CHARTERED INSTITUTE OF PERSONNEL MANAGEMENT OF NIGERIA

STUDY PACK

ON

BUSINESS STATISTICS AND SOCIAL RESEARCH

INTERMEDIATE I

BUSINESS STATISTICS AND SOCIAL RESEARCH

INTERMEDIATE I

FOURTH EDITION
CHARTERED INSTITUTE OF PERSONNEL
MANAGEMENT OF NIGERIA

CIPM House, 1 CIPM Avenue, Off Obafemi Awolowo Way,
Opposite Lagos State Secretariat, Alausa, Ikeja, Lagos.
Tel: 07001237555
E-mail: info@cipmnigeria.org, exams@cipmnigeria.org
Website: www.cipmnigeria.org
Facebook/X/Instagram/Youtube: @cipmnigeria
LinkedIn: Chartered Institute of Personnel Management of Nigeria

FOREWORD

This fourth edition of the CIPM study pack is one of the learning resources recommended to persons preparing for certification through professional examinations. It is uniquely prepared to meet the knowledge standards of HR certification bodies and/or degree awarding institutions. The study pack is highly recommended to researchers, people managers and organisations responsible for human capital development in its entirety.

Each chapter in the text has been logically arranged to sufficiently cover all the various sections of this subject as itemised in the CIPM examination syllabus. This is to enhance systematic learning and understanding of the users. The document, a product of in-depth study and research, is practical and original. We have ensured that topics and sub-topics are based on the syllabus and on contemporary HR best practices.

Although concerted effort has been made to ensure that the text is up to date in matters relating to theories and practices of contemporary issues in HR, nevertheless, we advise and encourage students to complement the study text with other study materials recommended in the syllabus. This is to ensure total coverage of the elastic scope and dynamics of the HR profession.

Thank you and do have a productive preparation as you navigate through the process of becoming a seasoned Human Resources Management professional.

Olusegun Mojeed, FCIPM, fnli
President & Chairman of the Governing Council

ACKNOWLEDGEMENTS

On behalf of the President and Chairman of the Governing Council, Mr Olusegun Mojeed, FCIPM, fnli and the entire membership of the Chartered Institute of Personnel Management of Nigeria (CIPM), we acknowledge the intellectual prowess of Dr. Bolanle Azeez Oseni in writing this well-researched text for Business Statistics and Social Research Methods. The meticulous work of our reviewer, Dr. Sanni Ladan has not gone unnoticed and is hereby acknowledged for the thorough review of this publication.

We also commend and appreciate the efforts of members of the Education Committee of the Institute, under the chairmanship of Mr. Henry Onukwuba, MCIPM for their unflinching support.

Finally, my appreciation goes to my internal project team, led by the Director, Professional Standards and Development, Mr. Gbenga Samuel Odetunde ACIPM, the Team lead, Certification and Licensing, Mr. Samuel Eviewho, ACIPM and Team lead, Professional Examination, Mr. Yinka Oyedere, MCIPM for making this project a success.

Oluwatoyin Naiwo, FCIPM
Registrar/Chief Executive

TABLE OF CONTENTS

FOREWORD iii

ACKNOWLEDGEMENT iv

CHAPTER ONE 2

THE NATURE OF BUSINESS STATISTICS 2

1.0 Learning Objectives 2

1.1 Introduction 2

1.2 Nature of Business Statistics. 2

1.3 Uses and Importance of Statistics in Human Resource Management 3

CHAPTER TWO 5

STATISTICAL DATA 5

2.0 Learning Objectives 5

2.1 Introduction 5

2.2 Type of Statistical Data 5

2.3 Methods of Data collection 6

2.3.1 Direct personal or Interview method 6

2.3.2 Indirect Oral Interview 7

2.3.3 Information through correspondents 7

2.3.4 Telephone interviews 8

2.3.5 Mailed Questionnaire method 8

2.3.6 Observation Method 8

2.3.7 The Experimental Method 9

2.4 Components of a good Questionnaire designs 9

2.5 Methods of collecting secondary data 9

2.6 Qualities of a Good Questionnaire 10

2.7 Factors that determine the Choice between Primary and Secondary Data 10
2.8 Sampling and Sampling Techniques 10
2.8.1 Sampling Techniques 11
2.8.1.1Probability sampling or random sampling 11
2.8.1.2Non-probability or Non-random sampling.. 12
CHAPTER THREE 13
CLASSIFICATION AND PRESENTATION OF DATA.. 13
3.0 Learning Objectives 13
3.1 Introduction 14
3.2 Types of classification 14
3.2.1 Geographical data 14
3.2.2 Periodical Classification 14
3.2.3 Qualitative Classification 15
3.2.4 Quantitative classification 15
3.3 Graphical and diagrammatic presentation of Data 20
3.3.1 Pictograms 20
3.3.3 The Histogram and Frequency Polygon 22
3.3.4 Bar Chart 24
3.3.5 Multiple and component Bar charts 25
3.3.6 Multiple Bar Charts 26
CHAPTER FOUR 29
MEASURES OF CENTRAL TENDENCY 29
4.0 Learning Objectives 29
4.1 Introduction 30
4.2 Mean 30
4.2.1 Arithmetic mean of ungrouped data 30
4.2.2 Arithmetic Mean of Grouped Data 31

4.2.3 Assumed Mean of Ungrouped Data 34

4.2.4 Assumed Mean of Grouped Data 35

4.2.5 Harmonic Mean 36

4.2.6 Geometric Mean 37

4.3 The Median 39

4.3.1 Median of grouped Data 40

4.4 The Mode 43

4.4.1 Mode of grouped data 45

4.5 Measures of Positional values 46

4.5.1 Quartiles 47

4.5.2 Deciles 50

4.6 Percentile 52

CHAPTER FIVE 57

MEASURES OF VARIABILITY 57

5.1 Learning Objectives 57

5.2 Introduction 57

5.2.1 The Range 58

5.2.2 The Mean Deviation 58

5.2.3 Variance and Standard Deviation 60

5.2.4 Standard Deviation 61

5.2.5 Coefficient of variation 64

5.2.6 Skewness and Kurtosis 66

5.2.6.1 Skewness 66

5.2.6.2Kurtosis 68

CHAPTER SIX 73

PROBABILITY AND PROBABILITY RULES 73

6.1 Learning Objectives 73

6.2 Permutation and Combination....74
6.2.1 Permutation 74
6.3 Probability and Probabilities Rules....81
CHAPTER SEVEN 87
PROBABILITY DISTRIBUTIONS....87
7.1 Learning Objectives 87
7.2 Bernoulli Trial 87
7.3 Binomial Distribution....88
7.4 Poisson Distribution....91
7.5 Normal Distribution 95
7.5.1 Properties of Normal Distribution....96
7.5.2 How to use the Normal Table....99
CHAPTER EIGHT 103
REGRESSION AND CORRELATION....103
8.1 Learning Objectives 103
8.2 Introduction....103
8.3 Correlation....106
CHAPTER NINE 117
SAMPLING AND DISTRIBUTION ESTIMATION 117
9.1 Learning Objectives 117
9.2 Introduction....121
9.3 Properties of a Good
Estimator……………………………………………………………..105
9.4.2 Interval Estimate 121
9.5 Estimation of population proportion 122
CHAPTER TEN....126
TEST OF HYPOTHESIS 126

10.1 Learning Objectives 126
10.2 Introduction 126
10.2.1 Types of Hypothesis 126
10.2.2 Test Procedures 126
10.2.3 Some Common Test 127
10.3 Analysis of Variance (ANOVA) 134
10.4 Contingency Table/Goodness Fit 137
CHAPTER ELEVEN 141
LABOUR TURNOVER 141
11.1 LEARNING OBJECTIVES 141
11.2 Introduction 141
11.3 Meaning of Labour Turnover 141
11.4 Measurement of Labour Turnover 142
11.5 Causes for Labour Turnover 143
11.6 Controllable Causes (Avoidable Causes) 144
11.7 Uncontrollable Causes (unavoidable causes) 144
11.8 Effects of Labour Turnover 144
11.9 Cost of Labour turnover to Employees 145
11.10 Costs to the Employer 145
11.11 Measures to Control Labour Turnover 145
11.12 Index Numbers 146
11.12.1 Price index 146
11.12.2 Types of Price Index 146
11.12.3 Quality Index 147
11.12.4 Factors affecting the construction of Index Numbers 147
11.13 Price Relatives 148
11.13.1 Quantity Relatives 148

11.13.2 Value Relatives 151
11.14 Methods of constructing Index 151
11.14.1 Simple Aggregate price Index 151
11.14.2 Simple Average of Relative Method 152
11.14.3 Weighted Aggregate Method 152
CHAPTER TWELEVE 157
THE NATURE AND IMPORTANCE OF RESEARCH 157
12.1 Learning Objectives 157
12.2 Introduction 157
12.3 The Importance of Research 157
12.4 Guidelines for selecting a re-searchable topic 158
CHAPTER THIRTEEN 160
PLANNING THE RESEARCH 160
13.1 Learning Objectives 160
13.2 Writing the Research proposal 160
CHAPTER FOURTEEN 163
LITERATURE REVIEW 163
14.1 Learning Objectives 163
14.2 Purpose of Literature Review 163
CHAPTER FIFTEEN 164
DATA ANALYSIS AND REPORT WRITING 164
15.1 Learning Objectives 164
15.2 Data Analysis 164
15.3 Descriptive Statistics 165
15.4 Inferential Statistics 166
15.5 Dissemination of Research Output 167
15.5.1 The common type of Scientific Papers include 168

GENERAL EXERCISES....171

REFERENCES: 179

LIST OF FIGURES

Figure 1: Classification of Sampling 11
Figure 2: Pie chart diagram of the distribution of value added in First Bank (Nig.) Plc in 2000 22
Figure 3: Histogram frequency polygon displaying the number of workers' monthly salaries. 24
Figure 4: Bar chart of Net Assets from 1990 to 1995 25
Figure 5: Component Bar chart showing the students' enrolment from 1994 - 1999 26
Figure 6: Multiple Bar Charts showing the scores of four CIPM candidates in three courses. 27

CHAPTER ONE
THE NATURE OF BUSINESS STATISTICS

1.0 Learning Objectives

At the end of this chapter, the students should be able to;

i. use common statistical terminology in the context of business applications;
ii. identify different types of sampling; and
iii. explain the distinction between categorical independent and dependent variables.

1.1 Introduction

The candidates at the end of the chapter should be able to collect information properly in making decisions. Also, problems concerned with the collection, analysis, and interpretation of data lie in the domain of the field of statistics; therefore, it is essential for anyone making business decisions on the basis of data to have a clear understanding of the field under considerations.

1.2 Nature of Business Statistics

The word "statistics" is defined as the method of data collection, organization, analysis and interpretation of numerical data to give meaningful results. Data is collected, processed and needed for planning and policy implementation.

It refers to facts derived from figures relations to any phenomenon such as population level, sales figures, profit, national income, state of health, student's enrolment in primary, secondary, and tertiary institutions and so on. Thus, statistics is synonymous with figures. Statistics as a discipline can be comprehensively defined as the scientific method for collection, organizing, summarizing, presenting and analyzing data for the purpose of making reasonable and conclusive decision under uncertainty.

The primary goal of statistics is to help give assistance in the process of decision making. Therefore, statistics as a key to decision making is a multidisciplinary related subject because its application is found in all disciplines of human knowledge especially when the available information can be transformed into numerical form of data such as management, agriculture, education, finance, banking, health and so on. Statistics is the catalyst to a societal development as guiding in the word in its present time. It can be classified into two categories viz: Descriptive statistics and Inferential statistics or Analytical statistics or statistical science. The method of describing the entire population through diagrams, graphs, pictogram (pictures), tables, is known as Descriptive statistics, while the method of examining fractional part of the population through mathematics and probability is called inferential statistics.

1.3 Uses and Importance of Statistics in Human Resource Management

The importance of statistics in human endeavor is numerous and cannot be over emphasized;

(i) Policy formulation: The role of formulation and planning of policies by top managers of various business organisation and the various federal or state government cannot be overemphasized. The volume of trade output and input of industries like manufacturing and mining, agriculture, wages, price, bank deposit, bank loan and advances, clearing house return, sales, purchase, production, finance, personal, accounting and market research. For each of these major areas, statistical data can be obtained. For instance, statistical tables and charts are frequently used by sales managers to present facts of sales; sampling methods and Markov chain are used by market researchers in determining the customer's preference for a particular brand and predict future change in market share of the market participants. Statistical quality control is employed by quality control expert to determine the level of conformity of newly manufactured products to a defined standard.

(ii) Government use in budgeting and control: statistical data and method also play an important role in budgeting and control by government organization for quick and easy decision on growth and survival of business. This is the fact that statistics supplies complex and unwieldy mass of data in such a way that these figures could be reduce to total, ratio, percentage, and average for quick understanding. For example, the Federal Government (FG) through its monetary and fiscal policies of the fifth development plan was trying to reduce the rate of unemployment, inflation, instability of price and underutilisation of the labour force.

(iii) Government uses statistics as an aid in the day- to- day conduct of the affairs of the state administration. Specifically, it is used in routine administrative control in policy and decision making in planning for accelerated economic growth and development.

(iv) In sports, for instance, a commentator in the radio television tells that they will give the statistics of the game (for example football) in a few minutes and before the twinkling of an eye they would show on the screen the number of free kicks, number of corner kicks, throwing, offside, yellow cards, red card and number of goals scored, shot at goal and so on. All these facts would tell us how tough or tense the match is or the side that is mounting pressure on the other. Even some past event of results of matches can be used to predict what will happen in future.

(v) In medical and pharmaceuticals like testing for the effectiveness of a drug. Statistical techniques can also be used to determine the dosage of drugs to be taken by patients. It is also useful to compare the effectiveness of two type of drugs on a patient.

(vi) In industry, statistics is used in quality control: It is used to control the quality of products being produced and to elevate new products before they are marketed.

(vii) In economic activities, it is useful in solving a variety of economic problems like, average, price, analysis of time series and demand analysis. It can also be use to facilitate the development of economic theory.

(viii) Meteorologist use statistical methodology to predict weather forecast future.

(ix) In agriculture, it is used to test the effect of feed on the average weight of chicken, the effect of the use of fertilizers on yield etc.

(x) The likely success of development policies in achieving their aims can be greatly enhanced by the use of statistics at key stages.

In conclusion, there is no short cut to planning of government personal projects. Accurate and timely statistics is the answer to all problems be it economic, social or behavioral.

CHAPTER TWO
STATISTICAL DATA

2.0 Learning Objectives:

At the end of this chapter, the students should be able to;

i. organise data into a regular or a grouped frequency table;
ii. understand data that are presented in a table; and
iii. to organise data into frequency distribution graphs, include bar graphs, histogram, polygon etc.

2.1 Introduction

There are two types of statistical data. They are Quantitative data or Numerical data and Qualitative data or categorical data. Quantitative Data may be found in almost all over daily activities for examples, enrolment of students, enrolment of candidates in profession examination such as chattered institute of personal management of Nigeria (CIPM), Chartered institute of Bankers of Nigeria (CIBN), Chartered institute of Taxation of Nigeria (CITN), Chartered institute Professional. Management (CIPN), distance covered in a number of kilometer and so on. Here, it should be noted that not all quantitative data is regarded as statistical data. Quantitative data /information suitable for statistical analysis must be a set of numbers showing significant relationship. Statistical data must be computed, analyzed and interpreted to give a meaningful result. A single data that cannot compare with or which shows no significant relationship with another data is not a statistical data.

Qualitative Data refers to the data that can be expressed as Yes or No, True or False, success or failure, good or bad, alive or dead. All the statistical data described above can be primary unprocessed data or secondary processed data. These types of data will be discussed fully in the later part of the text.

2.2 Type of Statistical Data

Basically, there are only two sources of data in statistics. The statistical sources refer to the data that are collected for some official purpose and this include censuses and officially conducted surveys. On the other hand, Non – statistical sources refer to the data that are collected for other administrative purpose or private sector.

However, statistical source refers to data that is gathered for official purposes, incorporate censuses, and officially administered survey. Finally in statistics, we have

(i) Internal sources which means when data is collected from reports and records of the organization itself. Examples; if a company publishes its annual report on profit and loss, total sales, loans, wages and so on.

(ii) External source is when data is collected from source outside the organization. For example, when a tour and travel company collects information on XYZ tourism from ABC Transport Corporation.

Types of Data

There are two types of data:

1. Primary data simply means first-hand or unprocessed or fresh information collected by an investigator and use for the purpose that is meant for. The primary data is the original data and it is a more reliable data in nature for example; the population census conducted by the government of Nigeria after Ten (10) years.

2. Secondary data refers to second-hand information or processed information. This secondary data is not originally collected by the investigator. It is not reliable like the primary data because they have been processed and published. For example, data extracted from the records of an institution, firm, industry etc.

2.3 Methods of Data collection

The following are the major methods of collecting primary data.

2.3.1 Direct personal or Interview method: under this method, the investigator or interviewer obtains first-hand information directly from the respondents. The investigator will personally visit the respondent or interviewer directly to collect the information or data.

Advantages:

(a) The data is first-hand information.

(b) The data collected is accurate and more reliable.

(c) There is room to modify questions to suit the level of the respondent that means it is flexible.

(d) Additional information may also be collection alongside with the referred information for future use.

Disadvantages:

(a) It is not suitable for a wide coverage area.

(b) It is time consuming.

(c) It is very expensive.

(d) Personal bias can be introduced in recording the opinions of the respondents.

2.3.2 Indirect Oral Interview.

In this method, the investigator or interviewer interview several persons who are directly or indirectly in touch with the informants.

Advantages:

(a) Wider area can be covered within limited time.

(b) It is economical i.e. not expensive.

Disadvantages:

(a) There is a possibility of fake response from the respondent.

(b) The degree of response may be low.

(c) Information collected from different respondent for the same questions may not be homogeneous and comparable.

2.3.3 Information through correspondents

Under this method, the interviewer or investigator may employ local agents or correspondents and trained them to collect data from the respondents.

Advantages:

(a) It is very useful where the field of investigator is very wide.

(b) It is more economical and less time consuming.

(c) It is very suitable for some special investigation.

Disadvantages:

(a) Lack of conformity.

(b) It is not reliable.

(c) Data are not accurate.

(d) It is more expensive.

(e) It is time consuming.

2.3.4 Telephone interviews

Under this method, information is collected through a telephone call.

Advantages:

(a) Large coverage can be achieved.

(b) It is more economical and less time consuming.

(c) The data collected through this method is reliable.

Disadvantages:

(a) Limited access to respondents.

(b) Lack visual feedback.

2.3.5 Mailed Questionnaire method

Under this method, a set of prepared questions related to the investigation is sent to respondents by post along with the instructions to fill. The respondents or the informants, after filling up the questionnaire, send it back to the investigator.

Advantages:

(a) Possibility of wide range coverage.

(b) It is quite economical.

(c) Data or information is free from bias.

(d) The information collected through this method is original.

(e) Secrecy is maintained.

Disadvantages:

(a) It is only useful where the majority of the respondents are educated.

(b) Low rate of feedback.

(c) There is tendency for mis-interpretation of the questions by some respondents.

(d) It is time consuming and non-economical.

2.3.6 Observation Method

This method is concerned with actual perception of situation and it is a direct method of accessing respondents' behavior. It seizes to assertion what people think or how they behave by direct contact with respondents over a period of time. There are two kinds of participants, in this method first, the judge or non-participants, this is because the judge or non-participant observers are usually detached from the situation they observing, so that they are not part of the group being investigating. Secondly, we have participant observer. It is common in social or behavioral setting.

Difficulties of the Observation Method

(a) Presence of observer which usually have a negative impact on the information collected.
(b) Problem of classifying the information for comparison and analysis.

2.3.7 The Experimental Method

This method is rightly controlled in which the researcher holds all variable constant while manipulating one of them to see its effect on the response variable which he is measuring. For example, in the days of competition in banking sector, a researcher might try to experiment with vary interest rates on the earnings of the bank, keeping all other variable like advertising budgeting, impact of personnel on customers, type of bank, buildings and so on constant.

2.4 Components of a Good Questionnaire Designs

Questionnaire is divided into three parts namely:
(i) **Introduction:** this introduction serves as explanation as to why somebody is conducting the research or investigations.
(ii) **Personal data:** like slang for names, sex, state of origin, local government, nationality, level of education, marital status, occupation and so on.
(iii**) Actual questions:** This refers to questions that will help the researcher with research problem. Recording of information from the questionnaire itself. The respondents can read and write (indirect-when the respondents cannot read and write)

2.5 Methods of Collecting Secondary Data:

(a) Government records and publications.

(b) Newspapers.
(c) Unpublished sources.
(d) Online journals, records and publications.
(e) Business and industry records yearly publication.
(f) Professional bodies annual publications.

(g) Nigerian year Book.

(h) Encyclopedia and so on.

2.6 Qualities of a Good Questionnaire

(i) The purpose of the study should always be clearly stated.

(ii) The questions should be simple and brief.

(iii) The questions should be unambiguously phrased in understandable term.

(iv) The questions should be designed so that they fall into a logical sequence of order.

(v) Clarity of the questions is essential.

(vi) Avoid embarrassing questions like how many man friend does one have?

(vii) There should be no leading questions.

2.7 Factors that Determine the Choice between Primary and Secondary Data

The following are some of the factors that determine the choice between primary data and secondary data.

(i) The purpose of the enquiry.

(ii) Time required for the exercise.

(iii) Standard of accuracy demanded.

(iv Funds at the disposal of investigator.

(v) Nature of statistical investigation.

2.8 Sampling and Sampling Techniques.

The process of selecting a required number of items or persons i.e. whether living things or non-living things from the entire population of items is called sampling. It has been stated earlier that, data can be generated from populations or samples. A population or universe is defined as a totality of elements or items under study. In other words, population or universe is the set-of items or measurements a researcher wants to study. For example, a set of all current account holders in a bank can constitute a population under a given situation. Under another situation, the subject of all current account holders who are enjoying a type of credit facility or the other can constitute the population.

To collect information from the entire units in the population, it may be expensive and time consuming; therefore, it requires an alternative approach of selecting a representative sample from the population which is studied. It should be noted that, a representative sample is one which is scientifically selected such that, it has all the characteristics of the population under study.

2.8.1 Sampling Techniques

There are two classification of sampling techniques. Sampling may be random or non- random. Random sampling is also called probability sampling while Non-random sampling is called Non-probability sampling.

The figure (1) below illustrates a diagrammatically, representation of the classification of sampling.

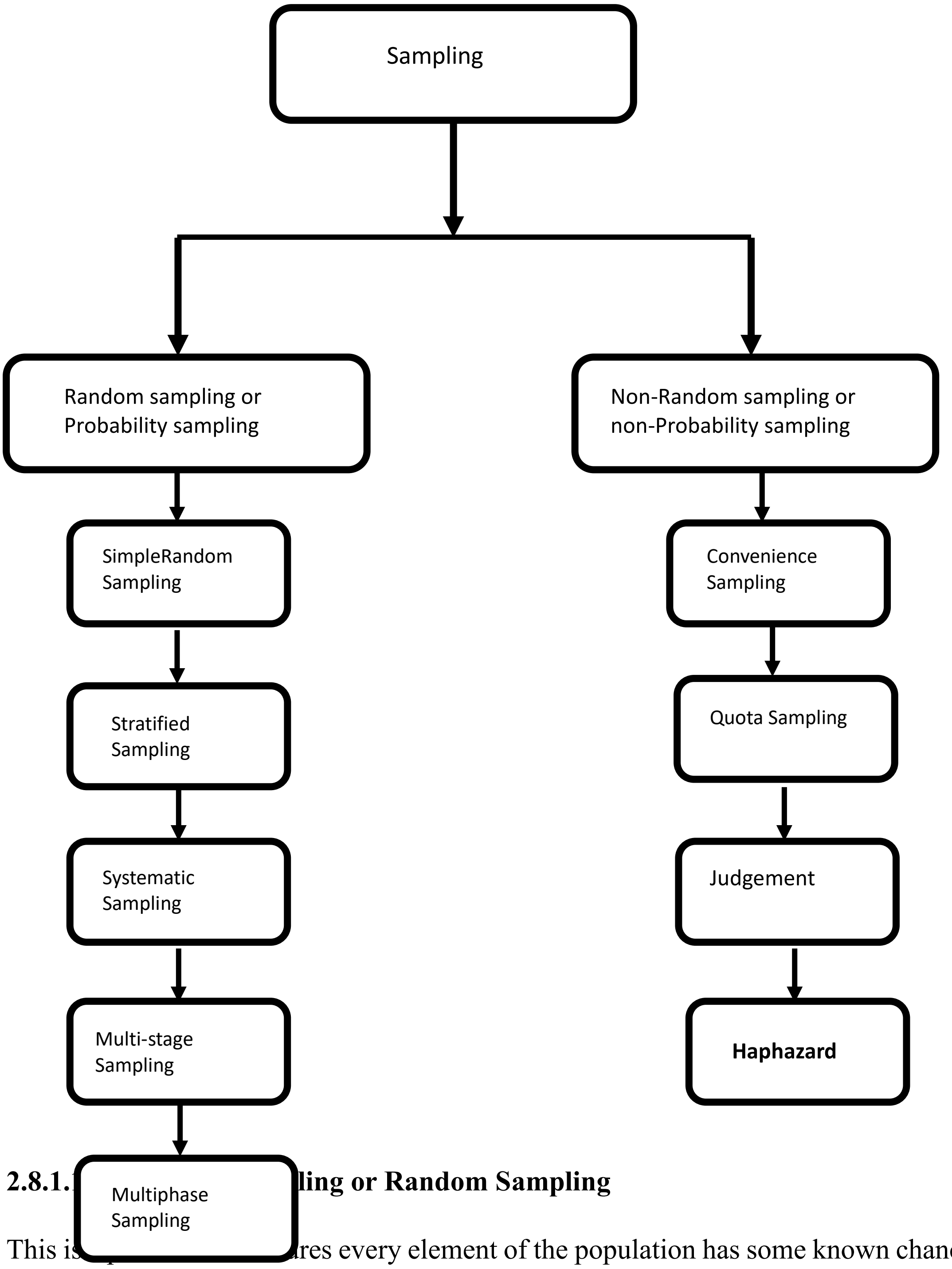


Figure 1: Classification of Sampling

2.8.1.[illegible] ling or Random Sampling

This is [illegible] res every element of the population has some known chance of being selected. Units selected using probability samplings are generated using random numbers which are numbers that have equal probabilities of selection. The following are probability sampling techniques:

1) **Sample Random Sampling (SRS)**: Simple random sampling is the process whereby all the units of the population being surveyed have equal probabilities or chance of being selected. For example, if we have a list of numbers, 1, 2, 3, 4, 5, 6, 7, 8, 9. The chances of selecting any of the numbers are 1/9.
2) **Stratified Sampling**: In stratified sampling the population is first divided into different groups randomly called strata and each group is referred to as stratum. Samples are then taken from each stratum or group using simple random sampling.
3) **Systematic Sampling**: It is a method of selecting the first unit of the sample On the random basis and choosing the additional unit at regular fixed intervals until the required sample size is attained.
4) **Cluster Sampling**: In this case, the population is divided into sampling units or clusters, for example, zones within educational or geographical areas. Having divided the population into clusters, we can use simple random sampling to select the clusters to be examined.
5) **Multistage Sampling**: This involves several stages. It is a series of sampling taken at successive stages. At the first stage of the sampling, primary units are selected and at the second stage secondary units are selected method.

2.8.1.2 Non-probability or Non-random sampling

This is a process whereby the selection is not based on chance (or probability), but on personal convenience, expert judgement or any other of conscious researcher selection method.

1) **Quota Sampling**: It involves the division of the population into a number of segments with the researcher arbitrarily selecting a quota of sample from each segment or cell. There are three reasons in using Quota sampling that makes it to have advantages over stratified sampling.

a) Decking how many segments or cells into which the population will be divided.
b) Decking what percentage of sample items that should be in anyone cell.
c) Selecting the sample item.

2) **Haphazard Sampling**: In this case, the interviewer makes a random selection according to the dictate of his mind the method encourages the introduction of bias into the selection of unit of interest.

3) **Judgement Sampling**: This is a system whereby samples are based on the researcher's intuition and judgement of what constitutes a representative sample or other subjective criteria. Sample representatives will depend on his or her good judgement and possibly a little bit of good luck.

4) **Convenience Sampling**: This is so named because it is convenient for the researcher to first select a few sample items, rather than go through the laborious process of obtaining a probability sample

Advantages of Sampling:

1. Sampling saves time and money
2. It encourages accuracy
3. Errors can be measured and readily handled statistical

CHAPTER THREE
CLASSIFICATION AND PRESENTATION OF DATA

3.0 Learning Objectives:

At the end of this chapter, the students should be able to;

a) choose the right Table or graph for the right data and audience;
b) ensure that graphics are self-explanatory; and
c) create graphs and Tables that are attractive.

