)

	Year				
	2010	2011	2012	2013	2014
(a) GDP	54.61	57.51	59.93	63.22	67.16
(b) PCI	3462.81	3542.46	3586.01	3674.75	3792.18

(2) Dependency ratios for four countries in the same year

	Country			
	A	B	C	D
DR (%)	103.2	121.48	43.34	30.75

(b) Dependency burden appears highest in country B and lowest in country D

(3) Table 12.3: Distn of total popn, number dead and number of Births in a country

	Population	Death	Birth
SR (%)	104.62	137.58	106.41
CBR			20.43
CDR		4.48	

(4) a(i) Crude Marriage Rate (CMR) for the males ever married is

$$\text{CMR} = \frac{23417936}{49387659} \times 100 = 47.42\%$$

This indicates that out of every 100 males reported as aged 10 years and above, about 47 (47.42) were reported as ever-married.

a(ii) Crude Marriage Rate (CMR) for the females ever married is

$$\text{CMR} = \frac{28975834}{48443784} \times 100 = 59.81\%$$

This indicates that out of every 100 females reported as aged 10 years and above, about 60 (59.81) were reported as ever-married.

b(i) Crude Divorce Rate (CDR) for males is

$$CRD = \frac{229627}{49387659}\ x\ 100 = 0.46\%$$

This indicates that out of every 100 persons (males) reported as aged 10 years and above, about 0.00 (0.46) was reported as divorced.

b(ii) Crude Divorce Rate (CDR) for females is

$$CRD = \frac{473987}{48443784}\ x\ 100 = 0.98\%$$

This indicates that out of every 100 persons (females) reported as aged 10 years and above, about 1.00 (0.98) was reported as divorced.

(5) $$Q_{01} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100 = 118.42$$

(6) $$P_{01} = \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = 112.8$$

(7)

Year	**Price of Wheat**	**Index nos (2004 =100)**
2004	50	100
2005	60	$\frac{60}{50} \times 100 = 120$
2006	62	$\frac{62}{50} \times 100 = 124$
2007	65	$\frac{65}{50} \times 100 = 130$
2008	70	$\frac{70}{50} \times 100 = 140$
2009	78	156
2010	82	164
2011	84	168
2012	88	176
2013	90	180

It should be noted that from 2004 to 2005, there is a 20% increase. From 2005 to 2006, there is a 24% increase etc

(8) Computation of chain indices

Commodity	2009	2010	2011	2012	2013
A	100	$\frac{16}{20}\times 100 = 80$	$\frac{28}{16}\times 100 = 175$	$\frac{35}{28}\times 100 = 125$	$\frac{21}{35}\times 100 = 60$
B	100	$\frac{30}{25}\times 100 = 120$	$\frac{24}{30}\times 100 = 80$	$\frac{36}{24}\times 100 = 150$	$\frac{45}{36}\times 100 = 125$
C	100	$\frac{25}{20}\times 100 = 125$	$\frac{30}{25}\times 100 = 120$	$\frac{24}{30}\times 100 = 80$	$\frac{30}{24}\times 100 = 125$
Total of link relatives	300	325	375	355	310
Average of link relatives	100	108.33	125	118.33	103.33
Chain index 2009 = 100	100	$\frac{108.33}{100}\times 100 = 108.33$	$\frac{125}{100}\times 108.33 = 135.41$	$\frac{108.33}{100}\times 100 = 108.33$	$\frac{103.33}{100}\times 160{,}23 = 165.57$

(9) Conversion of FBI to CBI

Year	FBI	Link Relatives	Chain Index
2000	376	---	376
2001	392	$\frac{392}{376}\times 100 = 104.26$	$\frac{376\times 104.26}{100} = 392$
2002	408	$\frac{408}{392}\times 100 = 104.08$	$\frac{392\times 104.08}{100} = 408$
2003	380	$\frac{380}{408}\times 100 = 93.14$	$\frac{408\times 93.14}{100} = 380$
2004	392	$\frac{392}{380}\times 100 = 103.16$	$\frac{380\times 103.16}{100} = 392$
2005	400	$\frac{400}{392}\times 100 = 102.04$	$\frac{392\times 102.04}{100} = 400$

It should be noted that the CBI are the same as FBI. In fact, this will always be true for a single fixed series of index numbers.

(10) $CPI = \frac{\sum p_1 q_0}{\sum p_0 q_0}\times 100 = \frac{174}{146.50}\times 100 = 118.77$

(11) $CPI = \frac{\sum PV}{\sum V} = \frac{17400}{146.50} = 118.77$