3.1 Introduction

This Chapter deals with the classification and presentation of data in a clear and understanding manner for making statistical inferences. Data collected using any of the sampling techniques need to be summarized and edited, Classified, tabulated and presented in Tables for better understanding before proceeding to the analysis.

After the data have been collected, summarized and edited, the next step is tabulation or classification of the data.

3.2 Types of Classification

The following are some of the classifications of data notable in the statistics field.

3.2.1 Geographical data: is made according to geographical differences. For example, the classification of data in terms of wards, cities, local government, constituencies, states, and so on.

Example 3.1: The Table below shows the population census for 5 local government areas in Oyo - State, Nigeria in 2003.

Table 1: The Population of 5 Local Government Areas in Oyo - State, Nigeria in 2003

Local Government	Population (Million)
Ibarapa	3.5
Akinyele	5.0
Oyo-west	4.0
Saki East	4.5
Ogbomoso North	3.5

3.2.2 Periodical Classification

It is a situation where data are classified over a specific period in the occurrence

Example: 3.1: The information in the Table below shows the profit in (₦' m) of Private Liability Companies in a state, between March and August 2000.

Table 2: The profits (N,' m) of Private Liability Companies in a state between March and August 2000

Months	Profit (₦'m)
March	4.5
April	3.6
May	5.2
June	4.0
July	3.5
August	4.2

3.2.3 Qualitative Classification

Some data are classified according to the attributes (in terms of qualities) such as good or bad, marital status, sex, education attachment, tribes, religious affiliation and so on.

Example 3.3: The Table below illustrates the result of an election in three local governments in the Ibarapa zone, and their distribution by sex.

Table 3: The results of an election in three local governments in the Ibarapa zone and their distribution by sex

Local Government	Male	Female	Total
Ibarapa North	50	30	80
Ibarapa central	80	45	125
Ibarapa East	70	35	105
Total	200	110	310

3.2.4 Quantitative classification

Classifications under the quantitative classification are measurable and continuous.

Examples of quantitative data are weights, length, height, number of items (living and non-living) and so on.

Example 3.3: The following data gives the amounts (₦'000) spent on feeding by twenty households in a certain community in February 2021.

32, 22, 19, 18, 43, 42, 40, 43, 18, 21

35, 38, 34, 41, 36, 25, 22, 45, 48, 26

Construct a frequency distribution table using a class interval of size 5 starting from 15 - 19, 20 – 24 and so on.

The above data can be classified thus:

Table 4: The frequency distribution Table of the amounts (N'000) spent on feeding by twenty households in a certain community in February 2021

Amount	**Tally**	**Number of Households**
15 – 19	III	3
20 - 24	II	2
25 – 29	II	2
30 – 34	II	2
35 – 39	III	3
40 – 44	~~IIII~~	5
45 – 49	II	2
Total		**20**

However, we can classify the data into various classes of frequency distribution if we have a large set of data.

The following are the steps involved in classifying raw data into various classes in a frequency Table.

Step 1: Calculate the range i.e., the difference between the maximum observation – Minimum observation (i.e., Maximum observation – minimum observation). For example, the range of the following sets of data 6, 7, 2, 12, 4, 8, 15, and 20 are given as:

$$Range = 20 - 2 = 18$$

Step 2: Determine the number of classes.

The number of classes is defined as 1+ 3.22 logN, where N is the total number of observations given.

For example, the number of classes in a given set of observations: 4, 5, 7, 3 and 1

i = 1+ 3.22 log (4+5+7+3+1)

= 1 + 3.22 log 20

= 1 + 3.22 (1.3010)

= 1 + 4.18922

= 5.189 Approximately 5.0

Step 3: Determine the size of the class intervals as

$$c = \frac{range}{Number\ of\ classes}$$

Example: 3.4: Determine the class interval of the following set of observations:

10, 12, 13, 21, 22, 15, 16, 18, 24, 12, 16, 14, 15, 25, 22,

18, 17, 11, 8, 9, 28, 24, 12, 14, 14, 21, 22, 18, 6, 7

Solution:

Range = 28 - 6 = 22

Number of classes = 1 + 3.22 log N

= 1+3.22 log (10 + 12 …+ 6+7)

= 1 + 3.22 log (484)

= 1 + 3.22 (2. 685)

= 1+8.6457

= 9.6457 Approximately 10

Hence, class interval (C)

$$c = \frac{22}{10} = 2.2 \cong 2$$

Step 4: Building up the classes by starting with the smallest data and using the class width obtained in Step 3.

Step 5: Assign data into various classes such that each datum belongs to one and only one class.

Step 6: Finally, determine the number of data belonging to each class.

Definitions of Some Useful Terms:

Class Limits: This is the lowest and highest value a class can take e. g in a class of 21-30, the lowest value the class can take is 21 and the highest value the class can take is 30.

Class Boundary: These are the lower and upper-class boundaries. That is, the lower class boundary of a class is the mid-point of the lower class limit of the immediate proceeding class.

For example, in a class of 31 – 40, the lower-class boundary is 31 i.e. the lower-class limit of the preceding class is

$$= \frac{31 + 30}{2} = 30.5$$

The upper class of a class is the midpoint of the upper class limit of that class and the lower class limit of the next higher class.

Relative frequency: this is the proportion of the frequency of a class to the total frequency multiplied by 100.

Class Midpoint: This is the average of the lower and upper-class limits of a class.

For example, the midpoint of class 41 – 50

$$= \frac{41 + 50}{2} = 45.5$$

Frequency: This is the number of times an observation appears in a given set of observations.

A frequency Distribution Table: This is a table indicating the classes and the respective frequencies.

Example 3.5: A group of 30 CIPM professional candidates participated in an Economic Theory test and their scores were recorded based on the assessment as follows: 12, 18. 19. 9. 10. 12. 15. 17. 8. 10. 25. 27. 29, 14, 7, 15, 11, 6, 13, 12, 14, 15, 13, 28, 24, 21, 22, 24, 23, 18.

Required:

(a) Classify the data above using classes 5-9, 10-14 and so on.
(b) Construct a frequency distribution table for the above.
(c) From (a) above, determine the following:

i. Range
ii. Frequency of the 2^{nd} class
iii. Midpoint of the 3^{rd} class
iv. Class width of the 1^{st}class
v. Upper and lower class limits of the 4th class
vi. Relative frequency of the 3^{rd} class
vii. Sum of the frequencies preceding the 4^{th} class.

Solution: (a) and (b)

Table 5: Frequency distribution Table of scores of 30 CIPM professional candidates who participated in an Economic Theory test

Class	Tally	Frequency
5–9	IIII	4
10–14	IIIIIIIIII	10
15–19	IIII II	7
20 – 24	IIII	5
25 – 29	IIII	4
		30

(i) Range = 29 – 6 = 23

(ii) Frequency of the second class is = 10

(iii) Midpoint of the 3^{rd} class $= \frac{15+19}{2} = 17$

(iv) Class width of the 1^{st} class = 9-5 = 4

(v) Upper class and lower class limits of the 4^{th} class are 20 and 24 respectively

(vi) Relative frequency of the 3^{rd} class

$$= \frac{frequency\ of\ 3rd\ class}{Total\ frequency} \times 100$$

$$= \frac{7}{30} \times 100 = 23.33\%$$

(vii) Sum of the frequencies preceding the 4^{th} class = 4 + 10 + 7 = 2

3.3 Graphical and Diagrammatic Presentation of Data

The next step after the data has been classified is to present the data in such a way to be understood by a layman. Data can be presented in three major forms namely tabular form, graphical form and in the form of equations.

Data can be presented graphically using various types of graphs or pictorial representations. The most commonly used are histograms, bar charts, pie charts, frequency polygons and cumulative frequency curves (Ogive)

Characteristics of a Good Statistical Graph

The following are the characteristics of statistical graphs. Statistical graphs must have;

(i) General row and column titles

(ii) A scale

(iii) A source of data

(iv) A key.

3.3.1 Pictograms: These are representations of data. We may present data by symbols.

For example, the data below represent the number of candidates registered for the CIPM professional examination in the following years: 1990, 1991, and 1992.

Table 6: The number of candidates registered for the CIPM professional examination in the following years: 1990, 1991, and 1992

Years	Number of candidates
1990	2.5 million
1991	3.0 million
1992	1.5 million

Represent the above data using a pictogram.

Solution:

The Pictogram shows the number of candidates registered for the CIPM professional examination in three years.

Table 7: Pictogram representation of the number of candidates registered for the CIPM professional examination in the following years: 1990, 1991, and 1992

Years	Number of Candidates
1990	Φ ΦΛ
1991	ΦΦΦ
1992	ΦΛ

Where $\Phi = 1$ million candidates

And $\Lambda = \frac{1}{2}$ million candidates.

3.3.2 Pie-chart: This is a representation of data in a circle expressed in degree. To construct a pie chart, we divide a circle of the convenient radius into sectors in proportion to the relative contributions of the items under consideration.

Example 3.6: The distribution of value added in First Bank (Nig.) Plc in year 2000 is as given below. Present this information in the form of a pie chart.

Table 8: The distribution of value added in First Bank (Nig.) Plc in year 2000.

Source	Amount (₦'000)	Computations	Values (degrees)
Employees	60	$= \frac{60}{720} \times 360^0$	30
Retained earnings	180	$= \frac{180}{720} \times 360^0$	90
Government	36	$= \frac{36}{720} \times 360^0$	18
Stakeholders	30	$= \frac{30}{720} \times 360^0$	15
Statutory reserves	120	$= \frac{120}{720} \times 360^0$	60

Depreciation	294	$= \frac{294}{720} \times 360^0$	147
Total	720		

Solution:

Total Value added = 720

Since 360 ^{O} is a cycle.

Each sector is computed as $= \frac{Number\ in\ a\ sector}{Total\ for\ all\ sectors} \times 360^{\circ}$

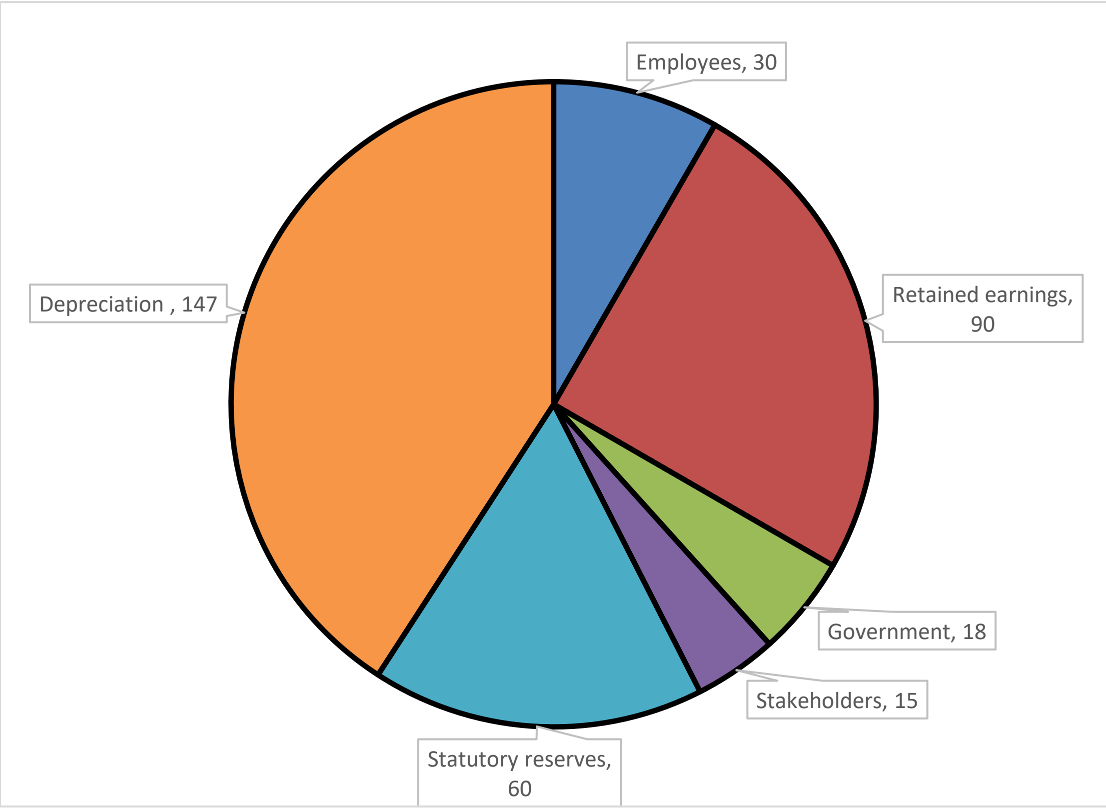


Figure 2: Pie chart diagram of the distribution of value added in First Bank (Nig.) Plc in 2000

3.3.3 The Histogram and Frequency Polygon:

The histogram is a graphical display of the frequency distribution. It looks more or less like a bar chart except that there are no spaces between the "bars" of the histogram. The broken line joining the midpoint of the heights of the histogram is called frequency-polygon

Example 3.7: The Table below shows the monthly salary of Land's Business Services workers

Table 9: The monthly salary of Land's Business Services workers

Monthly salary (₦'000)	**Number of Workers**
5 - 9	8
10 - 14	5
15 - 17	7
20 - 24	12
25 - 29	10
30 - 34	2
35 - 39	6

Display the information using (a) Histogram (b) Frequency Polygon

To find the histogram, we transform the monthly salary into a continuous class.

To do this, we subtract 0.5 from all lower-class limits and add 0.5 to all upper limits. The above Table then becomes;

Table 10: The class boundaries of monthly salary of Land's Business Services workers

Monthly Salary (Class Boundaries)	**Frequency**
4.5 - 9.5	8
9,5 - 14.5	5
14.5 - 19.5	7
19.5 - 24.5	12
24.5 - 29.5	10
29.5 - 34.5	2
34.5 - 39.5	6

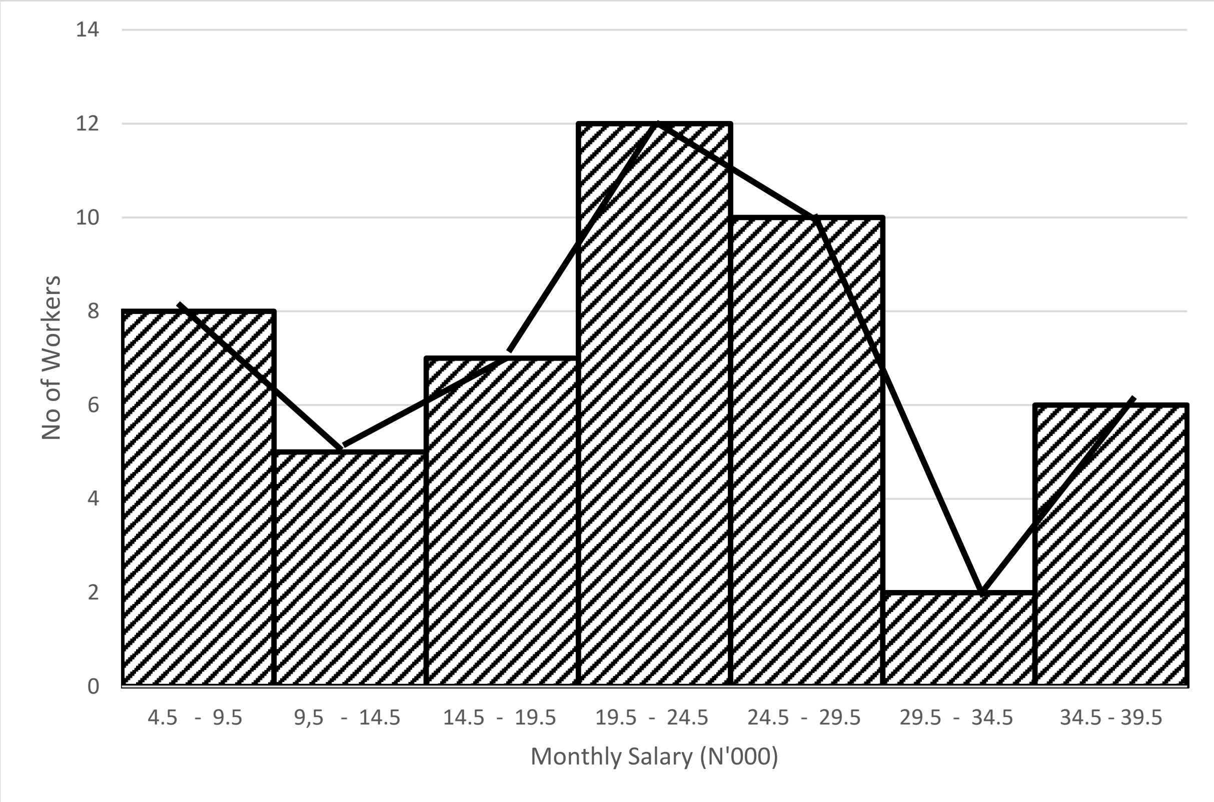


Figure 3: Histogram frequency polygon displaying the number of workers' monthly salaries.

3.3.4 Bar Chart

The Bar chart is more suitable for displaying the frequency distribution of information relating to

Qualitative or attribute data: This is because attribute data has distinct end points and cannot be merged as in the case of histogram, Bar chats can be classified as simple bar charts, component bar charts and multiple bar charts.

Example 3.8: Present the information in the Table below on a bar chart.

Table 11: Net assets (N'000)

Year	**Net Assets (₦'000)**
1990	35

1991	50
1992	25
1993	40
1994	60
1995	55

Solution:

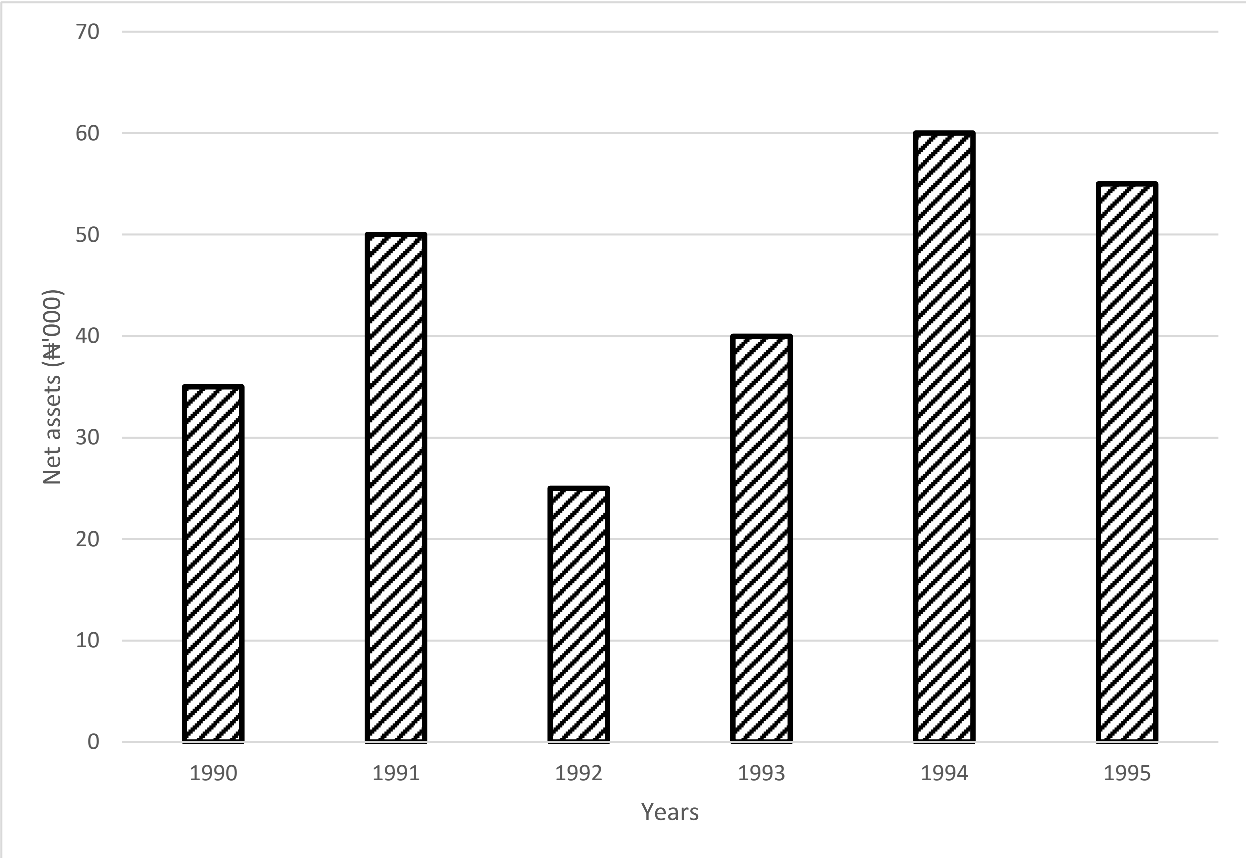


Figure 4: Bar chart of Net Assets from 1990 to 1995

3.3.5 Multiple and Component Bar Charts:

Component Bar chart: This is the case where the given set of data is divided into component parts. It provides a breakdown of the quantity into component parts.

Example 3.9: The data below represent students' enrolment by sex into ABC University in the year 1994 - 1999.

Table 12: Students' enrolment by sex into ABC University in the year 1994 - 1999.

Year	Male	Female
1994	40	30
1995	35	45
1996	20	40
1997	40	25
1998	30	30
1999	45	15

Represent the above data using the component bar chart.

Solution:

Number of Students enrolment

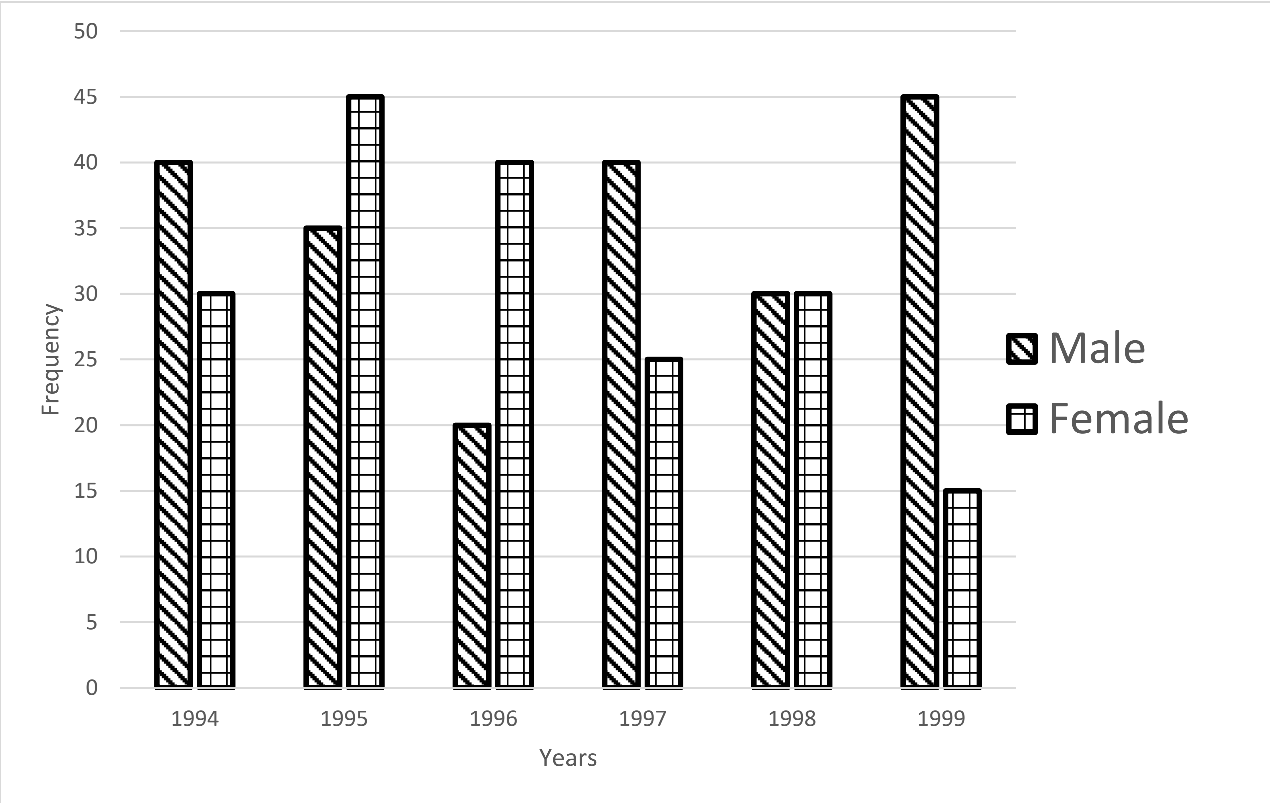


Figure 5: Component Bar chart showing the students' enrolment from 1994 - 1999

3.3.6 Multiple Bar Charts: Each component part is represented by a bar. The height of each bar is determined by the magnitude of the individual component.

Example 3.10: The data below reflects the scores of four candidates in the CIPM professional examination in three courses.

Table 13: The scores of four candidates in the CIPM professional examination in three courses.

Courses	Candidates			
	I	**II**	**III**	**IV**
Business Law	60	40	20	35
Taxation	30	20	25	20
ICT	20	30	35	45

Represent the above data using multiple Bar Charts.

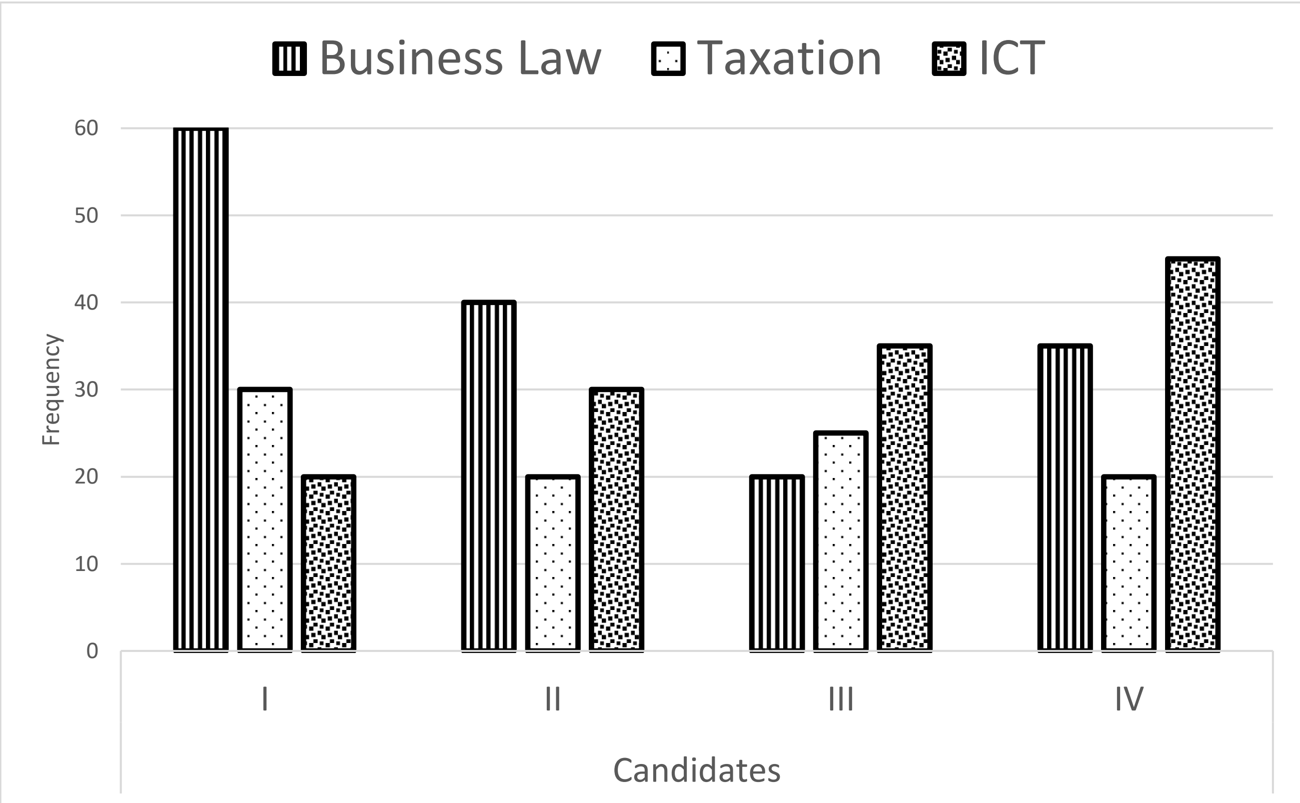


Figure 6: Multiple Bar Charts showing the scores of four CIPM candidates in three courses.

Example 3.11: In 1990, 15000 candidates entered for professional examinations. 25 percent of the candidates came from Lagos while the rest came from outside Lagos. Of those who came from Lagos, 65 passed the examination, 13 percent were referred and of the rest 0.7 were absent, 0.2 never received their result while others failed. Of those who came from outside Lagos, 0.4 passed the examination, 0.2 were referred and 0.3 never received their results. Of the rest, 8 percent were absent while all others failed the examination. Arrange the information above in a table.

Solution:

Total candidates = 15000

Those from Lagos = $\frac{25}{100} \times 15000 = 3750$

Those from outside Lagos = 15000 - 3750 = 11250

Of those who came from Lagos:

Those that passed = 65

Those referred = $\frac{13}{100} \times 3750 = 487.5$

The rest = 3750 – (65+487.5) = 3197.5

Those absent = $0.7 \times 3197.5 = 2238.25$

Those never received results $= 0.2 \times 3197.5 = 639.5$

Those that failed = 3197.5 – (2238.25 +639.5) = 319.75

Of those who came from outside Lagos

Those that passed = $0.4 \times 11250 = 4500$

Those referred = $0.2 \times 11250 = 2250$

Those never received results $= 0.3 \times 11250 = 3375$

The rest = 11250 – (4500+2250+3375) = 1125

Those absent = $\frac{8}{100} \times 1125 = 90$

Those that failed = 1125 – 90 = 1035

Table 14: CIPM candidates entered for professional examinations and their locations

	Lagos	Outside Lagos
Passed	65	4500
Referred	487.5	2250
Absent	2238.25	90
Never received result	639.5	3375
Failed	319.75	1035

CHAPTER FOUR
MEASURES OF CENTRAL TENDENCY

4.0 Learning Objectives

At the end of this chapter, the students should be able to;

i. Understand the differences between the mean, median and mode and use the differences in these values to analyse and compare the spread of the data; and
ii. Compute the mean, median and mode.

4.1 Introduction

In chapter one, we learnt that descriptive statistics involves the organisation and manipulation of data with a view to making some calculation that will help a researcher to summarise his data. A group of measures known as measures of central tendency is one such example. The aim of measures of central tendency is the identification of centrally located representative figure of a set of data. Among the measure of central tendency are mean, medium and mode.

4.2 Mean

There are three types of mean in statistics namely; the arithmetic mean, the harmonic mean and the geometric mean. The most commonly used in statistics is arithmetic mean. This is because of its use in statistical analysis like in computations of the measures of dispersion such as mean derivation, variance, standard deviation, co-efficient of variation, test of hypothesis and so on.

4.2.1 Arithmetic mean of Ungrouped Data

The arithmetic mean of a set of observations $x_1, x_2, \dots, x_n$ is given by

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

where $n =$ number of observations, $\sum$ = summation or addition.

Example 4.1: Find the mean of 8, 4, 10, 5, 3

Solution:

$$\bar{x} = \frac{\sum x}{n} = \frac{8+4+10+5+3}{5} = 6$$

Example 4.2: Find the mean of 12, 15, 16, 10, 12, 11

Solution: $\bar{x} = \frac{\sum x}{n} = \frac{12+15+16+10+12+11}{6} = \frac{76}{6} = 12.67$

Example 4.3: Find the average profit after tax of 20 banks in a given year in ₦'m

22, 32, 10, 19, 21, 30, 40, 16, 18, 19, 41, 50, 20, 30, 14, 12, 10, 18, 13, 1.

Solution:

$$\bar{x} = \frac{\sum x}{n} = \frac{22 + 32 + \cdots + 13 + 18}{20} = ₦22.2m$$

4.2.2 Arithmetic Mean of Grouped Data

If we have data set in its frequency distribution or grouped form. The arithmetic mean can be computed by the formula.

$$\bar{x} = \frac{\sum fx}{\sum f}$$

Example 4.4: Find the arithmetic mean of the profit after tax of the data in the table below using group mean?

Table 15

Profit after tax (₦'m)	Frequency (f)
9	1
10	2
12	1
13	1
14	1
16	1
18	2
19	2
20	1
21	1
22	1
30	2
32	1
40	1
41	1
50	1
-	20

Solution:

Profit after tax (₦'m)	Frequency (f)	fx
9	1	9
10	2	20

12	1	12
13	1	13
14	1	14
16	1	16
18	2	36
19	2	38
20	1	20
21	1	21
22	1	22
30	2	60
32	1	32
40	1	40
41	1	41
50	1	50
Total	**20**	**444**

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{444}{20} = \text{₦}22.2m$$

Example 4.5: Calculate the mean of the daily claims of the Table below:

Table 16

Time in days (x)	Number of claims (f)
20	50

21	65
26	40
29	20
30	5
34	10

Solution:

x	f	fx
20	50	1000
21	65	1365
26	40	1040
29	20	580
31	5	155
34	10	340
Total	**190**	**4480**

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{4480}{190} = 23.58$$

Example 4.6: Given the frequency distribution Table below, determine its mean.

Table 17

Overtime (Hours)	Frequency (f)
10 - 19	2
20 - 29	6
30 - 39	14
49 - 49	17
50 - 59	8
60 - 69	3

Solution:

x (mid value)	f	fx

14.5	2	29
24.5	6	147
34.5	14	483
44.5	17	756.5
54.5	8	436
64.5	3	193.5
Total	**50**	**2045**

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{2045}{50} = 40.9$$

4.2.3 Assumed Mean of Ungrouped Data

Given $x_1, x_2, \dots, x_n$ to be a set of observations, if A is an assumed mean, then the mean is given by

$$\bar{x} = A + \frac{\sum d}{n}$$

where $d = x - A$, A is the assumed mean and n is the number of observations

Example 4.7: Find the mean of 8, 12, 6, 2 and 8 using the assumed mean of (a) 10 (b) 4

Solution:

(a)

x	d = x – 10
8	-2
12	2
6	-4
2	-8
8	-2
Total	**-14**

$$\bar{x} = A + \frac{\sum d}{n} = 10 + \frac{-14}{5} = 10 - 2.8 = 7.2$$

(b)

x	d $= x$ - 4
8	4
12	8
6	2
2	-2
8	4
Total	**16**

$$\bar{x} = A + \frac{\sum d}{n} = 4 + \frac{16}{5} = 4 + 3.2 = 7.2$$

From, the examples (a) and (b) above, i.e. it can be seen that the mean are the same for any values of assigned mean

4.2.4 Assumed Mean of Grouped Data

The assumed Mean of grouped data is calculated by

$$\bar{x} = A + \frac{1}{\sum f}[\Sigma f(x - A)] = A + \frac{\sum fd}{\sum f}, where\ d = x - A$$

Example 4.8: An auditing company found that the frequency distribution of the annual salaries of the Chief Executives of 22 of its clients is as follows:

Table: 4.4

Salary range (N)	**Number of Chief Executives**
15,000 < 20,000	2
20,000 < 25,000	4
25,000 < 30,000	5
30,000 < 35,000	8
35,000 < 40,000	3

Determine the mean salary of the company.

Solution:

In this type of problem, we replaced the salary range (class interval) by replacing the midpoint of the values of the end points of each class. For example

$$\frac{15000 + 20000}{2} = 17{,}500 \ and \ so \ on.$$

f	x	fx(₦)
2	17,500	35,000
4	22,500	90,000
5	27,000	139,500
8	32,500	260,000
3	37,500	112,500
22	**-**	**635,000**

$$\bar{x} = \frac{\sum fd}{\sum n} = \frac{635000}{22} = ₦28863.64$$

4.2.5 Harmonic Mean

The harmonic mean of a set of observations $x_1, x_2, \ldots, x_n$ is simply defined as the reciprocal of the arithmetic mean of the reciprocals of a set of observations. This is given by;

$$HM = \frac{n}{\sum \frac{1}{x_i}}, i = 1, 2, \ldots, n \quad for \ ungrouped \ data$$

$$HM = \frac{\sum f}{\sum \frac{f_i}{x_i}}, i = 1, 2, \ldots, n \quad for \ grouped \ data$$

Example 4.9: Find the Harmonic Mean (HM) of the following set of data;

(a) $3, 4, 5, 1, 2, \frac{1}{10}, \frac{1}{5}$

(b) $4, 5, 10, 2, 4$

Solution:

$(a) \quad HM = \frac{n}{\sum \frac{1}{x_i}} = \frac{7}{\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{1}+\frac{1}{2}+10+5} = \frac{7}{0.33+0.25+0.2+1+0.5+10+5} = \frac{7}{17.28} = 0.4051$

$$(b) \quad HM = \frac{5}{\frac{1}{4}+\frac{1}{5}+\frac{1}{10}+\frac{1}{2}+\frac{1}{4}} = \frac{5}{0.25 + 0.2 + 0.1 + 0.5 + 0.25} = \frac{5}{1.3} = 3.85$$

Example 4.10:

The following group frequency distribution shows the cost involved in setting travel policy claims.

Table 18

Travel claims cost (₦)	**No. of Claim**
150 but less than 160	8
160 but less than170	17
170 but less than 180	12
189 but less than 190	10
190 but less than 200	3

Solution:

Class interval (₦)	**F**	**X**	$\frac{f}{x}$
150 <160	8	155	0.516
160 < 170	17	165	0.1030
170 <180	12	175	0.980
180 < 190	10	185	0.541
190 < 200	3	195	0.154
TOTAL	50	-	2.294

$$HM = \frac{\Sigma f}{\Sigma \frac{f}{x_1}} = \frac{50}{2.294} = ₦21.80$$

HM has limitations of assign too much weight to the smaller items and thus has limited scope. It is useful mean to measure rate of change and can be adapted to measures involving time and certain types of ratios and rates.

4.2.6 Geometric Mean

The geometric mean of a set of observations $x_1, x_2, \dots, x_n$ is obtained by multiplying the n observations and then taking their nth root. The Geometric Mean is given by $GM = \sqrt[n]{x_1, x_2, \dots, x_n}$ for ungrouped data, i.e.

$$\log GM = \frac{1}{n}\left(\sum \log x_i\right)$$

So that

GM = Anti log of $\frac{1}{n}(\sum \log x_i)$

GM is very easy to work with logarithm, since logarithms of a product equals sum of logarithms

Example 4.11: Find the geometric mean of data set 5, 3, 2. 1, 4.

$$GM = \sqrt[5]{5 \times 3 \times 2 \times 1 \times 4} = \sqrt[5]{120} = 2.61$$

Note that, if the frequencies of $x_1, x_2, \dots, x_n$ with respective $f_1, f_2, \dots, f_k$, the G.M is then given by

$GM = \sqrt[\Sigma f]{x_1^{f_1}, x_2^{f_2}, \dots, x_k^{f_k}}$, for grouped data

Working with logarithms

$$Antilog\ \frac{1}{\sum f} \sum f_i \log x_i \, , i = 1,2,3, \dots, k$$

Example 4.12: Find the Geometric Mean of the data in example 4.10 above.

Solution:

x_i	f_i	$\log x_i$	$f_i \log x_i$
155	8	2.1903	17.5224
165	17	2.2175	37.6995
175	12	2.2430	26.9160

185	10	2.2672	22.6720
195	3	2.2900	6.8700
-	**50**	**Total**	**111.6799**

:. G.M = Anti-log $\frac{1}{50}$

The following relationship exists among all the three types of means discussed above.

$$HM \leq GM \leq AM$$

4.3 The Median

This is simply defined as the middle number when a set of data were arranged in the order of magnitude. The arrangement can be in ascending or descending order That is, the arrangement can be from lowest number to the highest or from the highest to the lowest number.

Example 4.13. Find the median of the following set of observations:

(a) 4, 5, 6, 12, 10, 8, 7

(b) 24, 32, 18, 14, 15, 17, 21, 10, 15, 12

Solution:

Arrangements in ascending order are as follows

4, 5, 6, 7, 8, 10, 12

For n odd numbers

$$Median\ is\ \ \frac{(n+1)^{th}}{2} = \frac{(7+1)^{th}}{2} = 4th\ position$$

The observation in the 4^{th} position is median is 7

Alternatively, if the arrangement is in descending order,

12, 10, 8, 7, 6, 5, 4

Median = 7

(b) Arranging in ascending order

10, 12, 14, 15, 15, 17, 18, 21, 24, 32

$$Median = \frac{15+17}{2} = 16$$

4.3.1 Median of Grouped Data

The computation of median of grouped data is given by

$$Median = L_{me} + \left(\frac{\frac{N}{2} - Cf_{bf}}{f_{me}}\right) \times C$$

where,

f_{me} = Frequency of the median class

L_{me} = Lower class boundary of the median class

Cf_{bf} = Cumulative frequency before the frequency of the median class

C = Class interval of the median class

N = Summation of frequency

Example 4.14: The following Table shows the duration of unemployment and the number of employed people in a certain developing country in 1986;

Table 19

Duration in Weeks	**Number of unemployment**
< 1	71
1-2	65

2-3	49
3-4	31
4-8	74
8-13	68
13-26	93
26-39	47
39-52	25
>52	132

Find the median?

Solution:

Duration in Weeks	**Number of Unemployment**	$\boldsymbol{cum.frequency\ (cf)}$
0-1	71	71
1-2	65	71+65 = 136
2-3	49	136+49 = 185
3-4	31	185+ 31 = 216
4-8	74	216+74 = 290
8-13	68	290+68 = 358
13-26	93	358+93 = 451
26-39	47	451+47 = 498
39-52	25	498+25 = 523
52-65	132	523+132 = 655

$$Median = L_{me} + \left(\frac{\frac{N}{2} - Cf_{bf}}{f_{me}}\right) \times C$$

where,

L_{me} = lowest class boundary before the median class

Cf_{bf} = cumulative frequency before the median class

f_{me} = frequency of the median class

C = class size

$\frac{N}{2} = \frac{655}{2} = 372.5,$ this will fall in the class interval $8 - 13$

Hence,

$L_{me} = 8, C = 13 - 8 = 5, Cf_{fb} = 290, f_{me} = 68$

Median

$$= 5 + \left(\frac{327.5 - 290}{68}\right) \times 5$$

$$= 5 + \left(\frac{37.5}{68}\right) \times 5$$

$$= 7.76$$

Example 4.15:

A certain clinic reported the following data on ages of women who had abortions during a given period.

Table: 4.7

Ages	15-17	18-19	20-21	22-23	24-25	26-30
Frequency	20	25	20	8	6	13

You are required to determine the:

i. Harmonic mean

ii. Median

Solution:

Age	f	x	$\frac{f}{x}$	cf
15-17	20	16.0	1.25	20
18-19	25	18.5	1.35	45
20-21	20	20.5	0.96	65
22-23	8	22.5	0.36	73
24-25	6	24.5	0.24	79
26-30	13	28.0	0.46	92
Total	92	-	4.62	-

$$i.\quad Harmonic\ Mean = \frac{\sum f}{\sum \frac{f}{x}} = \frac{92}{4.62} = 19.91$$

$$ii.\quad Median = L_{me} + \left(\frac{\frac{N}{2} - cf_{bme}}{f_{me}}\right)C_{me}$$

$f_{me}, L_{mc}, C_m\ etc.$

where,

$$\frac{N}{2} = \frac{92}{2} = 46 \quad ; f_{me} = 20, L_{me} = 19.5 \quad C_{me} = 21.5 - 19.5 = 2, Cf_{bme} = 45$$

$$Median = 19.5 + \left(\frac{46 - 45}{20}\right) \times 2$$

$$= 19.5 + 0.1$$

$$= 19.6$$

4.4 The Mode

The mode of a set of data $x_1, x_2, \dots, x_n$ is that value which has the highest frequency or that value that occurs most frequently. The mode may not exist, even if it exists, it may not be unique.

Example 4.16: Find the mode of the following set of data;

(a) 15, 17, 18, 11, 19, 15, 18, 15, 10, 12, 25

(b) 2, 3, 4, 7, 0, 2, 0, 0, 3, 3, 7, 3, 3, 2, 2, 2

Solution:

(a)

X	*f*
10	1
11	1
12	1
15	3
17	1
18	2
19	1
25	1

Here, mode = 15, this is because it appears 3 times. It is unimodal.

(b)

x	*f*
0	3
2	5
3	5
4	1
7	2

Here, modes are 2 and 3. This is called bimodal

Example 4.17: The mean of seven numbers is 10. If the six of the numbers are 2, 4, 8, 14, 16 and 18, find the mode.

$$Mean\ \bar{X} = \frac{\sum X}{n}$$

$$10 = \frac{2 + 4 + 8 + 14 + 16 + 18 + x}{7}$$

$$70 = 62 + x$$

$$x = 70 - 62 = 8$$

The set of values are 2,4,8,14,16,18,8

The mode is 8

4.4.1 Mode of grouped data

The mode of a set of data $x_1, x_2, \dots, x_n$ with corresponding frequencies $f_1, f_2, \dots, f_n$ is given by

$$Mode = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times C_{mo}$$

where,

L_{mo} = lower class boundary of modal class

Δ_1= the difference between the frequency of the modal class and that of the frequency just immediately before the frequency of the modal class.

Δ_2= the difference between the frequency of the modal class and that of the frequency just immediately after the frequency of the modal class.

C_{mo} = class size or width of the modal class

Example 4.18: The table below gives the distribution of mark (%) obtained by 80 students in Quantitative Techniques in Management of Professional Examination in 2020.

Table 20

Marks	**Frequency**
10-19	7
20-29	2
30-39	12
40-49	20
50-59	13
60-69	10
70-79	6

Determine the mode

Solution:

$$Mode = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times C_{mo}$$

where,

$$L_{mo} = 40 - 0.5 = 39.5, \qquad C_{mo} = 49.5 - 39.5 = 10$$

$$\Delta_1 = 20 - 12 = 8, \qquad \Delta_2 = 20 - 13 = 7$$

$$Mode = 39.5 + \left(\frac{8}{8+7}\right) \times 10$$

$$= 39.5 + \frac{80}{15} = 39.5 + 5.33$$

$$= 44.83$$

Travel claims (₦)	Number of claims
150-160	4
160-170	8
170-180	12
180-190	6
190-200	5

Calculate the modal travel claim from the table above

$$Mode = 170 + \frac{4}{4+6} \times 10 = ₦174$$

However, the relationship that exists between mean, median and mode is that

$$Mean - Mode = 3(Mean - Median)$$

4.5 Measures of Positional values

Measures of Positional values are;

(i) Quartiles

(ii) Deciles

(iii) Percentiles.

These measures are naturally extensions of the median instead of dividing the set of data into two equal parts, those below and those above the median; we may divide it into three, four or more parts.

If the division into three parts, there will be two values such that one third of the data lie below one of them and two-third lie below the other. Similarly, if the data are divided into four equal parts, we will have three will values. One quarter of the data will lie below one of them, half will lie below another one which we have called the median and three quarters will lie below the last one.

4.5.1 Quartiles

Quartiles divide a distribution into four equal parts. The two quartiles are usually called the lower quartile (Q_1) while (Q_2) is the same as the median.

In an ungrouped set of data,

$$Q_1 \; is \; the \; \left(\frac{N}{4}\right)^{th} \; item$$

$$Q_2 \; is \; \left(\frac{2N}{4}\right)^{th} \; item$$

$$Q_3 \; is \; the \; \left(\frac{3N}{4}\right)^{th} \; item$$

But for a grouped set of data

$$Q_1 = \left(\frac{N}{4}\right)^{th} \; position$$

$$Q_3 = \left(\frac{3N}{4}\right)^{th} \; position$$

$$Q_2 \; is \; \left(\frac{2N}{4}\right)^{th} \; position$$

Example 4.19: The following table indicates the marks scored by 100 students in a Business Management of a professional examination.

Table 21

Marks	Number of students
25-34	10
35-44	3
45-54	30
55-64	34
65-74	8
75-84	15

Determine the following:

(a) Q_1

(b) Q_3

(c) Quartile Deviation (QD)

(d) Interquartile Range

(e) Semi-Interquartile Range.

Solutions:

Mark	F	Cf
25 – 34	10	10
35 – 44	3	10+3 = 13
45 – 54	30	13+30 =43
55 – 64	34	43+34 =77
65 -74	8	77+8 = 85
75 -84	15	85+15 =100
-		

$$(a)Q_1 = L_{q1} + \left(\frac{\frac{N}{4} - C_{fbq1}}{f_{q1}}\right) \times C_{q1}$$

$$\frac{N}{4} = \frac{100}{4} = 25$$

$L_{q1} = 44.5, C_{q1} = 54.5 - 44.5 = 10$

$Cf_{fbq1} = 13\ f_{q1} = 30$

$$Q_1 = 44.5 + \left(\frac{25 - 13}{30}\right) \times 10$$

$$= 44.5 + \left(\frac{12}{30}\right)\ x\ 10$$

$= 44.5 + 4$

$= 48.5$

$$(b) Q_3 = L_{q3} + \left(\frac{\frac{3N}{4} - C_{fbq3}}{f_{q3}} \right) \times C_{q3}$$

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

$L_{q3} = 54.5, C_{q3} = 64.5 - 54.5 = 10$

$cf_{bq3} = 34, \quad fb_3 = 43$

$$Q_3 = 54.5 + \left(\frac{32}{34}\right) \times 10$$

$$= 54.5 + 9.41$$

$$= 63.91$$

$$(c) \quad Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{63.91 - 48.5}{63.91 + 48.5}$$

$$= \frac{15.41}{112.41}$$

$$= 0.1371$$

(d) Interquartile $= Q_3 - Q_1$

$= 63.91 - 48.5$

$= 15.41$

(e) S.I.Q.R. $= \frac{Q_3 - Q_1}{2}$

$= 7.705$

Example 4.20: The following table shows the weekly profit of 50 randomly selected small scale enterprises in some West African cities in thousands of Naira:

Table 22

Profit (N'000)	**Number of enterprises**
160 – 179	6
180 – 199	10
200 – 209	20

210 – 219	14

Calculate the third quartile?

Solution:

Class	***f***	***cf***
160-179	6	6
180-199	10	16
200-209	23	36
210-219	14	50
	50	

$$\frac{3}{4}N = \frac{3}{4} \times 50 = 37.5$$

$$Q_3 = 199.5 + \frac{37.5 - 36}{14} \times 10$$

$$= 199.5 + 1.07$$

$$= 200.57$$

4.5.2 Deciles

This is defined as the nine points that divide a set of observations into ten equal parts.

Example 4.21: Given that $N = 40$ in an ungrouped data find

(a) 8th deciles

(b) 6th deciles

(c) 9th deciles

(d) 4th deciles

Solution: Since $N = 40$

(a) 8th deciles $= D_8 = \left[\frac{8}{10} \times 40\right]^{th}$ item or position

= 32th position

(a) 6th deciles $= D_6 = \left[\frac{6}{10} \times 40\right]^{th}$ item or position

= 24th position

(c) 9th deciles $= D_9 = \left[\frac{9}{10} \times 40\right]^{th}$ item or position

= 36^{th} position

(d) 4th deciles D_9 items or position

= 16th position

In general, the i^{th}deciles is defined as $\left(\frac{i}{10} \times N\right)^{th}$ item or position

For grouped data, i^{th}deciles is given by

$$D_{i^{th}} = L_{i^{th}} + \left[\frac{\frac{Ni^{th}}{10} - Cf_{bdi^{th}}}{f_{di^{th}}}\right] \times Cdi^{th}$$

Example 4.22: Find the 7^{th} deciles in the data in example 4.19 above.

Solution:

$$D_7 = L_{d7} + \left[\frac{\frac{7N}{10} - Cf_{bd7}}{f_{d7}}\right] \times Cd_7$$

Here,

$$\frac{7}{10} \times 100 = 70,$$

$$L_{d7} = 54.5$$

$$C_{d7} = 64.5 - 54.5 = 10$$

$$C_{fbd7} = 43$$

$$f_{d7} = 34$$

$$D_7 = 54.5 + \left[\frac{70 - 43}{34}\right] \times 10$$

$$= 54.5 + \frac{270}{34}$$

$$= 54.5 + 7.94$$

$$= 62.44$$

4.6 Percentile

This is defined as the 99 points that divide a set of data 100 equal parts, for example, 60^{th} percentile of N = 150 is given by

$$60^{th} = \left[\frac{60}{200} \times 150\right] th$$

= 90th Position

For grouped data, the ith percentile is given by

$$ith\ percentile = L_{p_{ith}} + \left[\frac{\frac{ith}{100} - Cf_{bpith}}{f_{ith}}\right] \times 5$$

Example 4.23: The table below shows the distribution of the electorates of a number of adults in elections.

Table 23

Electorates	Number of Adults
45-50	2
50-55	8
55-60	12
60-65	10
65-70	5
70-75	8

Determine the

(a) 30th percentile

(b) 55thpercentile

(c) 75th percentile

(d) 25th percentile.

Solution:

Electorates	Number of Adults	cf
40-45	2	2
45-50	8	2+8=10
50-55	12	10+12=22
55-60	10	12+10=32
60-65	5	32+5=37
65-70	8	37+8=45

$$(a)P_{30} = L_{p_{ith}} + \left[\frac{\frac{ith}{100} - Cf_{bpith}}{f_{ith}}\right] \times 5$$

Where,

$$\frac{30}{100}N = \frac{30}{100} \times 45 = 13.5, \quad LP_{30} = 50 \quad Cp_{30} = CP_{30} = 5, \quad Cfbp_{30} = 10,$$

$$fbP_{30} = 12$$

$$P30 = 50 + \left(\frac{13.5 - 10}{12}\right) \times 5$$

$$P_{30} = L_{p30} + \left(\frac{\frac{30}{100}N - Cf_{bp30}}{f_{bp30}}\right) \times 5$$

$$50 + \left(\frac{3.5}{12}\right) \times 5$$

$$= 50 + \frac{17.5}{12} = 51.458$$

$$(b)P_{55} = LP_{55} + \left[\frac{\frac{55N}{100} - CF_{bp55}}{fp_{55}}\right] \times 5$$

$$\frac{55}{100}N = \frac{55}{100} \times 45 = 24.75, \qquad LP_{55} = 55 \;\; Cp_{55} = 60 - 55, \;\; Cfbp_{55} = 22,$$

$$fbp_{55} = 10$$

$$P_{55} = 55 + \left[\frac{24.75 - 22}{10}\right] \times 5$$

$$= 55 + \left(\frac{2.75 \times 5}{10}\right)$$

$$= 55 + 1.375$$

$$= 56.375$$

$$(c)p_{75} = LP_{75} + \left[\frac{\frac{75N}{100} - Cf_{pb75}}{fp_{75}}\right] \times 5$$

$$\frac{75}{100}N = \frac{55}{100} \times 45 = 33.75,$$

$$LP_{75} = 60$$

$$Cp_{75} = 65 - 60 = 5$$

$$Cfbp_{75} = 32$$

$$fbp_{75} = 5$$

$$P_{75} = 60 + \left(\frac{33.75 - 32}{5}\right) \times 5$$

$$P_{75} = 60 + 1.75 = 61.75$$

$$(d)P_{25} = LP_{75} + \left[\frac{\frac{25N}{100} - Cf_{pb25}}{fp_{25}}\right] \times 5$$

Where,

$$\frac{25}{100}N = \frac{25}{100} \times 45 = 11.25,$$

$LP_{25} = 50$

$Cp_{25} = 55 - 50 = 5$

$Cfbp_{25} = 10,$

$fbp_{25} = 12$

$$P_{25} = 50 + \left(\frac{11.25 - 10}{12}\right) \times 5$$

$$= 50 + \frac{1.25 \times 5}{12}$$

$= 50 + 0.520$

$= 50.52$

Note the following relationship between quartiles, deciles and percentiles.

(i) $Q_{25} = D_{2.5} = P_{25}$

(ii) $D_2 = Median = D_5 = P_{50}$

(ii) $Q_3 = D_{7.5} = P_{75}$

Example 4.24: The following table gives the distribution of scores obtained by 40 students in an examination.

Table 24

Scores (%)	No. of students
10-19	5
20-29	7
30-39	10
40-49	12
50-59	4
60-69	2
Total	40

Required:

(a) Determine the mean deviation of the student's scores

(b) Calculate the coefficient of Skewness using the two common methods

Solution:

Scores	f	x	fx	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$	$f(x-\bar{x})^2$
10-19	5	14.5	72.5	22.25	111.25	2475.3125
20-29	7	24.5	171.5	12.25	85.75	1050.4375
30-39	10	34.5	345	2.25	22.5	50.625
40-49	12	44.5	534	7.75	93	720.75
50-59	4	54.5	218	17.75	71	1260.25
60-69	2	64.5	129	27.75	55.5	1540.125
	40		1470		439	7097.5

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1470}{40} = 36.75$$

$$S.D = \sqrt{\frac{7097.5}{40}}$$

$$= 84.25$$

$$MD = \frac{\sum f\,\lvert x-\bar{x}\rvert}{\sum f}$$

$$= \frac{439}{40}$$

$$= 10.975$$

Mode is

$$39.5 + \frac{(12-10)}{(12-10)+(12-4)}(49.5-39.5)$$

$$= 39.5 + \frac{2}{10}(10)$$

$$= 39.5 + 2$$

$= 41.5$

Median =

$$= 39.5 + \frac{(20 - 12)}{10} \times 10$$

$$= 39.5 + 8$$

$$= 47.5$$

(b) Method 1

Coefficient of Skewness

$$= \frac{\bar{x} - mode}{S.D} = \frac{36.75 - 41.5}{84.5}$$

$$= -0.0564$$

Method 2

Coefficient of Skewness

$$= \frac{3(\bar{x} - median)}{S.D}$$

$$= \frac{36.75 - 41.5}{84.5} = -0.3828$$

CHAPTER 5
MEASURES OF VARIABILITY

5.1 Learning Objectives:

At the end of this chapter, the students should be able to;

i. judge the variability of measures of central tendency
ii. make a comparative study of the variability of two series
iii. identify the causes of variability with a view to control it
iv. serve as a basis for further statistical analysis

5.2 Introduction

Measures of variability otherwise called measures of dispersion help us to determine the extent to which the figures of the distribution are clustered together or well spread out from the central values. Among the measures of dispersion are:

(i) Range

(ii) Mean Absolute Derivation (MAD)

(iii) Variance

(iv) Standard Deviation

(v) Inter-quartile range

(vi) Coefficient of variation

(vii) Skewers

(viii) Kurtosis

The measures are as briefly discussed as follows:

5.2.1 The Range

This is defined as the differences between the maximum value and the minimum values in a set of observation. For example, if data range from 86 to 42, then Range = 86 - 42 = 44. Also, if the data set values are: 12,10,22,45,8,26, the Range (R) = 45 – 8 = 37.

5.2.2 The Mean Deviation

The mean deviation of set of observation $x_1, x_{2,L}, x_n$ from the mean is the average of the deviations of the set of observations from the mean. The mean deviation is obtained by ignoring the signs of the deviations and simply regarding all of it as positive. The mean of these absolute deviations is the mean deviation. It deviation $\mathrm{x} - \bar{\mathrm{x}}$, of a set of observation from the mean ($\bar{x}$) is regarded as positive written as $|\mathrm{x} - \bar{\mathrm{x}}|$absolute deviation.

Hence, the mean deviation of observations $x_1, x_2, x_3, \ldots, x_n$is given by

$$\mathrm{MD} = \frac{\sum_{i=1}^{n} |\mathrm{x} - \bar{\mathrm{x}}|}{\mathrm{n}} \text{ where, } \bar{\mathrm{x}} = \frac{\sum_{i}^{n} \mathrm{x}}{\mathrm{n}} \text{, for ungrouped data.}$$

Example 5.1: Given the observation 14, 20, 6, 10, 22 .Obtain the mean deviation

Solution:

$$\overline{x} = \frac{\sum_{i=1}^{n} x}{n} = \frac{14+18+6+10+22}{5} = \frac{70}{5} = 14.$$

x	x-14	$\lvert x-14 \rvert$
14	0	0
18	4	4
6	-8	8
10	-4	4
22	8	8
-	-	24

$$\text{MAD} = \frac{\sum_{i=1}^{n} \lvert X - \bar{X} \rvert}{n} = \frac{24}{5} = 4.8$$

Mean Deviation of Grouped Data:

$$\text{MD} = \frac{\sum f \lvert x - \bar{x} \rvert}{\sum f}$$

Example 5.2.

Consider the table below:

Table 25

Incomes per day (Naira)	No. of casual workers
40-60	50
60-80	70

80-100	90
100-120	110
120-140	130
140-160	150

Calculate its mean deviation.

Solution:

x	F	Fx	$f\,|x - 104|$
50	60	300	3240
70	70	4900	2380
90	85	7650	1190
110	110	12100	660
130	95	12350	2470
150	80	12000	3680
-	500	52000	13620

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{52000}{500} = 104.$$

$$MD = \frac{\sum f\,|x - \bar{x}\,|}{\sum f}$$

$$= \frac{13620}{500}$$

$$= 27.24$$

This means that, the average difference between the daily incomes paid to these casual workers and the mean income is₦27.

5.2.3 Variance and Standard Deviation

The variance of a set of number observation is the arithmetic means squared deviation. When a finite population data is being used, this, mean square of (deviation) is referred to as the population variance denoted by σ^2 .when sample data is used, we commonly do, sample variance is obtained by dividing the sum of squared deviations by (n-1) not n.The sample variance is denoted by s^2. Suppose we have a set of observations $x_1, x_2, x_3, \ldots, x_n$and their arithmetic mean ($\bar{x}$), then the variance denoted by σ^2 is given by

$$\sigma^2 = \frac{\sum_i^n (x - \bar{x})^2}{n}, for\ a\ variance\ with\ finite\ population\ n.$$

$$And,\ s^2 = \frac{\sum_i^n (x - \bar{x})^2}{n-1}, variance\ for\ a\ sample\ of\ size\ n.$$

Also, suppose further that the set of data $x_1, x_2, x_3, \ldots, x_n$, have respective corresponding frequencies $f, f_2, f_3, \ldots, f_n$, then,

$$\sigma^2 = \frac{\sum_i^n f(x - \bar{x})^2}{\sum f}, for\ a\ variance\ with\ finite\ population\ n.$$

$$And,\quad s^2 = \frac{\sum_i^n f(x - \bar{x})^2}{\sum f - 1}, variance\ for\ a\ sample\ of\ size\ n.$$

5.2.4 Standard Deviation: The standard deviation is defined as the square-root of the variance. That is, it implies $\sqrt{\operatorname{var} iance}$.

Example 5.3: Obtain the variance and Standard deviation of tables in examples 5.1 and 5.2 above.

x	$(x-\bar{x})$	$(x-\bar{x})^2$
14	0	0
18	4	16
6	-8	64
10	-4	16

22	8	64
Total = 70	-	Total = 160

$$\text{Variance}(\sigma^2) = \frac{\sum_i^n (x - \bar{x})^2}{n}$$

$$= \frac{160}{70}$$

$$= 2.386$$

Hence, standard deviation $= \sqrt{2.386} = 1.512$

In example 5.2

x	f	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
50	60	2916	145800
70	70	1156	80920
90	85	196	17640
110	110	36	3960
130	95	676	64220
150	80	2116	169280
Total	500	Total	481820

Therefore, $\text{Variance}(\sigma^2) = \frac{481820}{500}$

$$\text{Standard deviation} = 31.042$$

Example 5.4: The mark of 100 applicants in an aptitude test in a university is stated in the Table below:

Table 26

Marks	40 -44	45 -49	50 -54	55 -59	60 -64	65 -69	70-74	75 -79
Applicants	3	9	12	16	24	24	9	3

Calculate:

(i) Variance.

(ii) Standard Deviation.

Solution:

Marks	x	f	fx	$(x-\overline{x})$	$(x-x)^2$	$f(x-\overline{x})^2$
40-44	42	3	126	-18.6	345.96	1037.88
45-49	47	9	423	-13.6	184.96	1664.64
50-54	52	12	624	-8.6	73.96	887.52
55-59	57	16	912	-3.6	12.96	207.36
60-64	62	24	1488	1.4	1.96	47.04
65-69	67	24	1608	6.4	40.96	983.04
70-74	72	9	648	11.4	129.96	1169.64
75-79	77	3	231	16.4	268.96	806.88
Total	.	100	6060	-	-	6804

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{6060}{100} = 60.6 \qquad And, \qquad s^2 = \frac{\sum_i^n f(x - \bar{x})^2}{\sum f - 1}$$

$$= \frac{6804}{499} = 13.64$$

Hence, Standard deviation = √13.64 = 3.6926

5.2.5 Coefficient of Variation

This is defined as 100 times the ratio of the standard deviation to the mean. Coefficient of variation (C.V) is mathematically represented by

$$\mathrm{C.V} = \frac{\sigma}{\bar{x}} \text{ x } 100$$

Variance is an absent concept that increases with the amount of the variability. The coefficient of variation is useful in comparing dispersion in data sets, that have marked by different means and comparing dispersion in set data.

Example 5.5: Find the coefficient of variation of data set 1, 2, 3, 4, 5.

Solution:

x	x- x	$(\mathbf{x} - \bar{\boldsymbol{x}})^2$
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4
Total	-	10

$$\text{Variance}(\sigma^2) = \frac{\sum_i^n (x - \bar{x})^2}{n} = \frac{10}{5} = 2$$

$\sigma = \sqrt{2} = 1.414$

$$\text{Hence, } \mathrm{C.V} = \frac{\sigma}{\bar{x}} \text{ x } 100 = \frac{1.414}{3} = 47.13\%$$

Example 5.6: Consider the Table below:

Table 27

No. of calls.	frequency
0	10
1	26
2	34

3	42
4	60
5	48

Obtain the following:

(a) The mean.

(b) Standard deviation.

(c) Coefficient of variation.

Solution:

x	f	fx	$(\mathbf{x}-\bar{\boldsymbol{x}})^2$	$\mathbf{f}(\mathbf{x}-\bar{\boldsymbol{x}})^2$
0	10	0	10.1124	101.124
1	26	26	4.7524	123.5624
2	34	68	1.3924	47.3416
3	42	126	0.0324	1.3608
4	60	240	0.6724	40.344
5	48	240	3.3124	158.9952
Total	**220**	**700**	**-**	**472.728**

(a) $\bar{x} = \frac{\sum fx}{\sum f} = \frac{700}{220} = 3.18$

(b) Standard deviation:

$$= \sqrt{\sigma^2} = \sqrt{\frac{\sum_i^n f(x-\bar{x})^2}{\sum f}} = \sqrt{\frac{472.728}{220}} = \sqrt{314.4555} = 17.73$$

.

(c) Coefficient of variation:

$$\mathrm{C.V} = \frac{\sigma}{\bar{\mathrm{x}}} \times 100 = \frac{17.73}{3.18} \times 100 = 557.55\%$$

Example 5.7: Given that, the mean and standard deviation of bullocks are 1097.5171 and 90.34 respectively, the mean and standard deviation of sheep are 261.15 and 47.75 respectively. Determine which of the two are more variable in weight.

Solution:

Bullocks;$\bar{x}$ = 1097 .5171, σ =90.34

$$\text{C.V} = \frac{\sigma}{\bar{x}} \text{ x } 100 = \frac{90.34}{1097.5171} = 8.2315$$

Sheep; $\bar{x}$ =261 .15, σ = 47 .75

$$\text{C.V} = \frac{\sigma}{\bar{x}} \text{ x } 100 = \frac{47.75}{261.15} \text{ x } 100 = 18.25\%$$

Hence, weights of sheep are more variable, since the bigger the variance, the wider the variation.

5.2.6 Skewness and Kurtosis

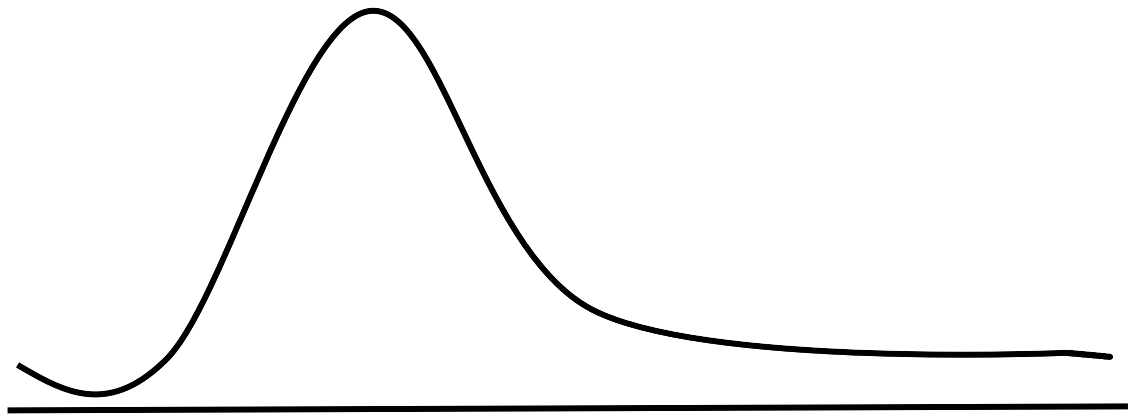
The shape of a frequency polygon describes the quantitative features of distribution. The graphical structures of a distribution are commonly referred to as skewness and kurtosis. Furthermore, both skewness and kurtosis are dimensionless.

5.2.6.1 Skewness

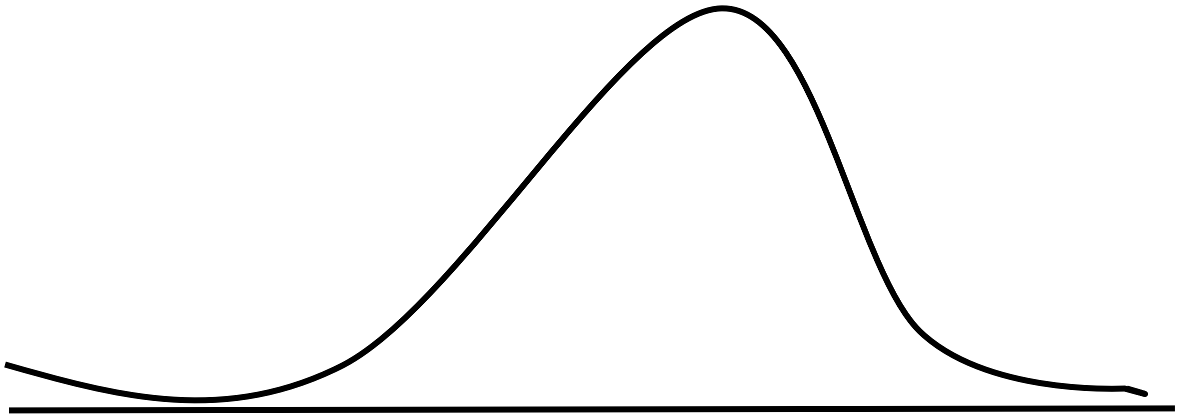
This simply means lack of symmetry in a distribution. skewness is a measure of a symmetry or departure from a symmetry. If more observation lies on one side of the mean than the other side, the distribution is said to be skewed.

Positive skewness: This means if more observations lie to the left of the mean, the distribution has a high tail goes to the right.

This is Positive skewness.

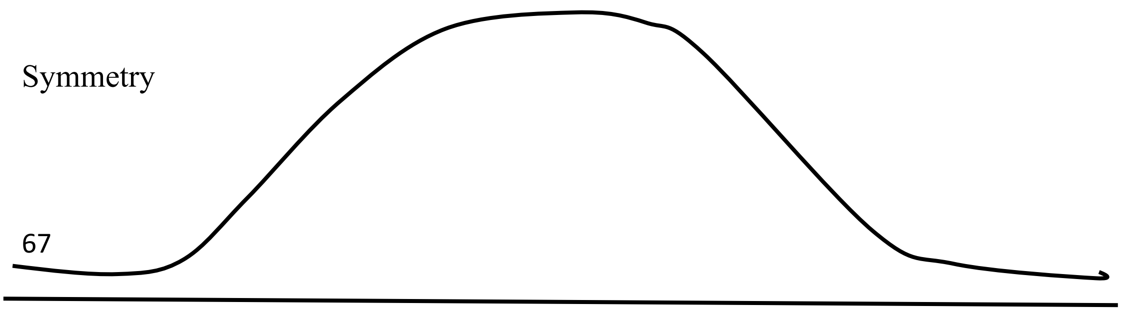


Negative skewness: This means if more observations lie to the right of the mean, the longer tail goes to the left. That is



In a symmetrical distribution, the mean, the median and the mode which are measures of central tendency coincide.

Symmetry



The following are the two most commonly measures of Skewness.

(i) Skewness = $\frac{3(\text{ Mean} - \text{Mode})}{\text{Standard Deviation}}$

(ii) Pearson's coefficient of skewness = $\frac{3(\text{ Mean} - \text{Mode})}{\text{Standard Deviation}} = \frac{3(\text{ Mean} - \text{Median})}{\text{Standard Deviation}}$

The Pearson's coefficient of Skewness lies between negative and positive values.

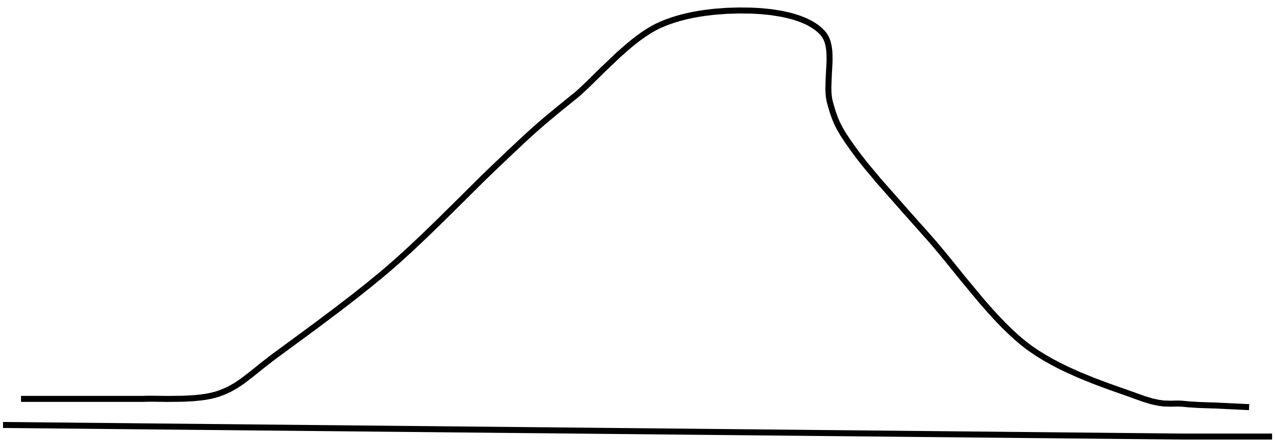
Example 5.8: In a hypothetical data given that mean is - 93, median = 93.05 and standard deviation is 16.13, compute the Pearson's coefficient of Skewness.

$$\text{Solution: } = \frac{3(\text{Mean} - \text{Mode})}{\text{Standard Deviation}} = \frac{3(-93-93.5)}{16.13} = -34.69$$

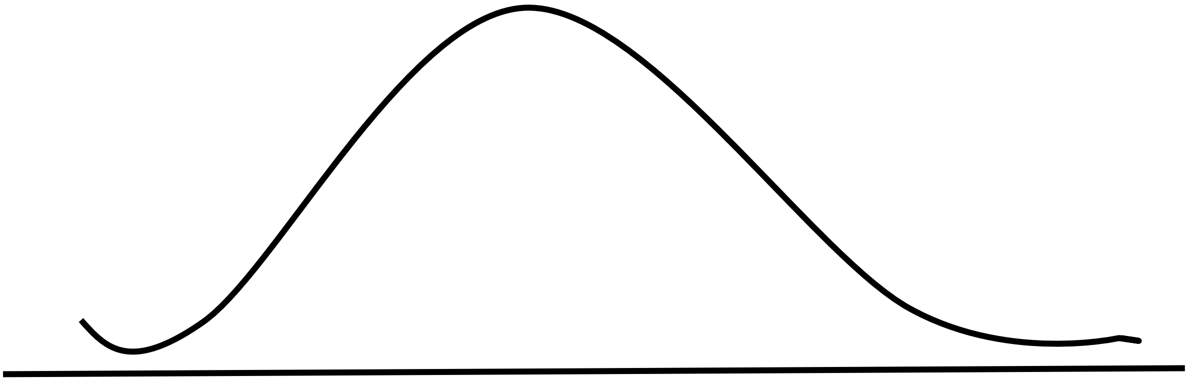
5.2.6.2Kurtosis

The term "Kurtosis" refers to the statistical measure that describes the shape of either tail of a distribution, i.e., whether the distribution is heavy-tailed (presence of outliers) or light-tailed (paucity of outliers) compared to a normal distribution. There are three identifiable kurtosis curves.

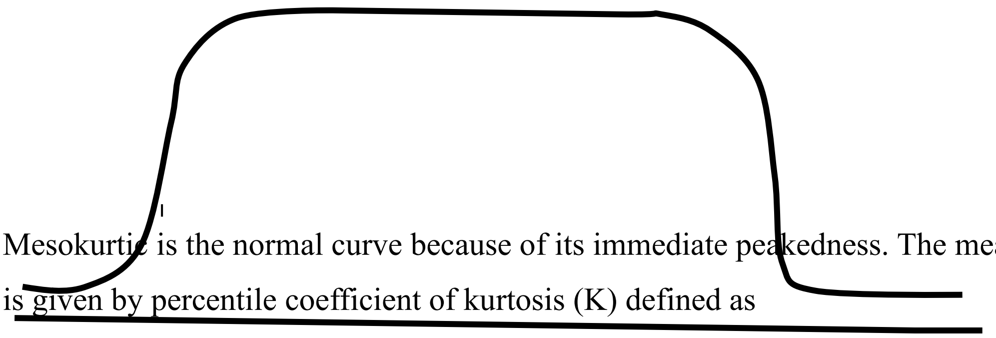
(i) Leptokurtic or heavy tailed distribution



(ii) mesokurtic



(iii) platykurtic or short tailed distribution



Mesokurtic is the normal curve because of its immediate peakedness. The measure of kurtosis is given by percentile coefficient of kurtosis (K) defined as

$$K = \frac{\frac{1}{2}(Q_3 - Q_1)}{p_{90} - p_{10}} \text{ or}$$

$$K = \frac{\sum \frac{(x-\bar{x})^4}{n}}{s^4}$$

Example 5.9: Below is a table showing the annual turn-over of branches of a supermarket in Nigeria.

Table: 5.4

Annual Turn-over (₦'000)	**Branch of Supermarket**
0 and under 10	6
10 and under 20	10
20 and under 30	8
30 and under 40	4
40 and under 50	5
50 and under 60	7
60 and under 70	9
70 and under 80	11

From the above Table, you are required to

i. Compute the coefficient of variation

ii .Determine the inter-quartile range.

Solution:

class	Frequency (f)	x	fx	$(X-\bar{X})^2$	$f(X-\bar{X})^2$	cf
$0-10$	6	5	30	1393.5289	8361.1734	6
$10-20$	10	15	150	746.9289	7469.2890	16
$20-30$	8	25	200	300.3289	2402.6312	24
$30-40$	4	35	140	53.7289	214.9156	28

40 – 50	5	45	225	7.1289	35.6445	33
50 – 60	7	55	385	160.5289	1123.7023	40
60 – 70	9	65	585	513.9289	4625.3601	49
70 – 80	11	75	825	1067.3289	11740.6179	60
Total	60	-	2540	-	35973.334	-

$$\text{Mean } (\bar{x}) = \frac{\sum \text{fx}}{\sum \text{f}} = \frac{2540}{60} = 42.33$$

$$\text{Variance } (\sigma^2) = \frac{\sum_i^n f(x-\bar{x})^2}{\sum f} = \frac{35973.334}{60} = 599.5556$$

$$\text{Standard Deviation (S.D)} = \sqrt{599.5556} = 24.4858$$

Hence, Coefficient of variation is given by,

(i) $\text{C.V} = \frac{\sigma}{\bar{x}} \text{ x } 100 = \frac{24.4858}{42.33} = 57.85\%$

a(ii) $Q_1 = L_{q_1} + \left(\frac{\frac{N}{4} - cfb_{q_1}}{fq_1}\right)C$

Where, Q_1= The first quartile.

L_{q_1}= lower class boundary of the first quartile class.

cfb_{q_1}= cumulative frequencies before the first quartile class.

fq_1 = frequency of the first quartile class.

C = class size.

N = total frequencies.

Solution:

First quartile class = N/4 = 60/4 =15. I.e, class [10 - 20].

L_{q_1}= [10 + 10]/2 = 20/2 = 10.

cfb_{q_1}= 6.

fq_1 = 10.

C = 20 - 10 or 30 - 20 = 10.

N =60.

Hence,

$$Q_1 = 10 + \left(\frac{\frac{60}{4} - 6}{10}\right)10 = 19.$$

$$Q_3 = L_{q_3} + \left(\frac{\frac{3N}{4} - cfb_{q_3}}{fq_3}\right)C$$

Where, Q_3= The third quartile.

L_{q_3}= lower class boundary of the third quartile class.

cfb_{q_3}= cumulative frequencies before the third quartile class.

fq_3 = frequency of the third quartile class.

C = class size.

N = total frequencies.

Third quartile class = 3N/4 = 3(60/4) =3(15)=45. I.e, class [40 - 50].

L_{q_3}= [40 + 40]/2 = 80/2 = 40.

cfb_{q_3}= 28.

fq_3 = 5.

C = 20 - 10 or 30 - 20 = 10.

N =60.

$$Q_3 = 40 + \left(\frac{45 - 28}{5}\right)10 = 74.$$

$$Q_3 - Q_1 = 74 - 19 = 55.$$

Hence, the interquartile range is 55.

CHAPTER 6

PROBABILITY AND PROBABILITY RULES

6.1 Learning Objectives:

At the end of this chapter, the students should be able to;

i. explain the concept of probability;
ii. calculate the probability of simple events;
iii. calculate the probability of compound events;
iv. calculate the probability of complementary events; and
v. apply basic probability concepts.

6.2 Permutation and Combination

6.2.1 Permutation: This is a process of arrangement whereby the order of arrangement is taking into consideration. Thus, the permutation of n-objects taken r at a time is the number of arrangement possible. It is denoted by np_r.For instance, under the four letters a,b, c, d. In how many ways can you pick two from this letters?

Solution

a	b	c	d		
ab	ac	ad	bc	bd	cd
ba	ca	da	cb	db	dc

=12 ways.

Also, if 3 letters are taken at a time we then have the following arrangement.

abc, acd, bcd, cba, dca,dcb

= 6 ways.

In general, the number of permutation of n different object taking r at a time is denoted bynp_r and it is expressed by $np_r = \frac{n!}{(n-r)!}$

Note that n! $= n(n-1)(n-2)\ldots(n-r+)$ and o! $= 1.$

For example, $5! = 5(5-1)(4-1)(3-1)(2-1)1 = 5(4)(3)(2)(1)1. = 120$ ways.

Also, when r = n

$$np_n = \frac{n!}{(n-n)!} = 1$$

However, if we have n items consisting of $n_1, n_2, \ldots, n_k$ similar items, where

$n_1 + n_2 + \ldots + n_k = n,$ the number of permutation is $\frac{n!}{n_1!n_2!\text{L } n_k!}$ of possible arrangements.

Example 6.1: In how many ways can the letters of the word “accounting” bearranged ?.

Solution:

a, o,u,t and g occur once while n and c occurs twice.

So that the numbers of permutation or arrangements possible are:

$=\frac{10!}{2!2!1!1!1!1!1!1!}$

$= \frac{10\times9\times8\times7\times6\times5\times4\times3\times2!}{2\times1\times2\times1\times1\times1\times1\times1\times1\times1} = 907200\, ways.$

Example 6.2: In how many ways can the word "MANAGEMENT" be arranged?.**Solution:**

We have 10 letters altogether.

M occurs 2 times.

A AA occurs 2 times.

E occurs 2 times.

G and T occur once .

Arrangement $= \frac{10!}{2!\ 2!\ 2!\ 2!\ 1!\ 1!}$

$= \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{2\times1\times2\times1\times2\times1\times2\times1\times1\times1}$

$= 10\times 9\times 7\times 6\times 5\times 4\times 3 = 226800\, ways.$

Example 6.3: In how many ways can two sets of items with 5 and 3 element be arranged taking all at once.

Solution:

Arrangement = $5 * 3 = 15$ ways.

Example 6.4: How many different arrangements can be done from the word "cylinder" by taking (a) 3 letters at a time.

(b) 5 letters at a time.

(c) All of the letters at a time.

Solution: (a) 3 letters at a time from 8 letters:

$$8p_3 = \frac{8!}{(8-3)!} = \frac{8!}{5!} = \frac{8*7*6*5!}{5\,!} = 8*7*6 = 336\ ways.$$

(b) 5 letters at a time from 8 letters

$$8p_5 = \frac{8!}{(8-5)!} = \frac{8\,!}{3\,!} = \frac{8*7*6*5*4*3\,!}{3\,!} = 8*7*6*5*4 = 6{,}720\ ways.$$

(c) 8 letters at a time taken from 8 letters

$$8p_8 = \frac{8!}{8-8!} = \frac{8\,!}{0\,!} = 8*7*6*5*4*3*2*1 = 40{,}320\ ways.$$

Example 6.5: The Honorable Minister for Education is invited to Induction Ceremony of the Chartered Institute of Personnel Management (CIPM). In how many ways can the Honorable Minister, the President of CIPM and 5 other guests be seated at high Table (in a row) if the Honorable Minister and the President are seated next to each other.

Solution:

The Honorable Minister and the President can be seated next to each other in 2 ways, by exchanging the sittings of 5 other guest., hence, total number of ways is

$$= 2p_2 * 5p_5 = 2*120 = 240\text{ways}.$$

Example 6.6: Compute the following permutations.

(a)$10p_7$ $(b)12p_4$ (c) $20p_2$ $(d)5p_3$

Solution: (a) $10p_7 = \frac{10\,!}{(10-7)!}$

$$= \frac{10!}{7!}$$

$$= \frac{10\times9\times8\times7!}{7!} = 720\ ways.$$

(b) $12p_4 = \frac{12!}{(12-4)!}$

$$= \frac{12!}{8!}$$

$$= \frac{12\times11\times10\times9\times8!}{8!} = 11{,}880\ ways.$$

(c) $20p_2 = \frac{20!}{(20-2\,)!}$

$$= \frac{20!}{18!}$$

$=\frac{20 \times 19 \times 18!}{18!} = 380\ ways$

(d) $5p_3$ $=\frac{5!}{(5-3)!}$

$=\frac{5!}{2!}$

$=\frac{5\times4\times3\times2!}{2!} = 60\ ways.$

Example 6.7: The Chartered Institute of Personnel Management (CIPM) wishes to constitute 3 committee members in the executive. How many ways can the three committee of 3, 2, and 3 are constituted if no member of the executive is to serve on more than one committee?

Solution:

Total = 8

1st committee = 3

2nd committee = 2

3rd committee = 3

This can be done in

$=\frac{8!}{3!\ 2!\ 3!}$

$=\frac{8\times7\times6\times5\times4\times3\times2\times1}{3\times2\times1\times2\times1\times3\times2\times1}$

= 8 x 7 x 5x 2

= 56 x 10

= 560 ways.

Example 6.8: Six members of association and two non –members are to be seated round a table. In how many ways can this be done if each member can sit anywhere?

Solution:

This is done in (8 -1)! ways

$=7!$

$= 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$

$= 5040$ ways.

Combination: This is a process of selection when the order of selection is not important. For instance, consider the four letters a, b, c, and d in permutation. In how many selections can you have by taking 2 letters at a time.

Solution:

a b c d

$ab = ba,$
$ac = ca,$
$ad = da,$
$bc = cb,$
$bd = db,$
$cd = dc.$

$= 6 \textit{ selections}$

In general, selection of n dissimilar items taken $r \leq n$ at once have

$$nc_r = \frac{n!}{r!(n-r)!} = \frac{np_r}{r!}$$

In combination, selection can be done with replacement (WR) or without replacement (WOT).

Example 6.9: A committee of 3 is to be chosen from 3 males and 4 females. In how many ways Can this be done so that the committee includes

(a) at least a male

(b) at least one female

Solution:

(a) 1 male 2 females $=3\,C_1\,.\,4C_2$

2 males 1 female $=3C_2.4C_1$

3 males 0 female $=3C_3.4C_0$

but

$$3\,C_1\,.\,4C_2=\frac{3!}{1!2!}\times\frac{4!}{1!3!}=3\times 6=18$$

$$3\,C_2\,.\,4C_1=\frac{3!}{1!2!}\times\frac{4!}{1!3!}=3\times 4=12$$

$$3\,C_3\,.\,4C_0=1\times 1=1$$

$hence(a)=18+12+1=31$ ways

(b) 1 female 2 males $=4\,C_1\,.\,3C_2$

2 females 1 male $=4\,C_2\,.\,3C_1$

3 females 0 male $=4\,C_3\,.\,3C_0$

But

$$4\,C_1\,.\,3C_2=\frac{4!}{1!3!}\times\frac{3!}{1!2!}=4\times 3=12$$

$$4\,C_2\,.\,3C_1=\frac{4!}{2!2!}\times\frac{3!}{1!2!}=\frac{4\times 3\times 2!}{2\times 1\times 2!}\times\frac{3\times 2!}{1\times 2!}=6\times 3=18$$

$$4\,C_3\,.\,3C_0=\frac{4!}{1!3!}\times\frac{3!}{0!3!}=4\times 3=12$$

Hence (b)

$=12+18+12$

$= 42$ ways

Example 6.10: In how many ways can a committee of five be selected from a group of ten ?

Solution:

$$10C_5 = \frac{10!}{5!5!}$$

$$= \frac{10\times 9\times 8\times 7\times 6\times 5!}{5!\times 5\times 4\times 3\times 2\times 1}$$

$$= 9\times 7\times 2\times 2$$

$$= 224 \text{ ways}$$

Example 6.11: Suppose we have 4,3,7 and 12 .In how many possible selection can the three numbers be drawn if the selection is done by

(a) with replacement

(b)without replacement.

Solution:

Here N = 4, n = 3.

(a) Selection with replacement $N^n = 4^3$=64 ways

(b) Selection without replacement

$$NC_n = 4C_3$$

$$= \frac{4!}{1!3!}$$

$$\frac{4\times 3!}{1\times 3!} = 4 ways$$

Example 6.12: Compute the following:

(a) $5C_3$

(b) $12C_4$

(c) $10C_3$

(d) $8C_3$

Solution:

(a) $5C_3 = \frac{5!}{3!2!}$

$=\frac{5\times 4\times 3\,!}{3!\times 2\times 1}$

=10 ways.

(b) $12C_4 = \frac{12!}{4!8!}$

$=\frac{12\times 11\times 10\times 9\times 8!}{8!\times 4\times 3\times 2\times 1}$

=495 ways.

(c) $10C_3 = \frac{10!}{3!7!}$

$=\frac{10\times 9\times 8\times 7!}{3\times 2\times 1\times 7!}$

=120 ways

(d) $8C_3 = \frac{8!}{3!5!}$

$=\frac{8\times 7\times 6\times 5!}{3\times 2\times 1\times 5!}$

=56

6.3 Probability and Probabilities Rules

Probability of an event occurring is the ratio of the number of favourable outcomes to the total number of outcome.

Hence, P(Event A) = P(A) = $\frac{fovourable\ \ outcomes}{Total\ \ number\ \ of\ \ trials}$

This is called classical approach or relative frequency approach.

The relative frequency approach is a development on the classical approach and the experiment is performed for a long period of time. There is a sufficient reason, so that : P(A) $\lim_{n\to\infty}(\frac{s}{n}) = p$; where s is the frequency of success in an experiment and n represent the total number of times the experiment is performed.

***Axioms or Rules of probability.**

This is also called laws or rules of probability. Let A be an event, the rules of probability are as follows:

(i) $p(A) = 1$ (This signifies a sure event)

(ii) $p(A) = 0$ (Implies that probability of p lies between 0 and 1 inclusive).

(iii) $0 \le p(A) \le 1$

If we have two events A and B,

(iv) P(A or B) = $P(A \cup B)$ = P(A) + P(B) – $P(A \cap B)$)

If the two event are not independent

(v) $P(A\ or\ B) = P(A) + P(B),$ for mutually exclusive events A and B.

(vi) The probability that a particular event A will not occur is = 1 –P (A). that is, one minus the probability that the event A will occur.

It can be written as $P(A^1) = 1 - P(A)$

Sample Space and Sample Point

The set of all possible outcome of a statistical experiment is called a sample space denoted by "S" or "Ω".

Each outcome in a sample space is called a sample point. For instance, given the experiment of rolling a die, the sample space is

S = (1 , 2, 3 ,4, 5, 6,) where, 2 is a sample point, 5 is a sample point etc.

The sample space in the experiments of tossing two coins is defined as

S = {HH, TH, HT, TT}.
HH is a sample point.

Example 6.13: A basket contains 5blue balls, 2 red balls, and 4 green balls. If a ball is selected at random, what is the probability that
(i) it is a blue ball.

(ii) it is a red ball.

(iii) it is a green ball.

Solution: Total = 5 +2 + 4 =11.
(i) Since there are 5 blue balls

$$\text{P(Blue)} = \frac{5}{11}$$

(ii) Since there are 2 red balls

$$\text{P(Red)} = \frac{2}{11}$$

(iii) Since there are 4 green balls

$$\text{P(Green)} = \frac{4}{11}$$

Example 6.14: What is the probability of:

(i) Selecting a blue ball or green ball

(ii) Selecting a blue ball and red ball in the above experiment.

Solution: (i) P(selecting a blue ball or green ball is

$$= P(B) + P(G)$$

$$= \frac{5}{11} + \frac{4}{11} = \frac{9}{11}$$

(ii) P (selecting a blue ball and red ball) is

=P (B)$\times$P(R) Since they are independent events

.

$$= \frac{5}{11} \times \frac{2}{11} = \frac{10}{121}$$

Example 6.15: A box contains 3 white balls and 4 black balls. Three balls are drawn at random with replaceable. Obtain the probability that (i) 3 white balls are drawn (ii) 1 white and 2 black balls (iii) at least 2 black ball

Solution:

Total number of balls = 3+ 4 = 7. Since the sample is with replacement, n=3, no. of white = 3, no. of black = 4.

$$S = N^n = 7^3 = 343$$

(i) $$\frac{3C_3.4C_0}{343} = \frac{\frac{3!}{3!0!} \times \frac{4!}{410!}}{343} = \frac{1}{343}$$

(ii) $$\frac{3C_1.4C_2}{343} = \frac{\frac{3!}{1!2!} \times \frac{4!}{212!}}{343} = \frac{3 \times 6}{343} = \frac{18}{343}$$

(iii) $$\frac{4C_1.3C_2}{343} \text{ or } \frac{4C_2.3C_1}{343} \text{ or } \frac{4C_3.3C_0}{343} = \frac{\frac{4!3!}{1!3!2!1!}}{343} + \frac{\frac{4!3!}{2!2!1!2!}}{343} + \frac{\frac{4!3!}{3!1!0!3!}}{343}$$

$$= \frac{(4 \times 3)}{343} + \frac{(6 \times 3)}{343} + \frac{(4 \times 1)}{343}$$

$$= \frac{34}{343}$$

Example 6.16: What is the probability that an integer x $(1 \le x \le 25)$ chosen at random is divisible by both 2 and 5?

$(1 \le x \le 25)$

Solution:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25

The numbers that are divisible by both 2 and 3 are 6, 12 18 and 24

$$p = \frac{4}{25}$$

$$X : N(82, 8^2)$$

$$\begin{aligned} p(X < 70) &= p(\frac{X - 82}{8} < \frac{70 - 82}{8}) \\ &= \Phi(\frac{70 - 82}{8}) \\ &= \Phi(-\frac{12}{8}) = \Phi(-1.5) \\ &= 1 - \Phi(1.5) \\ &= 1 - 0.9332 \\ &= 0.0668 \qquad\qquad (A) \end{aligned}$$

Example 6.17: A basket contains 9 apples, 8 bananas and 7 oranges. A fruit is picked from the basket at random. Find the probability that it is neither an apple nor an orange.

Solution:

Apples = 9

Bananas = 8

Orange =7

Total = 24

P(Apples) = $\frac{9}{24}$

P(Oranges) = $\frac{7}{24}$

$$P(\text{A or O}) = \frac{9}{24} + \frac{7}{24}$$

$$= \frac{2}{3}$$

$$\text{Pr(neither apple nor orange)} = 1 - \frac{2}{3}$$

$$= \frac{1}{3}$$

Example 6.18:

A box contains 2 red, 3 black and 4 white taxation policies. If three policies are drawn one at a time from the box, and each policy drawn is not replaced before the next drawing. You are required to calculate the probability that:

(i) The three policies are of different colour.

(ii) One policy is black and two are whites.

iii) At least two of the policies drawn is white?

Solution:

2 red, 3 blacks and 4 whites

Total number of policies = 2+3+4 = 9 = N

Policies number = n = 3

$$ {}^{9}_{3}C = \frac{9!}{2!(9-2)!}$$

$$\text{Sample space } = \frac{9\times8\times7!}{2\times1\times7!}$$

$$= 36$$

i. prob. of 3 different colours

$$= \frac{2\times3\times4}{36}$$

$$= \frac{24}{36}$$

$$= 0.67$$

ii. prob. of 1 black and 2 whites is

$$= \frac{C_1^3 \times C_2^4}{36}$$
$$= \frac{3\times 6}{36}$$
$$= \frac{18}{36}$$
$$= 0.5$$

iii. prob. of at least 2 whites is

$$= \frac{C_1^2 \times C_2^4 + C_1^3 \times C_2^4 + C_3^4}{36}$$
$$= \frac{(2\times 6)+(3\times 6)+4}{36}$$
$$= \frac{12+18+4}{36}$$
$$= \frac{34}{36}$$
$$= 0.94$$

CHAPTER SEVEN
PROBABILITY DISTRIBUTIONS

7.1 Learning Objectives:

At the end of this chapter, the students should be able to;

i. distinguish between discrete random variables and continuous random variables; and
ii. use probability distribution for discrete and continuous random variables to estimate probabilities and identify unusual events.

7.2 Bernoulli Trial

Any one trial whose outcomes can be classified as Yes or No, True or False, Success or Failure, Effective or Defective, Alive 0r Dead and so on is called Bernoulli trial. In Bernoulli trial, the probability of success is p and the probability of failure is $(1-p)=q$.

The probability distribution of Bernoulli trial is given by;

$$P(X=x)=f(x)=p^{x}(1-p)^{1-x}; x=0,1$$

7.3 Binomial Distribution

This simply means the repetition of Bernoulli trials. That is, when the Bernoulli is performed more than once $(n \geq 20)$

The probability distribution of Binomial distribution is given by;

$$P(X=x)=f(x)=nC_{x}p^{x}(1-p)^{n-x}; x=0,1,\ldots,n$$

The properties of a binomial distribution is that, the mean is $'np'$ and the variance is $'npq'$

Example 7.1: The probability that an audit Assistant commits an error in an audit assignment is 0.45. What will be the probability that in 8 audit assignments?

(a) 2 errors will be committed

(b) no error will be committed

(c) at most 6 errors will be committed

(d) at least 2 errors will be committed

Solution:

Here, $p=0.45,\ q=1-p=1-0.45=0.55,\ n=8$

(Using binomial distribution)

$$P(X=x)=nC_{x}p^{x}(1-p)^{n-x}$$

(a) $P(X=2)=8C_{2}(0.45)^{2}(0.55)^{8-2}$

$$=\frac{8!}{2!6!}\times 0.45^{2}\times(0.55)^{6}$$

$$=28\times 0.45^{2}\times 0.55^{6}$$

$$=0.1569$$

(b) $P(X=0)=8C_0\times0.45^0\times0.55^8$

$=1\times0.55^8$

$=0.00837339$

$=0.0084$

(c) $P(X\le6)=P(X=0,1,2,3,4,5,6)$

$=1-p(X>6)$

$=1-\left[P(X=7+P(X=8)\right]$

But

$P(X=7)=8C_7(0.45)^7(0.55)^{8-7}$

$=8\times0.45^7\times0.55$

$=0.01644$

$=0.0164$

$P(X=8)=8C_8(0.45)^8(0.55)^{8-8}$

$=1\times0.45^8\times0.55^0$

$=0.01345$

$=0.0135$

Therefore, the P(at most 6 errors will be committed) is

$=1-\left[P(X=7+P(X=8)\right]$

$=1-(0.0164+0.0135)$

$=1-0.0299$

$=0.9701$

(d) P(at least 2 errors will be committed) is

$P(X \geq 2) = P(X = 2,3,4,5,6,7,8)$

$= 1 - P(X < 2)$

$= 1 - P(X = 0,1)$

But

$P(X = 1) = 8C_1(0.45)^1(0.55)^{8-1}$

$= 8 \times 0.45 \times 0.55^7$

$= 0.05481$

$= 0.0548$

$P(X \geq 2) = P(X = 2,3,4,5,6,7,8)$

$= 1 - P(X = 0,1)$ (since $P(X = 0) = 0.0084$)

$= 1 - (0.0548 + 0.0084) = 0.9368$

Example 7.2: If 0.05 of the tyres purchased by a Cab company last less than 10 months and if 10 tyres are purchased, calculate the probability that

(a) at most 2 of the tyres will expire in 10 months.

(b) at least 2 tyres will expire in 10 months.

Solution:

$n = 10, p = 0.05, q = 1 - 0.05 = 0.95$

Using binomial distribution,

$(a) P(X \leq 2) = P(X = 0,1,2)$

$= P(X = 0) + P(X = 1) + P(X = 2)$

Where,

$P(X = 0) = 10C_0(0.05)^0(0.95)^{10-0}$

$$=1\times 0.95^{10}$$

$$=0.5987$$

$$P(X=1)=10C_1(0.05)^1(0.95)^{10-1}$$

$$=10\times 0.05\times 0.95^9$$

$$=0.3151$$

$$P(X=2)=10C_2(0.05)^2(0.95)^{10-2}$$

$$=45\times 0.05^2\times 0.95^8$$

$$=0.0746$$

$$\therefore P(X\le 2)=P(X=0,1,2)$$

$$=P(X=0)+P(X=1)+P(X=2)$$

$$=0.5987+0.3151+0.0746$$

$$=0.9884$$

$$(b)P(X\ge 2)=P(X=2,3,4,5,6,7,8,9,10)$$

$$=1-P(X=0,1)$$

$$=1-(0.5987+0.3151)$$

$$=0.0862$$

7.4 Poisson Distribution

This distribution describes the situation where it is meaningful to be concerned with the probability "p" of the occurrence of an event but less meaningful to worry about the probability of the event not occurring. Instead of using binomial distribution, we can use Poisson distribution under the following conditions;

(i) when the probability of an event occurring is very small that is when

$(p \le 0.01)$

(ii) the sample size' n'is too large

$(n \ge 20)$

For instance, in the banking hall, we are interested with the probability of arrival of customers than with their non-arrival and so on.

The probability distribution of a Poisson distribution is given by

$$P(X = x) = f(x) = \frac{e^{-\lambda}.\lambda^{x}}{x!}; x = 0, 1, 2, \ldots$$

Where,

$\lambda = np = mean / or\ average$

The mean and variance of Poisson distribution are equal = np

Example7.3: A computer manufacturer knows that 0.3 percent of every batch of 3000 computer will be defective. What is the probability that out of a batch of 1000

(a) exactly 4 will be defective.

(b) not more than 3 will be defective.

(c) at least 2 will be defective.

Solution:

Here, the appropriate probability distribution is a Poisson distribution

$X : p(\lambda)$

$$\lambda = np = 1000 \times \frac{0.3}{100} = 3$$

$$(a) P(X = 4) = \frac{e^{-3} \times 3^{4}}{4!} = \frac{81}{24} \times e^{-3}$$

$$= \frac{81}{24} \times 0.0498$$

$$= 3.375 \times 0.0498$$

$$= 0.1681$$

$$(b) P(X \le 3) = P(X = 0,1,2,3)$$

$$= \frac{e^{-3}.3^0}{0!} + \frac{e^{-3}.3^1}{1!} + \frac{e^{-3}.3^2}{2!} + \frac{e^{-3}.3^3}{3!}$$

$$= e^{-3}\left(1 + 3 + \frac{9}{2} + \frac{27}{6}\right)$$

$$= e^{-3}(\frac{6+18+27+27}{6})$$

$$= 13 \times 0.0498$$

$$= 0.6474$$

$$(c) P(X \ge 2) = P(X = 2,3,4,\cdots\cdots,999,1000)$$

$$= 1 - P(0,1)$$

$$= 1 - P(X = 0) - P(X = 1)$$

$$= 1 - \left[\frac{e^{-3}.3^0}{0!} + \frac{e^{-3}.3^1}{1!}\right]$$

$$= 1 - (1+3)e^{-3}$$

$$= 1 - 4e^{-3}$$

$$= 1 - 4 \times 0.0498$$

$$= 1 - 0.1992$$

$$= 0.8008$$

Example 7.4: The mean arrival rate of Customers at a telephone exchange is 10 customers per hour. Calculate the probability that:

(a) no person

(b) 3 persons

(c) 5 persons arrived at an exchange within an hour.

Solution:

$$mean = average = \lambda = 10$$

If X denotes the number of arrival

$$(a) P(X = 0) = \frac{e^{-0} 10^{0}}{0!}$$

$$= 1 \times 0.00005$$

$$= 0.00005$$

$$(b) P(X = 3) = \frac{e^{-3}.10^{3}}{3!}$$

$$= \frac{0.00005 \times 1000}{3 \times 2 \times 1}$$

$$= 0.0076$$

$$(c) P(X = 5) = \frac{e^{-5} 10^{5}}{5!}$$

$$= \frac{100000 \times 0.00005}{5 \times 4 \times 3 \times 2 \times 1}$$

$$= \frac{5}{120}$$

$$= 0.0416666$$

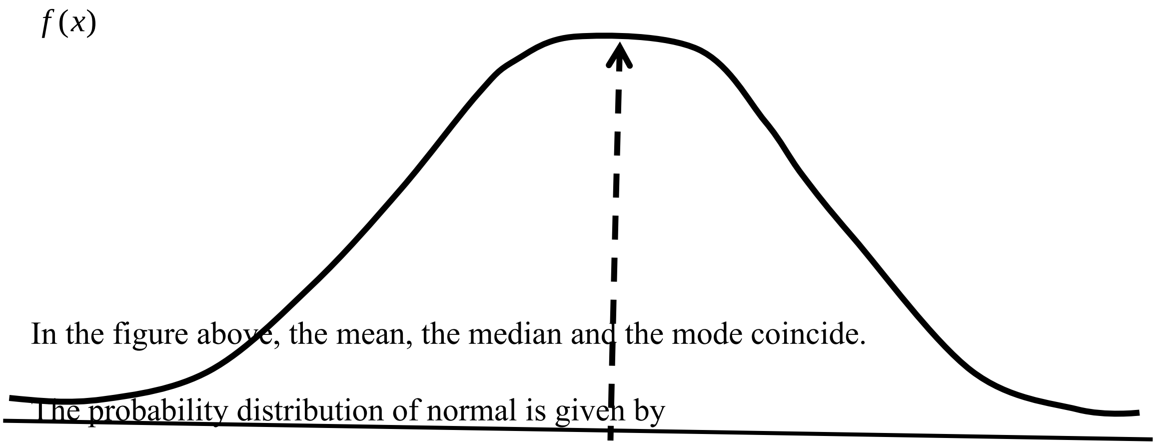
`

$$= 0.0147$$

7.5 Normal Distribution

This is the most widely distribution used in sciences, social sciences and humanities. This distribution fits many natural processes. It is a bell shaped and symmetrical about its mean. A random variable X has a normal distribution with mean μ and variance (σ^2). It is conveniently written as $X_{\sim} N(\mu, \sigma^2)$

The shape of the normal distribution is shown below as;



In the figure above, the mean, the median and the mode coincide.

The probability distribution of normal is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\propto}^{+\propto} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} dx, \quad -\propto < x < +\propto \qquad \text{(i)}$$

The expression above becomes standard form distribution as

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad provided \;\; z : \; N(0,1) \qquad (ii)$$

The expression in (i) above is transform to equation (ii) when

$\mu = 0, \sigma^2 = 1$

And

$z = \dfrac{x-\mu}{\sigma}$ (normal variate or standard score).

The computations of probability in normal distribution is based on the standard score derived from standard normal distribution.

7.5.1 Properties of Normal Distribution

(i) The total area under the curve equal to 1

(ii) It is symmetrical in shape

(iii) The mode is maximum and its occurs at point on the horizontal axis

(iv) Special property is that, the normal distribution can be approximated to binomial distribution.

Example 7.5: The amount of money withdrawal from a bank branch daily is normally distributed with mean 100 (₦ '000) and standard deviation 20. Find the probability that the amount of money withdrawn in a particular day is

(a) at most ₦120,000

(b) between ₦120,000 and ₦140,000

(c) not less than ₦125,000

Solution:

$$\mu = 100$$

$$\sigma^2 = 20$$

Let X represent the amount of money withdrawn

$$(a)P(X \leq 120) = P(\frac{X-100}{20} - P(\frac{120-100}{20})$$

$$= P(Z \leq 1)$$

$$= \Phi(1)$$

$$= 0.8413$$

$$(b)P(120 \leq X \leq 140) = P(X \leq 140) - P(X \leq 120)$$

$$P(\frac{140-100}{20}) - P(\frac{120-100}{20})$$

$$= P(Z \le 2) - 0.8413 \qquad \textit{from 'a' above}$$

$$= \Phi(2) - 0.8413$$

$$= 0.9772 - 0.8413$$

$$= 0.1359$$

$$(c) P(X \ge 125) = 1 - P(X \le 125)$$

$$= 1 - P(\frac{X-100}{20} \le \frac{125-100}{20})$$

$$= 1 - P(Z \le 1.25)$$

$$= \Phi(1.25)$$

$$= 1 - 0.8944$$

$$= 0.1056$$

Example 7.6: Test on electrical lamps of a certain type indicated that length of life's are normally distributed with a mean of 1500 hours and variance 8100 hours. What is the probability of lamps which can be expected to burn in:

(a) more than 1200 hours

(b) less than 1400 hours

Solution:

$\mu = 1500$

$\sigma^2 = 8100 \rightarrow \sigma = 90$

$$(a) P(X > 1200) = 1 - P(X < 1200)$$

$$= 1 - P(\frac{X-1500}{90} < \frac{1200-1500}{90})$$

$$=1-P(Z<-\frac{30}{90})$$

$$=1-\Phi(-3.33)$$

$$=1-(1-\Phi(3.33))$$

$$=1-1+\Phi(3.33)$$

$$=\Phi(3.33)$$

$$=0.9995$$

$$(b)P(X<1400)=P(\frac{X-1500}{90}<\frac{1400-1500}{90})$$

$$=P(Z<-\frac{100}{90})$$

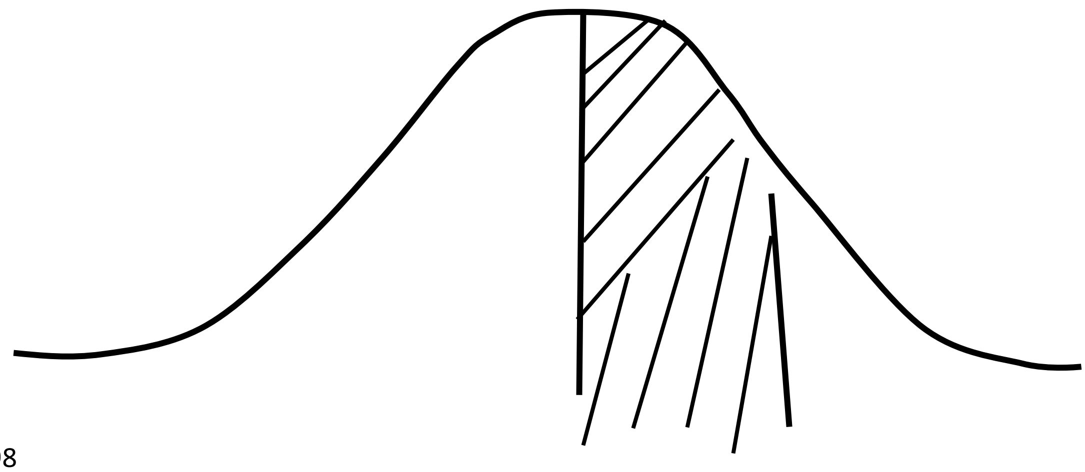
$$=1-\Phi(1.1)$$

$$=1-0.8643$$

$$0.1357$$

The area under the normal curve is referred to as standard a normal table which gives the probabilities. The probability is denoted by

$$\Phi(z)=\int_0^z p(z)dz$$



Example 7.7: The scores of candidates in an examination are considered as having a normal distribution with a mean of 82 and a standard deviation of 8. What is the probability that a candidate drawn from the class would have received a grade less than 70?

Solution:

$X : N(82,8)$

$$P(X < 70) = P(\frac{X-82}{8} < \frac{70-82}{8})$$

$$= P(Z < -\frac{12}{8})$$

$$= 1 - \Phi(1.5)$$

$$= 1 - 0.9332$$

$$= 0.0668$$

7.5.2 How to use the Normal Table

Locate

$P(z < 0.24, say,,$

Locate

0.2

And go to 5th column with leading 0.04, this gives 0.0948

Example 7.8:

Use the standard normal table to find the area of the following:

(a) To the left of z = 0.21

(b) To the right of z = 0.21

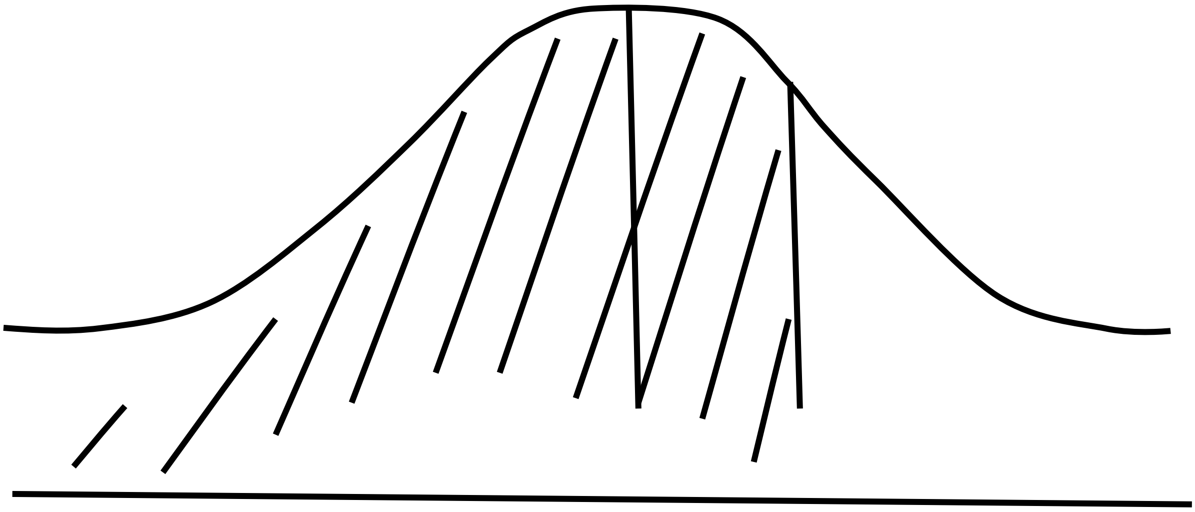
(c) To the left of z= 0.21

(d) To the right of z = -0.21

Solution:

It is advisable to draw the diagram to check whether your probabilities are reasonable

(a) The area to the left of z = 0.21

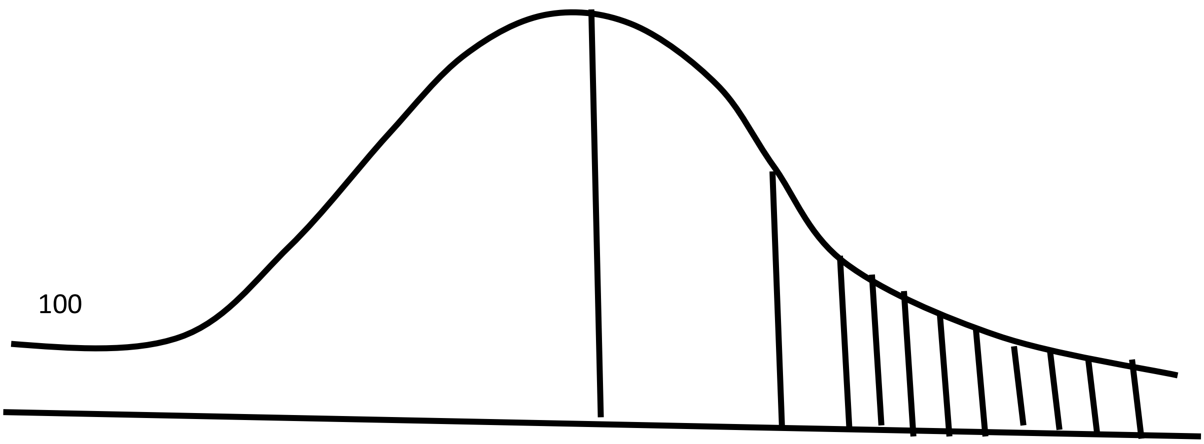


Z = 0.21

The area to the left of z = 0.21 is 0.500 plus the entry in Table corresponding to z = 0.21 which is 0.0832

Ans = 0.500 + 0.0832 = 0.5832

(b) The area to the right of z = 0.21



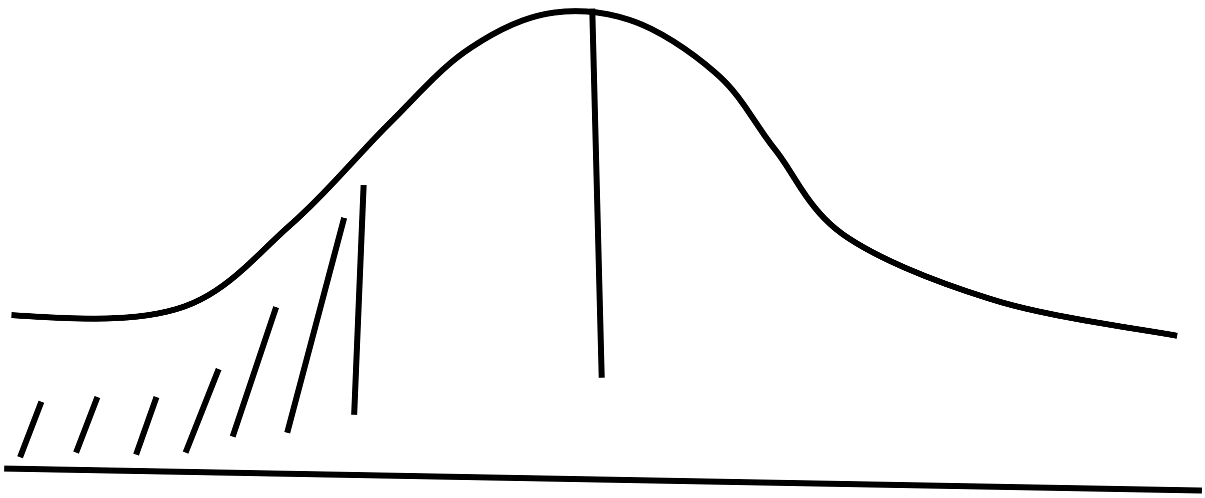
Z = 0.21

The area to the right of z = 0.21 is 0.500 minus the entry in the Table corresponding z = 0.21,

= 0.500 - 0.0832

$P(Z > 0.21) = 0.4168$

(c) The area to the left of z = - 0.21



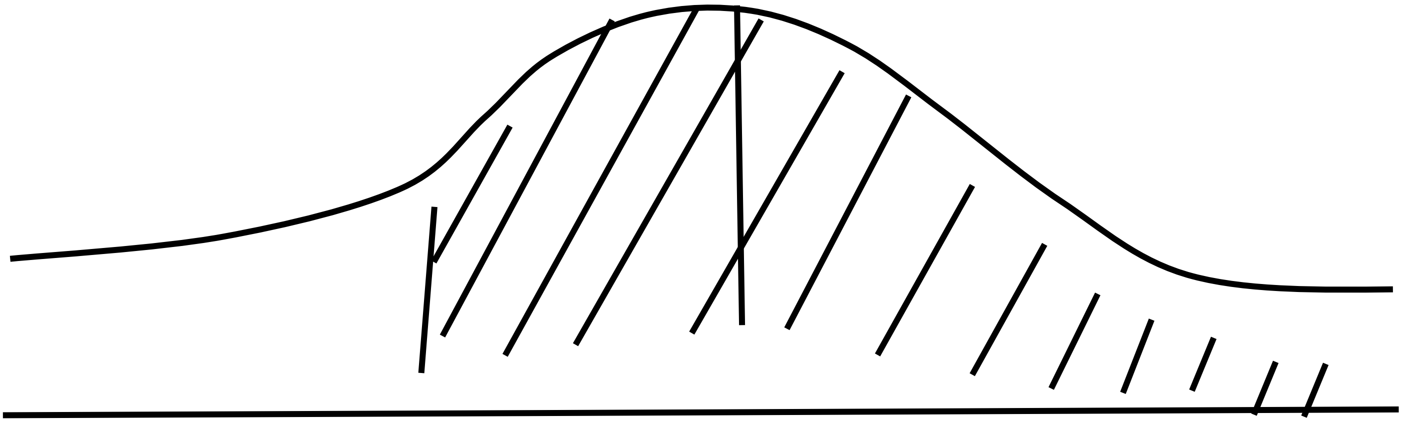
Z = - 0.21

The area to the left of z = - 0.21 is 0.500 minus the entry in the Table corresponding z = - 0.21 thus,

= 0.500 - 0.0832

$\therefore P(z < -0.21) = 0.4168$

(d) The area to the right of z = - 0.21



Z = -0.21

The area to the right of z = - 0.21 is 0.500 plus the entry in the Table corresponding z = 0.21,

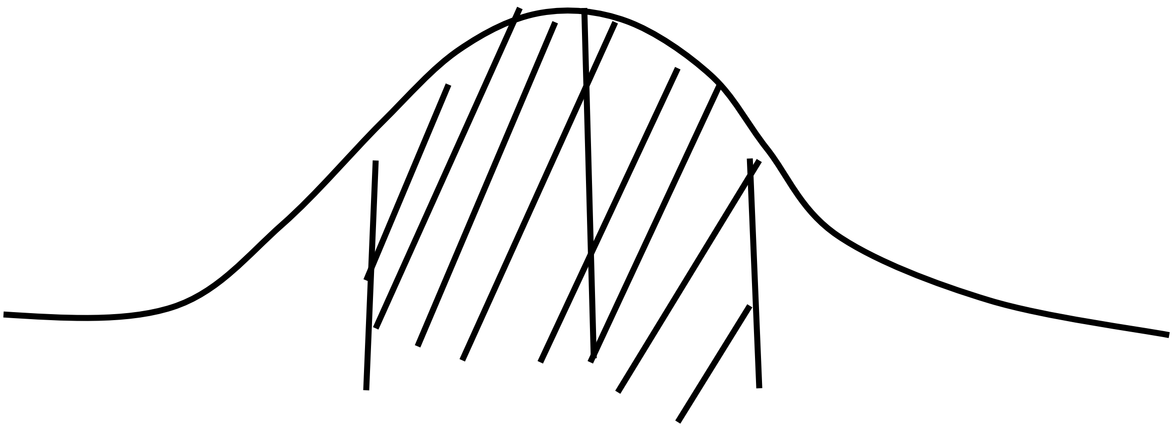
Thus, 0.500 + 0.0832 = 0.5832

$\therefore P(z > -0.21) = 0.5832$

Example 7.8:

Find the area under normal curve between z = - 1.37 and z = 1.71

Solution:



Z = -1.37   Z = 1.71

The area between -1.37 and 1,71 is the sum of the entry in the Table of normal distribution corresponding to z = -1.37 and z = 1.71

= 0.4147 + 0.4564

= 0.8711

CHAPTER EIGHT
REGRESSION AND CORRELATION

8.1 Learning Objectives:

At the end of this chapter, the students should be able to;

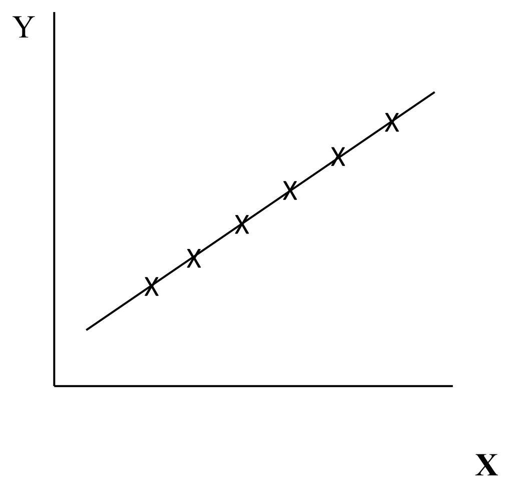
i. understand how the least square regression line is defined;
ii. interpret the slope (or gradient) of the regression line;
iii. graph the scatter plot and the regression line on the same grid;
iv. understand and describe how two different variables can be related;
v. determine if two different variables have a linear or nonlinear relationship;
vi. describe the relationship between different variables in terms of positive and negative correlation; and
vii. describe how strongly correlated two different variables are used or the proximity to a line of best fit or a scatter graph or a scatter graph without a line of best fit.

8. 2 Introduction

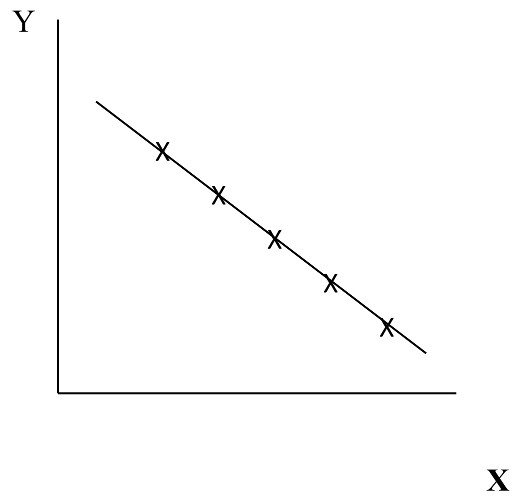
The predicting value of one variable in terms of a known value of another variable is referred to as regression. For instance, let assume that the manager had already noted the changes in systematic manner; he should be able to deduce the value of sales given a particular level of expenditure using graphical relationship. The vertical or y- axis is used for the independent or response variable while horizontal or x –axis is used for the independent variable. This type of diagram or graphical relationships is known as scatter diagram.

Thus, this relationship ranges from a perfect positive relationship to perfect negative and finally to no relationship. Therefore, a perfect relationship is one where a single straight line can be drawn through all the points.

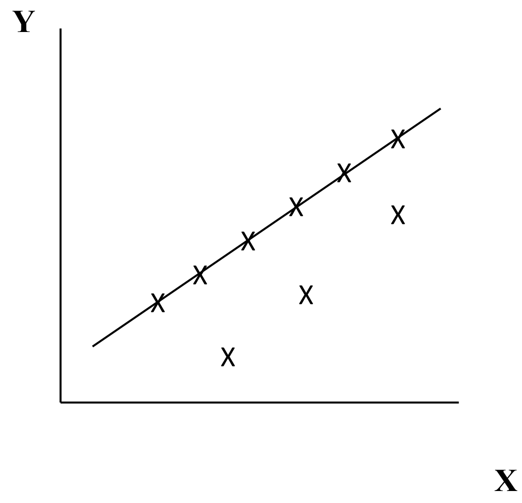
.



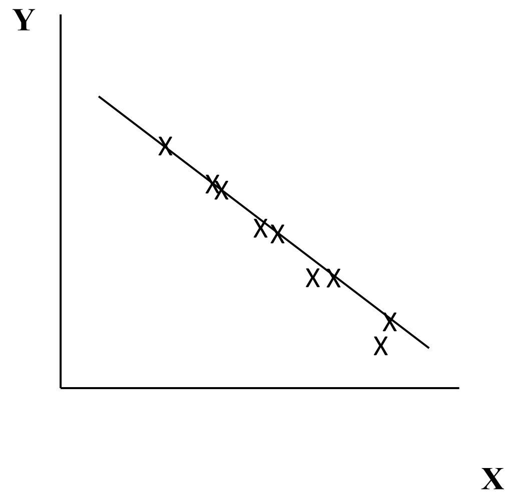
Perfect Positive Relationship



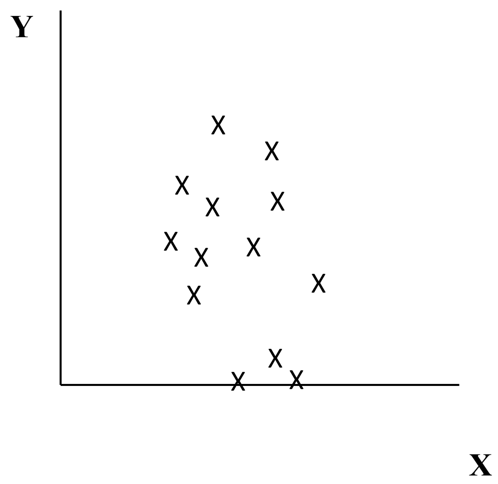
Perfect Negative Relationship



+ve relationship



- ve relationship



No relationship

8.3 Correlation

This is the relationship or association that exists between dependent and independent variable e.g. height and weight, income and expenditure, demand and supply and so on. The variables may be perfectly correlated, partly correlated and uncorrelated. The degree of correlated between two variables can be measured. Hence, correlation coefficient of correlation is the measure of linear relationship between two variables.

In practice, there are two common measures of correlation viz:

(i) The Pearson product moment correlation coefficient denoted by “r”.

(ii) The spearman’s rank correlation coefficient denoted by” r_k ”.

Equation of a Straight Line

The equation of a straight line y= a +bx is the central to all activities in regression and correlation.

Where,

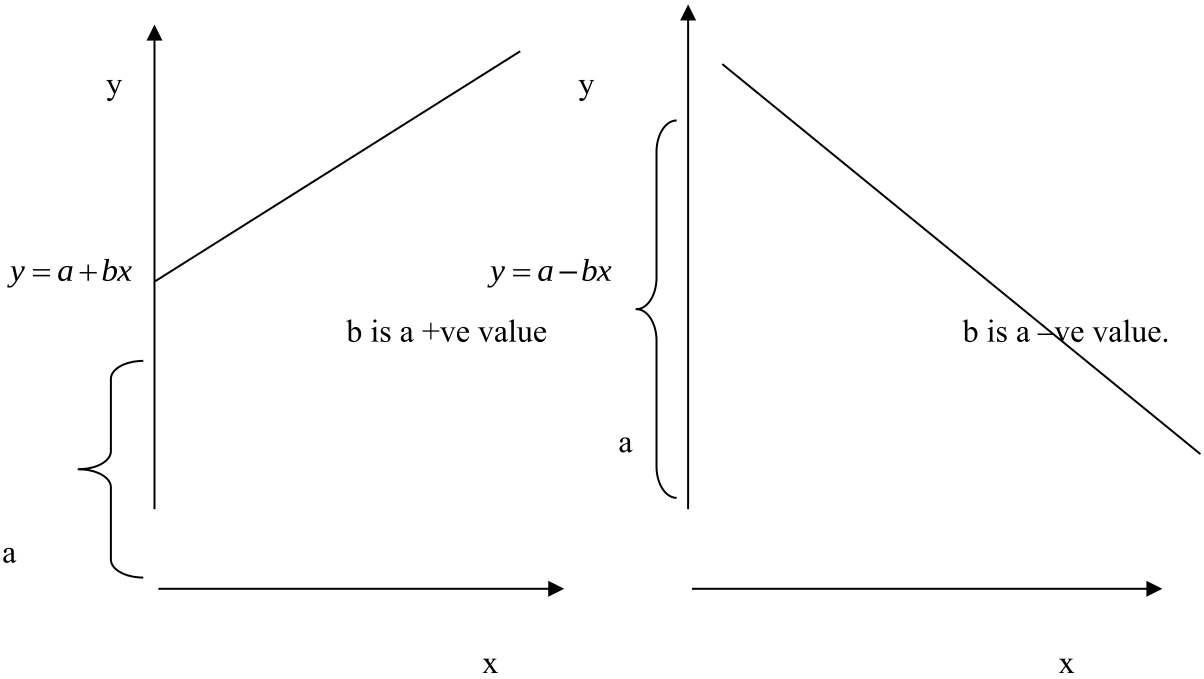
y = the dependent variable.

x = the independent variable.

a = the point where the line cuts the y –axis.

b = the slope or gradient of the line.

The appropriate values of “a” and “b” can be estimated from the data and from a given value of x, the value of y can be determined.



A positive slope **A negative slope**

Regression Analysis or Curve Fitting

This is the technique used to establish line of "best fit" to the observed data. The data can be shown as a scatter diagram with lines is drawn, free hand; thus, the lines are drawn based on the judgment of individuals. With this approach, there is no unique regressive line is drawn. To estimate the line of best fit, we use the method of least squares.

The Least Squares Method

The equation of a straight line $y = a + bx$, where, x, y, a and b are defined earlier, the values of "a" and "b" were estimated by

$$\hat{b} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$; n represent the sample size.

and

$\hat{a} = \bar{y} - \hat{b}\bar{x}$

if we regressed y on x, we then have $y = a + bx$

The alternative method of estimating "a" and "b" is to solve the equation (i) and (ii) below simultaneously

$$\sum y = na + b\sum x \qquad (i)$$

$$\sum xy = a\sum x + \sum x^2 \qquad (ii)$$

Correlation Co-efficient

We want to have some measure of how good a fit a line has to data.

The correlation coefficient provides us with this measure.

1. The Pearson Product Moment correlation coefficient r is given by

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[(n\sum x^2) - (\sum x)^2\right] \times \left[(n\sum y^2) - (\sum y)^2\right]}}$$

The value of "r" is in the range ; $-1 \le r \le +1$.

Interpretation of the "r" value

(i) if r = 0, no correlation exist between x and y.

(ii) if r =1, there is a perfect +ve correlation between x and y.

(iii) if r = -1, there is a perfect –ve correlation between x and y.

(iv) if r < 1, there is a +ve correlation between x and y but not perfect.

(v) if r < -1 ,there is a –ve correlation between x and y but not perfect.

Hence, the nearer the value of r to 1, the stronger the relationship or association.

2. Spearman's rank correlation coefficient

$$r_k = 1 - \frac{6\sum d^2}{n(n^2-1)}, \quad d = r_x - r_y \quad or \quad d = r_y - r_x$$

Example 8.1: The table below shows the incomes and expenditures of 8 workers in (#'000).

Table 28

Incomes(x)	Expenditures (y)
2	4
3	6
5	8
8	13
9	15
12	18
14	24
17	28

(a) Regress y on x and determine the equation of best line of fit.
(b) Predict the income when the expenditure is #40,000.
(c) Estimate the expenditure when the income is #7,000.

Solution:

x	y	xy	x^2
2	4	8	4
3	6	18	9
5	8	40	25

8	13	104	64
9	15	135	81
12	18	216	144
14	24	336	196
17	28	476	289
70	116	1333	812

$$\hat{b} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$\hat{b} = \frac{8(1333) - 70(116)}{8(812) - 70^2}$$

$$\hat{b} = \frac{10664 - 8120}{6496 - 4900}$$

$$\hat{b} = \frac{2544}{1596}$$

$$\hat{b} = 1.59$$

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

$$= \frac{116}{8} - 1.59 \times \frac{70}{8}$$

$$= 14.5 - 1.59(8.75)$$

$$= 14.4 - 13.91$$

$$= 0.59$$

The equation of line of best fit is given by:

$$y = 0.59 + 1.59x$$

(b)

when y = 40,

$$40 = 0.59 + 1.59x$$

$$1.59 = 40 - 0.59$$

$$1.59x = 24.79$$

$$x = \frac{24.79}{1.59}$$

$$= 15.5911$$

$$= 15.59$$

(c)

when

$$x = 7,$$

$$\hat{y} = 0.59 + 1.59(7)$$

$$\hat{y} = 11.72 \times 1000$$

$$\hat{y} = N11,720$$

Example 8.2: The table below shows the age and cost of 6 people.

Table 29

Age (x)	Cost (y)
5	160
10	80
15	95
20	120
25	140
30	170

Solution:

x	y	xy	x^2	y^2	r_x	r_y	$d = r_x - r_y$	d^2
5	160	800	25	25600	6	2	4	16
10	80	800	100	6400	5	6	-1	1
15	95	1425	225	9025	4	5	-1	1
20	120	2400	400	14400	3	4	-1	1
25	140	3500	625	19600	2	3	-1	1
30	170	5100	900	28900	1	1	0	0
105	765	14025	2275	103925	-	-	-	20

Determine the relationship between x and y using :

(a) Using Pearson's Moment

(b) Using spearman's method.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[(n\sum x^2) - (\sum x)^2\right] \times \left[(n\sum y^2) - (\sum y)^2\right]}}$$

$$= \frac{6(14025) - (105)(765)}{\sqrt{\left[6(2275) - 105^2\right] \times \left[6(103925) - 765^2\right]}}$$

$$= \frac{84150 - 80325}{\sqrt{[13650 - 11025] \times [623550 - 585225]}}$$

$$= \frac{63825}{\sqrt{2625 \times 38325}}$$

$$= \frac{3825}{\sqrt{100,603,125}}$$

$$= \frac{3825}{10030.11}$$

$$= 0.38$$

(b) Using Spearman's rank

$$r_k = 1 - \frac{6\sum d^2}{n(n^2-1)}, \ d = r_x - r_y / r_y - r_x$$

$$= 1 - \frac{6(20)}{6(6^2-1)}$$

$$= 1 - 0.57$$

$$= 0.43$$

There is a weak relationship between age and cost.

Example 8.3: The data below relates the weekly maintenance cost (#) to the age (in months) of ten machines of similar type in manufacturing company.

Table: 8.3

Machine	1	2	3	4	5	6	7	8	9	10
Age	5	10	15	20	30	30	30	50	50	60
Cost	190	240	250	300	310	335	300	300	350	395

You are required to calculate the correlation coefficient between ages and cost using Pearson's product moment method and interpret your findings.

Solution:

X	Y	X^2	Y^2	XY
5	190	25	36100	950
10	240	100	57600	2400
15	350	225	122500	5250
20	300	400	90000	6000
30	310	900	96100	9300
30	335	900	112225	10050
30	300	900	90000	9000
50	300	2500	90000	15000
50	350	2500	122500	17500

60	395	3600	156025	23700
300	3070	12050	973050	99150

Pearson's | Product moment correlation coefficient is given by

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$r = \frac{10(99150) - 300(3070)}{\sqrt{[10(12050) - (300)^2][10(973050) - (3070)^2]}}$$

$$= \frac{991500 - 921000}{\sqrt{30500 \times 305600]}}$$

$$= \frac{70500}{96544.29}$$
$$= 0.73$$

The Pearson's correlation coefficient is positively high ($(r = 0.7)$ close to 1. We can then say that there is evidence of good agreement between the age and cost. Thus, this indicates that the cost estimates are likely to be very reliable.

CHAPTER NINE

SAMPLING DISTRIBUTION AND ESTIMATION

9.1 Learning Objectives:

At the end of this chapter, the students should be able to;

i. understand the importance of sampling and how results from samples can be used to provide estimates of population characteristics such as population total, population mean, population variance

9.2 Introduction

Inferential statistics involves drawing meaningful conclusions about characteristics of the population using values of summary statistics computed from the samples drawn from the populations.

Some Useful Terms:

1) **Estimation**: Estimation of population characteristics (such as the sample mean and sample standard deviation).
2) **Parameter**: A parameter refers to the quantitative characteristics of the population e.g. population mean (N), population variance (σ^2), population proportion (p) and so on.
3) **Statistic**: This is a function of observation from the samples e.g. sample mean ($\bar{x}$), sample variance (s^2), sample proportion (p) and so on.

9.3 Properties of a Good Estimator

(a) **Unbiased** – A estimate is said to be unbiased if the mean of the sampling distribution of the sample statistic is the same as the population parameter otherwise, it is biased.

(b) **Efficiency**– .Estimator $\underline{T(x_1, x_2, \ldots x_n)}$ is said to be efficient estimator of its expectation $m(\theta)$ if there is equality in the information Inequality.

(c) **Consistency** – This means that as the sample size increases, the estimate of the parameter value is close to the actual value.

(d) **Sufficiency** – An estimator is said to be sufficient if it utilizes all the information in a sample relevant to the estimation of the population parameter.

Table 30

	Sample statistics	Population parameter
Arithmetic means	$\bar{x}$	μ
Standard deviation	s	σ
Number of items	n	N

Example 9.1: The elements of a given population X are; 4, 6, 7, 5, 2, 8, 3 for $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ respectively.

(a) Calculate the population mean and population variance

(b) If x_3, x_4, x_5, x_7 are selected at first sample. Obtain the sample mean and sample variance

(c) Can we conclude that the sample mean is an unbiased estimate of population mean?

Solution:

$x = 4, 6, 7, 5, 2, 8, 3$

$$\mu = \frac{\sum x}{n} = \frac{35}{7} = 5$$

$$\sigma^2 = \frac{\sum (x-\bar{x})^2}{n}$$

x	$x-\bar{x}$	$(x-\bar{x})^2$
4	-1	1
6	1	1

7	2	4
5	0	0
2	-3	9
8	3	9
3	-2	4
-	-	28

Therefore,

$$\sigma^2 = \frac{28}{7} = 4$$

(b) $x = 7, 5, 2, 3$

$$\bar{x} = \frac{\sum x}{n} = \frac{20}{4} = 5$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n}$$

x	$(x - \bar{x})$	$(x - \bar{x})^2$
7	2	4
5	0	0
2	-3	9
3	-2	4
-	-	17

$$s^2 = \frac{17}{4} = 4.25$$

(c) Hence, the sample mean is an unbiased estimate of the population mean.

Example 9.2: A computer firm produces a large quantity of computer components and wishes to known that the average life time of the component. A random sample of 75 is life time and found that the sample life-time is 100 with a sample standard deviation of 10.

Compute the followings;

(a) The estimate of the sample mean

(b) The standard error of the mean

(c) The standard error of the mean when the sample size was 10,000? (Hint: standard error = $\frac{s}{\sqrt{n}}$.)

Solution:

(a) $\bar{x} = 100 years$

$\mu = 100 years$

(b) standard error of the mean = $\frac{s}{\sqrt{n}}$

$$= \frac{10}{\sqrt{75}}$$

$$= \frac{10}{8.66}$$

$$= 1.155$$

(c) when n= 10000

Standard error of the mean now becomes

$$= \frac{10}{\sqrt{10000}} = 0.1$$

This confirms that the increase in the sample size reduces the standard error.

9.4 Estimation

There are two types of estimation namely point estimation and interval estimation. **9.4.1 Point Estimation** refers to a single value of a statistics that is used to approximate a population parameter. The statistic that one used to obtain point estimation is known as

Estimator and the value of statistic is called the **estimate.**

9.4.2 Interval Estimate -This is called confidence interval estimation. The confidence that the population parameter does lie. For instance, confidence limits for population means are based on the sample means and the standard error of the means. Therefore, the confidence interval (CI) for population mean is given by

$$CI : \overline{x} \pm Z_{\alpha} SE(\overline{x})$$

And $SE(\overline{x}) = \frac{s}{\sqrt{n}}$,where$\bar{x}$ = population mean, α = level of significance, s = standard deviation of the sample and n= sample size .

Example 9.3: Find the confidence limits for the population mean if the sample mean of sample size of 144 is 25 and sample standard deviation 24.

Solution:

$$C.I : \mu = 25 \pm \frac{1.96(24)}{\sqrt{144}}$$

$$= 25 \pm \frac{1.96(24)}{12}$$

$= 25 \pm 3.92$

$= (21.08 \quad 28.92)$

The confidence limits for population mean are 21.08 and 28.9

9.5 Estimation of population proportion

This is a fraction of the population that has a certain characteristics. For instance, the population has size N =1000 and 10 samples were selected for a competition. The proportion of selected person is $\frac{10}{1000}$ which is 0.01

It should be noted that, population proportion is a parameter that describes a percentage value associated with a population. Hence, the confidence interval of population proportion parameter is given by

$$C.I : \hat{p} \pm Z_{\alpha} SE(\hat{p})$$

where

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$\hat{P}$ is the sample proportion and P is the population proportion, n is the sample size.

Example 9.4: A random sample of 20 taken from a population is size N= 400 and 50 percentage is defective. Set up a 95 % confidence interval for the population proportion

Solution:

n =20, N= 400

$$\hat{p} = \frac{n}{N} = \frac{20}{400} = 0.05$$

$$1 - \hat{p} = 1 - 0.05$$

$$= 0.95$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= \sqrt{\frac{0.05(0.95)}{20}}$$

$$= \sqrt{0.002375}$$

$$= 0.049.$$

The 95% confident and the population proportion is

$$\hat{P} \pm 1.96 SE(\hat{P})$$

$$= 0.05 \pm 1.96\ (0.049)$$

$$= 0.05 \pm 0.096$$

$$=(-0.096 \quad 0.146).$$

Difference Between Two population means.

Given two population I and II

populations

	I	II
Sample size	n_1	n_2
sample mean	$\bar{x}_1$	$\bar{x}_2$
sample variance	s_1^2	s_2^2

The confidence interval of the difference between the two populations is given by

$$C.I : (\bar{x}_1 - \bar{x}_2) \pm Z_\alpha SE(\bar{x}_1 - \bar{x}_2),$$

where. $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

So that C. I now become

$$C.I : (\bar{x}_1 - \bar{x}_2) \pm = Z_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example 9.5: The training manager of a commercial bank found out that in a banking appreciation course for 20 supervisors, the average score was 60 percent with a standard deviation 15 and the training manager of a merchant bank found out that the average mean was 55 percent with a standard deviation of 12 for 10 supervisors. Determine a 5% confidence interval for the average score?

Solution:

Since n is small (i.e.<30), we use student's t-distribution to construct the confidence interval.

$$C.I:(\bar{x}_1-\bar{x}_2)\pm t_{0.05(d.f=n_1+n_{2-2)})}SE(\bar{x}_1-\bar{x}_2)$$

But $SE(\bar{x}_1-\bar{x}_2)=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

$$=\sqrt{\frac{15^2}{20}+\frac{12^2}{10}}=\sqrt{\frac{225}{20}+\frac{144}{10}}$$

$$=\sqrt{11.25+14.4}$$

$$=\sqrt{25.65}=5.06$$

Hence, $C.I:15\pm t_{0.05}(28)\times 5.06$

$$15\pm 1.701\times 5.06$$

$$15\pm 8.607$$

$$(6.393, 23.61)$$

Difference between Two Populations Proportions

Given two population I having sample size n_1, sample proportion p_1 and p0pulation II with sample size n_2, sample proportion p_2.

The confidence interval for difference of two population proportion is given by

C.I = $(p_{1-}p_2)\pm Z_\alpha\ SE(p_{1-}p_2)$

where

$$SE(p_1 - p_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

So that

$$C.I : (p_1 - p_2) \pm Z_\alpha \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

CHAPTER TEN
TEST OF HYPOTHESIS

10.1 Learning Objectives:

At the end of this chapter, the students should be able to;

i. understand the difference between hypothesis and a theory;
ii. develop scientific hypothesis;
iii. identify the steps of hypothesis testing; and
iv. define null hypothesis and alternative hypothesis, level of significance, test statistic and so on.

10.2 Introduction

This is another branch of inferential statistics that follows the estimation of population parameters. This hypothesis testing is also called significance testing. Definition – A hypothesis is a belief or opinion tested by statistical methods.

10.2.1 Types of Hypothesis

There are two types hypothesis namely, Simple hypothesis and composite hypothesis.

A hypothesis is simple if is assigns a specific value to the parameter under study, e. g

$\mu = 15$kg.

Hypothesis is composite if it does not assign a specific value to the parameter under study, e. g.

$\mu \neq 15$kg.

The hypothesis to be tested is called Null hypothesis denoted by H_o and Alternative hypothesis denoted by H_1.

10.2.2 Test Procedures

The following are the test procedures or the procedure involved in hypothesis testing.

Step 1: State the Null Hypothesis

Step 2: State the alternative hypothesis

Step 3: State the level of significance i.e. x%

Step 4: State an appropriate test and compute its value

Step 5: Draw the appropriate conclusion/decision.

In decision taken in hypothesis testing, there are only four possible outcomes. They are

1. When accept a true hypothesis. This is a correct decision.

2. When reject a false hypothesis. This is a correct decision

3. When reject a true hypothesis. This is an incorrect decision.

4. When we accept a false hypothesis. This is an incorrect decision.

However, it should be noted that we are likely to commit two types of errors in taking decision in hypothesis testing.

(a) Incorrect decision in (3) above is called a Type I error.

(b) Incorrect decision in (4) above is called a Type II error.

10.2.3 Some Common Test

*** One sample Test for the Mean**

This is given by

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

This is called Z- test or Normal test. This is appropriate when n is large (n ≥ 30 and

Population standard deviation is known. Otherwise, σ is estimated by $s = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$ the sample standard deviation.

When the samples size is very small (i.e .< 30) and σ is estimated by s, the appropriate test is given by $t-test = \frac{\overline{x}-\mu}{s/\sqrt{n}}$

This t -test is also used to test for equality of means.

*** Difference between Two Sample Means**

Under Ho: $\mu_1 - \mu_2 = 0$

The appropriate test is

$$Z = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad if \quad n \geq 30 \quad and \sigma = s$$

And

$$t = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad if \quad n_1, n_2 \ < 30$$

Example 10.1:

A machine fills packets with detergents which are summarized to have a mean weight of 28kg. a random sample of 35 packets is taken and the mean weight is found to be 24.5kg with a standard deviation of 4kg. Test the hypothesis that the sample mean does not equal to population mean at $\propto = 5\%$ level of significance.

Solution:

$$H_0 : \mu = 28$$
$$versus$$
$$H_1 : \mu \neq 28$$

$n = 35, \quad \overline{x} = 24.5, \quad s = 4$

$$Z = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{24.5 - 28}{4/\sqrt{35}}$$

$$= \frac{-3.5}{0.1143}$$

$$= -30.62$$

$$|Z| = |-30.62| = 30.62$$

But $Z_{5\%} = 1.96$

Reject Ho and conclude that the sample mean is not equal to population mean at 5% level of significance.

Note that, in testing the hypothesis for difference between two populations, if the standard deviations are not equal, we can assume equal variance by using pooled variance given as

$$\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Example 10.2: The average profit after tax of a random sample of 8 merchant banks was found to be ₦13 million with a standard deviation of N3.5million for a random sample of 12 commercial banks, the average profit after tax was estimated as N28.4 million with a standard deviation of N5.8 million. Is there any significant difference between these estimate at 5 percent level? Using

(a) unequal variance and

(b) equal variance.

Solution:

(a) Using unequal variance

Let n_1 represent Merchant Banks

Let n_2 represent Commercial Banks

$\bar{x}_1 = 13,\ s_1^2 = 3.5^2,\ n_1 = 8\ \ and\ \bar{x}_2 = 28.4,\ s_2^2 = 5.8^2,\ n_2 = 12$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{13 - 28.4}{\sqrt{\frac{3.5^2}{8} + \frac{5.8^2}{12}}}$$

$$= \frac{-15.4}{\sqrt{1.53125 + 2.80333}}$$

$$= \frac{-15.4}{\sqrt{4.33458}}$$

$$= -2.082$$

$$|t| = 2.082$$

(b) using equal variance

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where,

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(8-1)3.5^2 + (12-1)5.8^2}{8 + 12 - 2}$$

$$= \frac{7(3.5^2) + 11(5.8^2)}{18}$$

$$= \frac{85.75 + 370.04}{18}$$

$$= \frac{455.79}{18}$$

$$= 25.32$$

Hence, $s = \sqrt{25.32} = 5.03$

$$t = \frac{13 - 28.4}{5.03\sqrt{\frac{1}{8} + \frac{1}{12}}}$$

$$= \frac{-15.4 - \bar{x}_2}{5.03\sqrt{0.125 + 0.083}}$$

$$= \frac{-15.4}{5.03 \times \sqrt{0.2083}}$$

$$= \frac{-15.4}{77.462}$$

$= -0.1988$

$|t| = 0.1988$

but $t_{5\%}(d.f) =$

Decision:

Example 10.3: In the table given below:

Table 31

	Population	
	I	II
Sample size	6	11
Sample mean	12.5	15.0
Sample variance	22.8	36.4

$$\frac{5(22.8)+10(36.4)}{5+6}=31.87$$

Pooled variance $\sigma=\sqrt{31.87}=5.64$

$$test=\frac{12.5-15.0}{5.64\sqrt{\frac{1}{6}+\frac{1}{11}}}=-0.87$$

Example 10.4: A salesman selected a random sample of size 16 of his daily sales and found the average to be 10. If it is known that the population average daily sales of this agent is 8.5 with a standard deviation 5. If the test hypothesis to be tested is that the true population average daily sale of the agent does not equal to 8.5, formulate the Null and Alternative hypothesis.

$$H_0:\mu=8.5, H_1:\mu\neq 8.5$$

***One Sample Test for Population proportion:**

We use the similar procedures of one sample mean population.

The test is,

$$Z=\frac{\hat{p}-P}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}, \qquad if \quad n\geq 30 \quad and$$

$$t=\frac{\hat{p}-P}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}, \qquad if \quad n<30$$

Where P the population is proportion and $\hat{p}$ is the sample proportion.

***Difference between Two Populations proportions**

Given two Populations I and II as shows in the table below;

	Population I	Population II
Sample size	n_1	n_2
Sample proportion	p_1	p_2

Population proportion P

$$Z = \frac{(p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}, \qquad if \quad n \geq 30, \; and\, t \;\; if \;\; otherwise.$$

Example 10.5: A firm claim that 45% of the customers use its products. A market research analyst randomly arranged 500 customers and discovered that 240 customers used the product. Is the claim justified at 1% and 5% level of significance?

Solution:

$$H_0 : P = 45\% \quad against$$

$$H_1 : P \neq 45\%$$

$$\hat{p} = \frac{240}{500} = 0.48$$

$$SE(\hat{p} - P) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.48 \times 0.52}{500}}$$

$$Z = \frac{0.48 - 0.45}{\sqrt{\frac{0.48 \times 0.52}{500}}}$$

$$= \frac{0.03}{0.032} = 0.9375$$

But $Z_{5\%} = 1.96$

Accept H_0 since $Z_{calculated}$ is less than $Z_{tabulated}$ at 5% level of significance.

Example 10.6: An advertising company take a sample of 800 people and finds that 180 knows that Mercedes Germany car is fuel efficient and economical. After an advertising campaign a further sample of 920 people is taken and found that 208 know that Mercedes Germany car is fuel efficient and economical. The company claimed that there has been increase in the number of Nigerians having awareness that Mercedes Germany car is fuel efficient and economical. Can the claim be rejected at the 5 percent level of significance?

Solution:

$$H_0 : p_1 = p_2 \ versus$$
$$H_1 : p_1 > p_2$$

$$p_1 = \frac{180}{800} = 0.225, \quad q_1 = 1 - 0.225 = 0.775$$

$$p_2 = \frac{208}{920} = 0.226, \quad q_2 = 1 - 0.226 = 0.774$$

$$p_1 - p_2 = 0.775 - 0.774 = 0.001$$

$$= \frac{116}{8} - 1.59 \times \frac{70}{8}$$

$$= 14.5 - 1.59(8.75)$$

Hence,

$$Z = \frac{p_1 - p_2}{SE(p_1 - p_2)} = \frac{0.001}{20.20} = 0.00005$$

But $Z_{5\%} = 1.96$

Accept H_0 since $Z_{calculated}$ is less than $Z_{tabulated}$ at 5% level of significance. No, the claim cannot be rejected.

10.3 Analysis of Variance (ANOVA)

The Analysis is variance is used to test the equality of more than three variables. It is defined as the partitioning the total sum of squares into components and their interactions with their degree of freedom (d.f)

The hypothesis to be tested is $H_0 : \tau_i = \tau_i^{'}$ all treatments are equal against.

$H_1 : \tau_i \neq \tau_i^{'}$ for some treatment i.e. some treatments are of unequal effects.

Table 32: ANOVA Skeletal

Sources of Variation	Degree of Freedom	Sum of square	Mean square Error	F-ratio
Treatment	$t-1$	$Trt.SS$	$\dfrac{Trt.SS}{t-1}$	$\dfrac{Trt.MSE}{MSE}$
Error	$t(r-1)$	$Error.SS$	$\dfrac{Error.SS}{t(r-1)}$	-
Total	$rt-1$	$Total\ SS$	$\dfrac{Total\ SS}{rt-1}$	-

The model and the validating assumptions are taken for granted in the test – ratio

$$F = \frac{mean\ square\ error\ \ for\ \ treatment}{mean\ \ square\ \ error}$$

This ratio is called F- test i.e. F – Distribution with degree of freedom (t – 1), (rt – 1). Check the F – table under these degrees of freedom to obtain the F-tabulated table and compare the value with the value of ANOVA table to take decision at $\alpha\%$ level of significance.

If $F_{tab.} > F_{cal.}$ Accept H_o, otherwise Reject H_o.

Note that t = number of treatment and

r = the number of replicate

Example 10.7: Carry out the analysis of variance in the Table below at 5 percent level of significance?

Table 33

Treatment	Replicate
A	3, 5
B	6, 8, 12
C	4, 6, 8, 10

Solution:

Treatment	Total
A	$8,\ r_1 = 2$
B	$26,\ r_2 = 3$
C	$28,\ r_3 = 4$

$$N = 9, \quad GT = 62$$

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},\ i = 1,2,3,\ \ j = 1,2,\text{L}\ \ ,r_i$$

$$\text{Correction Factor (CF)} = \frac{GT^2}{N} = \frac{(8+26+28)^2}{9}$$

$$CF = \frac{62^2}{9} = 427.11$$

$$\sum y_{ij}^2 = 3^2 + 5^2 + 6^2 + 8^2 + 12^2 + 4^2 + 6^2 + 8^2 + 10^2$$

$$\text{Total Sum of Squares (TSS)} = 494 - CF$$

$$= 494 - 427.11$$

$$= 66.69$$

$$\text{Treatment Sum of Squares (Trt.)} \quad = \frac{\sum y_{i.}}{3} - CF$$

$$= \frac{8^2}{2} + \frac{26^2}{3} + \frac{28^2}{4} - 427.11$$

$$= 26.22$$

And that

Error Sum of squares (ESS) = Total SS - Treatment SS

$$= 66.89 - 26.22$$

ANOVA TABLE

Source	$d.f$	Sum of squares	MSE	$F-ratio$
Treatment	3-1=2	26.22	$\frac{26.22}{2}=13.11$	$\frac{13.11}{6.78}=1.93$
Error	8-2=6	40.67	$\frac{40.67}{6}=6.78$	-
Total	9-1=8	66.89	-	-

$F_{tab}=F_{5\%}(2,6)=5.14$ (from table)

We accept the Ho and conclude that the treatments have equal effect.

10.4 Contingency Table/Goodness Fit

This is also called Chi-squares goodness of fit test. This test is used for association and equality of proportions and variances. It is the test that used to show whether there is a relationship or association between two variables. The term used to describe relationship between attributes is independence.

Therefore, the chi-squares - statistic) which is test of goodness of fit is defined by;

Chi-squares $=\sum\frac{(Observed-Expected)^2}{Expected}$, $with\ (r-1)(c-1)$ degree of freedom, where

Expected $=\frac{(row\ \ total)\times(column\ \ total)}{Overall\ Total}$

where r is the number of rows and c is the number of columns.

Under Chi-squares test, H_0 is that the attributes under study e.g independence while H_1 is that the attributes under study e.g. not independence.

Example 10.8:

Educational level	Highest deposit	Low deposit
University	32	48
Secondary	40	60
Primary	75	55
Nursery	107	83

Determine whether a relationship exists between the level of deposits and educational attainment?

Educational level	High deposit	Low deposit	Row total
University	32	48	80
Secondary	40	60	100
Primary	75	55	130
Nursery	107	83	190
Column total	254	246	500

H_0 : Independence exists between level of deposit and educational attainment

Against

H_1: Independence does not exist between level of deposit and educational attainment.

The entries in the table are observed frequencies

To find the Expected frequencies we use

$$E = \frac{row\ total\ \times column\ total}{overall\ total}$$

$$E(32) = \frac{80 \times 254}{500} = 40.64$$

$$E(40) = \frac{100 \times 254}{500} = 50.80$$

$$E(75) = \frac{130 \times 254}{500} = 66.04$$

$$E(107) = \frac{100 \times 254}{500} = 96.52$$

$$E(48) = \frac{80 \times 246}{500} = 39.36$$

$$E(60) = \frac{100 \times 246}{500} = 49.20$$

$$E(55) = \frac{130 \times 246}{500} = 63.96$$

$$E(83) = \frac{190 \times 254}{500} = 93.48$$

So. that

O	E	(O-E)	$\frac{(O-E)^2}{E}$
32	40.64	74.6496	1.8369
40	50.80	116.6400	2.2961
75	66.04	80.2116	1.2157
107	96.52	109.8304	1.1379
48	39.36	74.6496	1.8966
60	49.20	116.6400	2.3707
55	63.96	80.2816	1.2552
83	93.48	109.8304	1.1749
-	-	-	13.184

Chi-square calculated is

$$= \sum \frac{(Observed - Expected)^2}{Expected}$$

$$= 13.184$$

But, Chi-square $\left[(4-1)(2-1)\right] = Chi - squares\ at 5\% = 7.81\ \ from table$

H_0 is rejected and conclude that independence does not exist between level of deposit and educational attainment.

CHAPTER ELEVEN
LABOUR TURNOVER

11.1 LEARNING OBJECTIVES:

At the end of this chapter, the students should be able to;

i. control over time-keeping and time-booking;
ii. time and motion study;
iii. control over idle time and overtime; and
iv. control over labour turnover

11.2 Introduction

Labour turnover refers to the rate at which employees leave employment. The concept is a major problem in industries and particularly in the field of organizational behaviour. In every organization, employees constantly join and leave for one reason or other. The relation between the number of persons joining the organization and leaving due to resignation, retirement or retrenchment to the average number of pay-roll is labour turnover.

The labour turnover is caused by several factors, some may be controllable, and some may not be controllable. Labour turnover may prove to be a costly process, harmful to the efficiency of worker and impairs the quality of production. Labour turnover is a serious obstacle to the full utilisation of country's human and material resources. High Labour turnover is not desirable as it affects both employers and workers adversely.

11.3 Meaning of Labour Turnover

Labour Turnover can be defined as the proportion of employees who leave an organization over a set of period (usually on 1-year basis), expressed as percentage of the total workforce number. Labour turnover is the overall change in the number of people employed in a business entity during a particular period. It takes into consideration the number of exiting personnel, new joiners and the total number of workers as listed in the payroll at the end of a given period. It can be evaluated by relating the number of employees leaving their employment during a period of time to the total or average numbers employed in that period.

Labour turnover is a measurement of the extent to which old employees leave and new employees enter into the services of an organization in a specific time period. Employees who are not satisfied with their career in the present organization may seek suitable employment in other organizations. Labour turnover may result into movements, shifting or migration of workers from one unit to another or from one industry to other.

11.4 Measurement of Labour Turnover

Labour turnover rate can be measured by the application of any one of the following three methods: Accession rate, separation rate and composite rate.

If the number of employees working in the organization remains the same, the separation rate is equal to accession rate. If the number of persons in the organization is increasing, accession rate will be greater than the separation rate. If the employment is decreasing, the separation rate will be greater than accession rate.

The term accession means, the total number of employee's added to the employment during the period (i.e., additions to the pay-roll).

The term 'separation' implied severance of employment at the instance of the employees (i.e. quits discharges, retirements, deaths etc), replacements (i.e. one accession plus separation) and average work force i.e., the number at the beginning of a period plus the number at the end of the period divided by 2.

$$AccessionRate = \frac{TotalAccessionsintheYear}{Averagenumberofemployeesfortheyear} \times 100$$

$$SeperationRate = \frac{TotalSeperationintheYear}{Averagenumberofemployeesfortheyear} \times 100$$

$$CompositeRate = \frac{TotalAccessionsperyear\ + Totalseperationperyear}{Averagenumberofemployeesfortheyear} \times 100$$

Example 11.1: In Klimanjaro Eateries Ltd. the average workforce per month is 4000 and there are 80 accession and 100 separations during that period. Calculate the labour turnout using the accession, separation and composite methods.

1. $Accessionrate = \frac{80}{4000} \times 100 = 2\%$

2. $Seperationrate = \frac{100}{4000} \times 100 = 2.5\%$

3. $Compositerate = \frac{80+100}{4000} \times 100 = 4.5\%$

Example 11.2: Company A manufactures tricycle, last year the company employed an average of 10 staff to replace 13 staff that left. What is the turnover rate?

$$\frac{13}{50} \times 100 \ = 26\%$$

Example 11.3: In a particular organisation, at the beginning of the month of January,1987 were 900 workers and at the end of the month 1,100. During the month 10 workers left, 40 persons were discharged and 150 workers were recruited. Of these, 25 workers were recruited in the vacancies of those leaving, while the rest were engaged for an expansion scheme. Calculate the labour turnover rate using the separation method.

$$Seperationrate = \frac{10 + 40}{1000} \times 100 = 5\%$$

$$Note: Averagenumberofworkersduringthemonth = \frac{900 + 1{,}100}{2} = 1000$$

11.5 Causes for Labour Turnover

Employees may leave the organisation on their own interest or they are discharged. The causes for Labour turnover may be classified as controllable (avoidable) and uncontrollable (unavoidable). A high rate of labour turnover is undesirable and there is need to study the various causes for labour turnover.

11.6 Controllable Causes (Avoidable Causes)

Among the controllable causes redundancy is the most important due to the seasonal nature of the trade or Industry, shortage of materials, lack of proper manpower, recruitment and promotion policies.

The other controllable causes are employees' dissatisfaction with the job, low wages, hours of work and work conditions and bad relations with supervisors and fellow workers.

11.7 Uncontrollable Causes (unavoidable causes)

Labour turnover also arises due to certain reasons which are beyond the control of management. Employees may leave the organisation because of the following uncontrollable causes.

i) Better Job opportunities outside the organisation

ii) Retirement, death etc.

iii) Ill health, accident etc.

iv) Unsuitable Job (or) Misconduct on his part.

v) Shifting from the locality

vi) Personal affairs like marriage or pregnancy in case of female workers.

vii) Lack of proper housing or transport facilities.

The causes of labour turnover may be ascertained through personal interview with the employees leaving the organisation and opinion poll or surveys.

11.8 Effects of Labour Turnover

A high rate of labour turnover is a great handicap to the workers and employer. Some amount of labour turnover is rather inevitable and natural. Such type of labour turnover arises on account of retirement of old employees and accession of fresh blood. Turnover of this type is very small in all organizations. In most of the causes turnover arises on account of resignations and dismissals. Such type of Labour turnover proves to be very harmful. The loss to the workers is distinct and loss to the employer is still greater.

11.9 Cost of Labour turnover to Employees

High labour turnover is harm full to the employees in many respects:

i) On account of a change in service from one organisation to another, the workers are not able to enjoy the various advantages (pay increment, leave, provident fund, pension, gratuity etc) of continued employment in one concern.

ii) A worker (employee) who changes his job quite often loses the opportunity of promotion on the basis of seniority.

iii) When worker joins new organisation, the special skills and experience relating to previous employment is of no use and become meaningless.

iv) It is very difficult to worker to adjust with the new work environment when worker joins in a new organisation.

v) Movement of workers from one organisation to another also affects the solidarity among them adversely.

11.10 Costs to the Employer

High labour turnover is costly to the employers in the following ways:

i) High labour turnover leads to high replacement costs to the employer, due to repeated recruitment, selection and placement of employees.

ii) Results into more expenditure on orientation and training of workers.

iii) Loss of production in the time interval between separation of old employees and replacement by new.

iv) Firm has to incur high over time costs to meet delivery schedules.

v) Output suffers both in quality as well as quantity.

vi) Scrap and waste rates increases due to inexperience of new workers.

vii) Due to lack of execution of orders in time, because of labour turnover the market reputation of the company will suffer.

viii) Labour turnover leads to under utilisation of human resources of the country.

11.11 Measures to Control Labour Turnover

Abnormal rate of labour turnover is bad both for the worker and the industry. Remedial measures should be taken after ascertaining the exact reasons for leaving. Labour turnover may be reduced by proper planning of manpower requirements so that it is not redundant.

Other remedial measures are:

i) Improvement in recruitment and practices

ii) Use of proper tests and interviews while selecting the personnel.

iii) Impartial promotion and transfer policies.

iv) Adequate training to the as well as existing employees.

11.12 Index Numbers

An index number is a device or tool which measures the relative change in the magnitude whether decreasing r increasing of a group of related, distinct variable in two or more situations. For examples, index numbers may be used to measure changes in prices, qualities, quantities, wages, production, salaries, employment and so on over a period of time. The time could be monthly, quarterly or yearly. In constructing an index number, a particular year's is taken as the base year or reference year and that of the other years are expressed in terms of the base year. The index number of base year is represented by 100.

11.12.1 Price index

The price index is the one that simply measures changes in the level of prices of certain goods at different places over a given period of time. For example, the price of rams in Sokoto and Federal Capital Territory.

11.12.2 Types of Price Index

We have three types of price index. These are

1. The wholesale price that simply measures the changes in the whole price of goods
2. The retail price that simply measures the changes in the retail price of goods
3. The consumer price that simply measures the changes in the prices of certain goods bought by a group of people.

11.12.3 Quality Index

This measures changes in the actual physical quality of goods. These goods may be in respect of those imported into or exported from a country of goods produced by a factory or a group of factories within the country.

11.12.4 Factors affecting the construction of Index Numbers

The following ate the problems affecting the construction of Index Numbers;

1. **The purpose of the index.** This constitutes a problem because since an index number is like a tool, it is useless unless its used for the purpose it is meant for. For example, in the case of index numbers, having decided its purpose then a decision must also be taken about the prices it would cover, that is items to be included in the index and the choice of quantities to be included.

2. **The base period**. The problem of the choice of the base period is to select a period such that this period is a most representative one. This is important since comparisons are going to be tied to it.

3**. Choice of the weight.** Weights should be attached. There are two categories of choice of weights associated with the construction of index numbers.

(a) Base year weighting index number otherwise called Laspeyres index

(b) Current year index otherwise called Paaschee index

The choice of the one over the other depends on the nature of the index.

Advantages of Laspeyres

(i) The weights are the same always

(ii) Less work is involved in its calculation

Disadvantages of base year index

(i) The weights may be out dated

(ii) The index may become inefficient.

Furthermore, current year index has only one disadvantage of when prices of certain goods rise people often tends to understates the rise in price while the base year price index tends to overstates such rise in price since its based on an out of date purchasing power.

11.13 Price Relatives

This is defined as the ratio of the price of a single good in a given period to its price in another period called base period or reference period.

$$\text{Price Relative}(PR) = \frac{p_n}{p_0} \times 100$$

Where p_n is the price in a given period

And p_0 is the price of the base year or reference period.

11.13.1 Quantity Relatives

This is defined as the ratio of the quantity of a single good in a given period to its quantity in another period called base period or reference period.

$$\text{Quantity Relative}(QR) = \frac{q_n}{q_0} \times 100$$

Where q_n is the quantity in a given period

And q_0 is the quantity of the base year or reference period.

Example 11.4:

Table 34

Commodity	Price 2000	Price 2001	Qty 2000	Qty 2001
Yam	2.85	3.20	8	15
Plantain	1.40	1.80	10	12
Beans	2.80	2.50	15	18

Calculate the simple Average of Relative Price Index using 2000 as base year.

Solution:

$$QR = \frac{15}{8} + \frac{12}{10} + \frac{18}{15} = 4.275$$

$$SAQR = \frac{4.275}{3} \times 100 = 302.5\%$$

Example 11.5: The prices of three products are tabulated as follows:

Table 35

Product	Price (Naira) 2000	Qty (kg) 2000	Price (Naira) 2001	Qty (kg) 2001
Gari	12	18	22	10
Beans	16	25	18	12
Rice	20	20	12	8

Calculate the simple average of price relative index using 2000 as the base year?

Solution:

$$PR = \frac{22}{12} + \frac{18}{16} + \frac{12}{20} = 3.558$$

$$SAPR = \frac{3.558}{3} \times 100 = 233.6\%$$

Example 11.6: Calculate the price relatives and quantity relative of the information in the

Table below:

Table 36

Year	Price(Naira)	Quantity sold (kg)
1990	100	40
1991	150	80
1992	200	100
1993	250	150

Using price and quantity in 1992 as the base period.

Solution:

$$PR(1990) = \frac{100}{200} \times 100 = 50\%$$

$$PR(1991) = \frac{150}{200} \times 100 = 75\%$$

$$PR(1992) = \frac{200}{200} \times 100 = 100\%$$

$$PR(1993) = \frac{250}{200} \times 100 = 125\%$$

Also, for quantity

$$QR(1990) = \frac{40}{100} \times 100 = 40\%$$

$$QR(1991) = \frac{80}{100} \times 100 = 80\%$$

$$QR(1992) = \frac{100}{100} \times 100 = 100\%$$

$$QR(1993) = \frac{150}{100} \times 100 = 150\%$$

11.13.2 Value Relatives

Let P represent the price of a good during a period, q represent the quantity or volume produced sold etc. during the same period, then the product of price and quantity is called value. Furthermore, if P_0 and q_0 denote the price and quantity of a good during a base period while P_n and q_n denote the corresponding price and quantity during a given period respectively, then the value relative index I define as

$$\frac{V_n}{V_0} = \frac{p_n q_n}{p_0 q_0}$$

$$= \frac{p_n}{p_0} \times \frac{q_n}{q_0}$$

$$= (\text{PR}) \times (QR)$$

11.14 Methods of Constructing Index

11.14.1 Simple Aggregate Price Index

define as $\frac{\sum p_n}{\sum p_0} \times 100$, where $\sum p_n$ is the sum of all good price pn the given period and

$\sum p_0$ is sum of all good price in the base period. This method is very easy to calculate but has the following disadvantages

(i). It does not take into account the relative importance of the various good (ii). The particular unit used in price quotation affects the value of the index.

11.14.2 Simple Average of Relative Method

Here several possibilities exist depending on the procedure used to average price relatives such as arithmetic mean, harmonic mean, geometric mean, weighted arithmetic mean, median, etc.

For instance, simple arithmetic mean of relative price index is defined as

$$= \frac{\sum \frac{p_n}{p_0}}{N}$$

Where $\Sigma(P_n/P_0)$ is the sum of all commodity price relative

N= Number of commodity price relative used.

11.14.3 Weighted Aggregate Method

Here, three possible formulae occur depending on whether base period, given period or typical period quantities denoted respectively by q and q are used.

(a) Laspeyre's Index or Base Period Method: Weighted aggregate price index with base year quantity weights

$$\text{LPI} = \frac{\sum P_n q_0}{\sum P_0 q_0} \times 100\%$$

(b) Paasche's Index or Current/ Given Period Method: Weighted aggregate price index year quantity weights

$$\text{PPL} = \frac{\sum P_n q_n}{\sum P_o q_n} \times 100\%$$

(c) Typical Period Method: weighted aggregate price index with typical period quantity weights

$$= \quad = \frac{\sum p_n q_t}{\sum p_0 q_t} \times 100$$

11.14.4: Fisher's Ideal Index

This is the geometric mean of Paasche's and Laspeyre's index number. It is defined as

$$\text{FII} \quad = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}}$$

Example 11.7:

Four commodities A, B, C and D are used in th computation of consumer price index (CPI). The data is as follows:

	1990		1995	
	Price	Quantity	Price	Quantity
Commodity	(P_0)	(q_0)	(P_n)	(q_n)
A	2.00	16	4.00	12
B	0.80	24	1.20	24
C	3.20	10	8.00	6
D	0.50	32	0.75	40

(i) Calculate Laspeyres and Paasches price indices. Why would you hesitate to adopt any of the two alternative indices as authentic?

(II) Compute and identify a compromise alternative index to the Laspeyre's and Paasche's index.

Solution:

Commodity	P_0	q_0	P_n	q_n	$P_n\, q_0$	$P_0\, q_0$	$P_n\ q_n$	$P_0\, q_n$
A	2.00	16	4.00	12	64.0	32.0	48.0	24.0
B	0.80	24	1.20	24	28.8	19.2	28.8	19.2
C	3.20	10	8.00	6	80.0	32.0	48.0	19.2
D	0.50	32	0.75	40	24.0	16.0	30.0	70.0
	-	-	-	-	196.8	99.2	154.8	82.4

$$\therefore \text{LPI} = \frac{\sum P_n q_0}{\sum P_0 q_0} \times 100\%$$

$$= \frac{196.8}{99.2} \times 100\%$$

= 198.4%

$$\text{And PPI} = \frac{\sum P_n q_n}{\sum P_o q_n} \times 100\%$$

$$= \frac{154.8}{82.4} \times 100\%$$

= 187.9%

(ii) Do it yourself (DIY).

(iii) A compromise alternative index is Fisher's Price Index

$$\text{FPI} = \sqrt{1.984 \times 1.879 \times 100}$$

= 193.1%

PART B - SOCIAL RESEARCH METHODS

CHAPTER TWELEVE
THE NATURE AND IMPORTANCE OF RESEARCH

12.1 Learning Objectives:

At the end of this chapter, readers should be able to;

i. identify research problem; and
ii. performing a literature review and write a theoretical/conceptual framework.

12.2 Introduction

Different authors define research in different ways, but we adopt the very simple explanation by Johnson (1996) who defines research as "A focused and systematic enquiry that goes beyond generally available knowledge to acquire specialized and detailed information, providing a basis for analysis and elucidatory comment on the topic of enquiry. The underline words are some of the basic characteristics of any type of research. Research simply means "To search "for new ideas.

12.3 The Importance of Research

The following are the major importance of research. These are:

(i) It generates novel ideas, builds credibility and develops understanding and analytical skills.
(ii) It helps to achieve your goal.
(iii) It reveals new ideas and facts.
(iv) It develops understanding and decision making.
(v) It helps understanding what's not working.
(vi) It builds credibility.

Finally, research is important to build a better understanding, decision making and discerning and analytical ideas to facilitate better results.

Types of Research

Basically, researchers are of two major types. They are basic research or pure research, and Applied Research.

1. **Basic Research**. This is generally for extending the base and frontier of knowledge in an area of study with little or no practical application.
2. **Applied Research**. This type of research aims at finding solutions to problems that are practical in nature. This means such research is action – oriented. This type of research may come either through personal observation and experience, or may be contracted out by someone, a body, an agency of government with a view to solving a given problem. It should be noted that, in basic research, the researcher or someone conducting a research does not know what his findings would be, he knows only the direction in which he \she is looking. In contrast with applied research, the researcher knows what he wants and tries to find out how far he can get there.

12.4 Guidelines for Selecting a Re-searchable Topic

There are five (5) major ways of identifying and selecting and starting a researchable topic. These are;
(i) Identify a general problem area of study.
The research must be an important issue to the researcher, his immediate environment and society at large since research involves time, efforts, resources both in human and materials.

(ii) Narrow down the general problem area to tentative specific ones. That is, the research question should be directly related to the problem the research is designed to find solution (s) to. If there is more than one question, each must be related the others and all the questions must be related to the overall issue or question.

(iii) Review relevant literature: A good research question must be answerable. That is, the research must have good chance of achieving the set objective, and if in the end it cannot, there must be very good reasons why it could not.

(iv) State problem clearly and concisely. This means that the problem must be clearly stated in understanding way.
After stating the research problem clearly, the next thing is to evaluate or appraise the extent or scope of the problem with a view to determining what has been done in a particular area of research and if there is any gap to be filled. The researcher needs to find out this through literature research review. A through literature search is an indispensable component of all

research, since it familiarizes the researcher with both the research that has been done in the area of study as well as for current research.

CHAPTER THIRTEEN
PLANNING THE RESEARCH

13.1 Learning Objectives:

At the end of this chapter, the students should be able to;

i. identify the research problem;
ii. performing literature review and write theoretical/ conceptual frame work;
iii. describe the design or approach to the problem;
iv. collect and analyse data and/ or design and validate a research design; and
v. draw conclusions and give recommendations.

13.2 Writing a Research Proposal

A research proposal should contain all the key elements involved in the research process which include sufficient information for the readers to digest and understand. In the planning of any type of research about seven (7) different stages are involved which form the basis of any research method.

The topic will therefore be structured in line with these stage that make up a typical research process, which are :

(i) Identifying the research question / problem. A research problem is the situation that causes the researcher to feel apprehensive, confused and ill- at- ease. It is the demarcation of a problem area within a certain a certain context involving the "WHO" or "WHAT", the "WHERE", the 'WHEN" and "WHY" of the problem situation.

There are three sources that usually contribute to problem identification namely,

(1) the researcher's own experience or the experience of other.

(2) Scientific literature. In this type of sources, the researcher may read about certain findings and notice that a certain field is not covered or noticed that a gap still remain to be filled.

(3) Certain shortcomings in previous theories could be a source of research, either to clarify, disprove or confirm such theories. To clarifying certain contradictory findings, at correcting a faulty methodology, faulty method of data collection, not using a suitable statistical technique, at reconciling conflicting opinions and so on, that emanating from previous researches.

(ii) Critical Appraisal of the research question /problem. Evaluate or appraise the scope of the problem with a view to determining what has been done in a particular field of study, if there is any gap to be filled.

(iii) Starting the research objectives. This is the statement of the objectives of the research. This is the objectives of s research project given a summary of what is to be achieved by the study. This is closely related to the statement of the problem by considering the body of facts through literature search.

In research objective, we have as general objective which refers to what the researcher expects to achieve by the study in general terms. But if the general objective is broken down into smaller and logically connected parts, this is referred to as specific objectives. Furthermore, when the objective is properly formulated, research objectives help to guide the researcher to develop research methodology, data collection, data analysis, and data interpretation. The essence of formulating the research objectives is that, it helps the researcher to be focus on the study, avoid the collection of data which are not strictly related to the study and finally organize the study in clearly defined parts.
Also, the research objectives must be measurable, achievable and feasible within the time frame and facilities with resources were available.
Furthermore, the use of action verbs that are specific enough to be evaluated such as, to find, to determine, to compare, to calculate, to describe and to established etc.

(iv) Research Design/Research Hypothesis. Research design is the glue that holds a research study together. It is used to structure the research to show how all the major parts of the research including the samples or groups, measures, treatments or programes and methods of their assignment work together to address the central research question. This research design is used to test for the validity of the result / output.
The following are the elements of research design. They are observations or measures, treatment, or programs, Groups, Assignment to group, Data.

13.3 Types of Research Designs

1) **Philosophical/ Discursive**: This covers a variety of approaches, but will draw primarily on existing literature, rather than new empirical data. This type of study could examine a particular issue, perhaps from an alternative perspective.
2) **Literature Review**: This is an attempt to summarize or comment on what is already known about a particular topic. By searching different sources together, synthesizing and analyzing critically, it essentially creates new knowledge or perspectives. Where a literature field is not sufficiently well conceptualized to allow this kind of synthesis, or where findings are largely qualitative (not adequately quantified), it may not be appropriate to attempt a systematic review.
3) **Case study**: This involves collecting empirical data, generally from one or some number of cases. This usually provides rich detail about those cases, of a predominantly qualitative nature. A case study aims to provide an insight into a particular situation and often stresses the experiences and interpretation of those involved. This may generate new understandings, explanations or hypothesis; this is because the data collected through this type of study design is not a true representativeness to generalize for the entire population.
4) **Survey**: A survey is a case where an empirical study involves collecting information from a larger number of cases, perhaps using questionnaires. A survey might make use of already available data that are collected for another purpose. A survey design may be cross- sectional (data collected at one time) or longitudinal (data collected over a period) this involve some quantitative analysis. This survey design is a measure of reliability of any instruments used.
5) **Evaluation**: An evaluation of a curriculum innovation or organisational change for example. An evaluation can be formative (designed to inform the process of development) or summative (to judge the effects): If evaluation relates to a situation in which the researcher is also a participant. It may be described as "action research". Evaluations will often make use of case study and survey methods while summative evaluation will ideally use experimental methods.
6) **Experimental:** This is the deliberate manipulation of an intervention in order to determine its effects. The intervention might involve individual, pupils, teachers, schools or some other unit. An experiment may compare a number of interventions with each other or may compare one (or more) to a control group.

CHAPTER FOURTEEN

LITERATURE REVIEW

14.1 Learning Objectives:

At the end of this chapter, the students should be able to;

i. Carry out a systematic literature review;
ii. formulate strong research questions;
iii. select relevant literature;
iv. draw conclusions and recommendations for further studies; and
v. write a review in academic English.

14.2 Purpose of Literature Review

1. Selection of relevant literature helps develops the skills of the researcher in the following areas:
   (i) It develops the ability to recognise and relevant resources, without getting lost in trivialities.
   (ii) It helps in gauging the quality of research material and in planning the research accordingly.
   (iii) It develops a critical altitude regarding others' research..
   (iv)It trains the researcher to be an astute observer.
   (v) knowledge of relevant literature helps the research to define the boundaries of his field .
2. Literature study assists the researcher to:
   (i) Select a research problem or theme. Relevant literature enables the researcher to discover specific gaps like inconsistencies, wrong designs and incorrect statistical conclusions.
   (ii) Define the boundaries of his field.
   (iii) Establish the size and extent of his research.
   (iv) Avoid unnecessary and unintentional repetition of research already undertaken.
   (v) Formulate his hypothesis with clearer insight.
   (vi) Carry out his research more purposefully.
   (vii) Better evaluate the significance of his own findings and discuss his results in

the light of previous studies.

CHAPTER FIFTEEN

DATA ANALYSIS AND REPORT WRITING

15.1 Learning Objectives:

At the end of this chapter, the students should be able to;

i. analyse and interpret data using ethically responsible approach; and
ii. use appropriate modules of analysis, to access the quality of inputs.

15.2 Data Analysis

Any good research will generate large volume of data depending on many factor including the population and sample size, number of treatments and variables, experiment design among others. Unless appropriate methods of analysis are used, the researcher may not come up with valid conclusions, Data collected are called raw data and such data are useless and cannot achieve research objectives until they have been processed in order to generate useful information.

Once research has been conducted and necessary data collected, the next stage is the collation, summary and analysis of data. Data analysis is very important since it allows for the interpretation of such data and guides the researcher in making deductions, conclusions, and recommendations based on the research. Data analysis can be done using different method that will help him achieve his research objectives. The researcher must have clear understanding of data collected so that the appropriate method of analysis could be used.

15.2.1 Application of Statistics in Data Analysis

Statistical analysis has found great usefulness in the treatment of research data. Thus, the researcher must find out which of the available statistical methods would be best suited to answer his research questions. **Statistics** is a mathematical science pertaining to the collection, analysis, interpretation, and presentation of data. It is applicable to a wide variety of academic disciplines, from the natural and social sciences to the humanities, science, technology and engineering and even to government and business.

15.2.2 Application of Appropriate Statistical Techniques

The following are the importance of selecting appropriate statistical techniques:

(i) It will aid adequate and appropriate sampling of sample population upon which the results of the research will be based.

(ii) It allows the use of appropriate research hypothesis and experimental design.

(iii) Allows for the testing of validity and reliability of test equipment and instruments.

(iv) Enables the researcher to appropriately allocate treatments to experimental groups in experimental research.

(v) Allows the incorporation of standard error which will determine the accuracy of methodology adopted.

(vi) It allows the making of valid inferences and conclusions about the research results.

The above benefits of statistics can only be achieved if the research data are analysed with appropriate statistical method. Different methods are available so the researcher must be conversant with the nature and application of each technique in order to use them successfully.

Statistical methods are classified into two main groups, namely; descriptive statistics and inferential statistics.

15.3 Descriptive Statistics

Descriptive statistics is the analytical method for organising, classifying, and summarising numerical data. We use descriptive statistics simply to describe what is going on in our data.

Some of the parameters for describing raw data include: frequency distribution such as histogram, frequency polygon, cumulative frequency curve and measures of central tendency like mean, mode, median, standard deviation etc. The choice of which of these to use will depend on the volume of data, the type of research design adopted, the objective of the research and the knowledge of statistical methods.

Frequency distribution is particularly good when the data to be handled is large. In such cases, the data is classified by creating class intervals and constructing a frequency distribution Table, which can then be used to construct histograms and distribution curves. This type of statistics may not provide much information about data, although the organisation of data in form of frequency distribution gives some indication as to the nature of data.

Application of central tendency will give better information about data, and allow more interpretations to be made about research results. The arithmetic mean for example, would appropriately describe the different types of data especially those collected at interval and ratio scales. Measure of variability or dispersion is another important characteristic that describe the distribution of data. The measure of dispersion indicates the distribution of a data set. Among the commonly used measures of dispersion are range, mean deviation and standard deviation. The standard deviation is the most widely used index of variability because it is the most stable, and all the observations in a data set are incorporated in the analysis. The standard deviation is particularly useful in replication of test to determine the standard error of experiment. It is related to the arithmetic mean by the expression:
$\bar{x} \pm S.D$

15.4 Inferential Statistics

In many research studies, every item in a population or batch cannot be taken as research subjects or samples. This is particularly true of researches that use questioning methods, surveys, and in experimental research where the inspection method is destructive. In such cases, the research study is conducted on samples which are carefully selected from the populations or batches. It is assumed that data collected from the representative samples are an expression of the respective parameters of the population from which sample were drawn. This means that deductions or statements regarding the broad population are made based on known samples measured. In inferential statistics, inferences about the population on the basis of the result of the result of the samples are made within the limit of specified probability.

This means that the inferences are made within the fluctuations due to sampling and some degree of uncertainty in their results. The process of drawing inferences about a population is the basis of inferential statistics. With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. We use inferential statistics to make judgment of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in the study. Thus, we use inferential statistics to make inferences from our data to more general conditions.

The common method in inferential statistics includes estimation predictions and hypothesis testing. Most of the major inferential statistics come from general family of statistical models known as the General Linear Model. This includes the test, Analysis of Variance (ANOVA), Analysis of covariance (ANCOVA), regression analysis, and many of the multivariate methods like factor analysis, multidimensional scaling, cluster analysis, discriminate function analysis and so on. Given the importance of the general Linear Model, it's a good idea for any serious researcher to become familiar with it workings.

This is a major problem of most researchers including this presenter. Many researchers especially from the social sciences have developed a lot of phobia for statistics and the limits the inferences and conclusion made from even very excellent research studies. It is very important that researchers should have a good understanding of basic statistics if they are to make good use of their research results.

Fortunately, there are now computer software packages available to lessen the mathematical burden on researchers. Some of these packages include statistical package for the social sciences (SPSS), Minitab and sign-test. They provide a number of functions such that once the data are input, different tests and calculations can be made.

15.5 Dissemination of Research Output

The last stage in a research process is the statement of the research results with either the researchers' peers or the general public. The success of a researcher is not measured only in his ability to carry out his research work, but also in his ability to communicate and disseminate the outcome of the research to his colleagues, the end user of the results and the society at large. A scientific research will not be regarded as completed until the results are made public through scientific publication.

15.5.1 Common Type of Scientific Papers

1) **Seminar Paper**
2) **Research Proposal**
3) **Conference Paper**
4) **Journal Articles** (To be published in learned journal after necessary peer review processes).
5) **Dissertation and Theses** (For the purpose of earning a Degree or Certificate).

It could also be in form of:

Newsletters, Monographs, Books

The structure and major components of a Scientific Paper include:

a) **Title** (Common to all Scientific Paper)
b) **Authors and their Affiliations or Addresses** (Mainly for journal articles and conference paper
c) **Abstract** (Common to all Papers)
d) **Introduction** (Common to all)
e) **Literature Review** (Mainly for Dissertations and Theses)
f) **Materials and Methodology** (Applicable to all Papers)
g) **Results and Discussion** (Common to all Scientific Papers)
h) **Conclusion and Recommendations** (Common to all Papers)
i) **Acknowledgements** (Common to all Papers)
j) **References** (Applicable to all Papers)
k) **Appendices (**Mostly applicable for Dissertation and Theses)

Title:

i. This is the topic of the study or scientific paper which gives a concise statement of the main topic.

ii. It is the first point of contact by those who may not even read the paper, but it could compel or discourage readership.
iii. It should normally be of about **12 to 15 Words**
iv. It should refer to the major variables or theoretical issues under investigation.
v. The title should be self-explanatory and unambiguous, even to those who are not experts in the area of study.
vi. It should not contain abbreviations or words that do not serve any useful purpose.

e.g. Do not use **Props.** To represent **Properties;** or "A study of" ………; or: An experimental investigation of". These will only increase the length of the title unnecessarily.

Authors' Name(s) and Affiliation(s)/Address (es)

The name(s) and contact address(es) of those that have made useful contributions to the planning and execution of the research work should be included on the title page. This is necessary so that the researcher(s) could be cited appropriated, and reached by readers who may need further clarifications or copies of the research paper.

Abstract:

This is regarded as a mini paper which provides a brief summary of each of the sections of the entire study. The number of words varies from one editorial board to another, but generally, it should normally be between 150 and 200. It should be written in past tense, clear and simple language without abbreviation or acronyms

Introduction:

This includes the nature and scope of the research problem or question being addressed. It also features a review of relevant background literature on the subject matter. It should orientate the readers on the need for the study and state briefly, but clearly, the objectives of study.

Materials and Methods:

This describes the experimental design and methodology used in carrying out the study as well as the step-by-step methods within the methodology adopted. To save space, details of standard methods are not necessary, but if any modification is used, such modification should be given. Sources of standard methods used should also be provided.

Results and Discussion:

This section caters for the presentation the findings of the research without repeating experimental details. Results may be presented in Tables or figures, but the same result should not be presented in both Tables and figures. Results should be presented with clarity. All Tables and figures should be properly labeled with appropriate titles to provides clues as to what each represents.

The conclusion should present the principles, relationships and generalisation arrived at by the results. Any exception or lack of correlations with previous studies should be pointed out with possible reasons for such lack of correlations. Also indicate how the results of current study agree with previous investigations. Appropriate reference with which results are compared should be acknowledge through proper citing in the discussion using the appropriate format.

Conclusion:

This gives the extent to which the objectives of the study have been achieved in clear, specific and general terms. The deductions made from the study should be given and also any further work that needs to be done to extend the frontier of knowledge in the area of study should be identified.

Referencing:

This section provides the lists of all the materials that were consulted in the course of the study. They may include journals, books, conference/seminar papers, personal communications or unpublished works. The style of presentation will depend on the editorial policy of the journal outlet. Two methods of referencing are common.

(i) **The American Psychology Association (APA) System:**

In this system, a reference is indicated in the body of the report in form of author(s) name(s) and the year of publication. For instance, Akinjayeju (2008), depending on where the reference appears in a sentence. For two authors, it is indicated as Azike and Akinjayeju (2008) or (Azike and Akinjayeju, 2008), respectively. In this system, the references are compiled under list of references by arranging them in alphabetical order using the surnames and initials of authors.

(ii) **The Modem Language Association (MLA) System**

This is also called the Author-and-No format in which references are indicated with numbers in superscript and in parenthesis in the body of the report. For example: The following statement is quoting the result of a previous study.

"It has been previously reported that nutritional status or students affects their academic performance" (1).

The author(s) of such report is/are indicated by the number 1. The list of references in this system is compiled serially and numbered as each reference appeared in the body of the work. The format for listing of the different literature items vary from one journal to another, so the researcher(s) and prospective author(s) should be familiar with the policy of each editorial board they wish to patronise.

GENERAL EXERCISES

1. Which of the following is NOT an element of secondary data?

   A. It is not time consuming

   B. It may not be reliable

C. It is readily misleading

D. It may not be accurate

2. Sampling can be broadly classified into TWO namely

A. Random Sampling and Probability Sampling

B. Random Sampling and Non-random sampling

C. Non-random sampling and Non-Probability sampling

D. Probability sampling and Simple random sampling

3.. If the coefficient of variation and the mean of set a of observation are 30% and 15 respectively, find the variance of the data set.

A. 15.25

B. 20.25

C. 22.25

D. 22.50

4..The rule of degree of association (r) between x and y is that, the value of r

A. $0 \leq r \leq 1$

B. $-2 \leq r \leq 2$

C. $-1 \leq r \leq 1$

D. $-1.5 \leq r \leq 2$

5.If the 25^{th} percentile n d 75th percentile of a data set are 10.5 and 45.5 respectively, find the Quartile deviance?

A. 0. 325

B. 0. 425

C. 0. 550

D. 0. 625

6.Find the variance of a data set 20, 12, 22, 16, 10?

A. 19.8

B. 20.8

C. 22.8

D. 22.6

7.The simplest measure of variability is the -------------.

A. Variance

B. Standard Deviation

C. Range

D. Quartiles

8. An exercise consists of rolling two dice.

Find the probability that the sum of the numbers thrown is 9?

A. 8/36

B. 6/36

C. 4/36

D. 3/36

9. If the P(A) = 0.5, P(B) = 0.4 and P(A/B = 0.2.

A. 0.62

B. 0.72

C. 0.82

D.0.97

10. The probability that a Broadcasting television A will transmit World Cup Live is ¾ and the probability that a broadcasting television B will transmit same programme live is 2/5. What is the probability that both will not transmit?

A. 3/16

B. 6/25

C. 3/20

D. 9/20

11. Find the number of arrangements of the letter in the World MERCEDES?

A. 5520 ways

B. 6720 ways

C. 7500 ways

D. 8720 ways

12. How many different Committees of 3 men and 2 women can be formed from the Advisers Association comprising of 5 men and 4 women?

13.The probability that a farmer who enjoys an agricultural loan from a bank will became a bad debtor is 0.04. If a bank has 200 farmers as customers, find the probability that 5 of them will become bad debtors.

A. 0.0511

B. 0.0512

C. 0.0611

D. 0.0612

14. At a given period I a certain Health Centre, there are 6 expectant mothers. What is the probability that 3 boys were delivered?

A. 0.6105

B. 0.5105

C. 0.4105

D. 0.3125

15. Binomial distribution can be approximated to a poission’sdistribution if,

A. “n” is small and “P” is large

B. “n” is small and “P” is very small

C. “n” is and “P” are too large

D. “n” and “P” are small

16. The mean arrival rate of customers in a banking hall is 10 persons per hour. Find the probability that3 persons will arrived the banking hall within an hour?

A. 0.0762

B. 0.0672

C. 0.0076

D. 0.0005

17. If the score can be consider as having a normal distribution with mean grade of 86 and a standard deviation 8. Where do you stand regarding the performance on this test if your score is 86?

A .0.72

B. 0.82

C. 0.60

D. 0.50

18.Acceptance of the null hypothesis when it is false is the

A. Type I error

B. Type II error

C. Rejection region.

D. Both Type I and Type II error

19. A sample of an items is randomly selected from a normal population with mean 7 and standard deviation 5. If the sample mean is 8, find the value of test-statistic.

A. 0.55

B. 0.60

C. 0. 65

D. 0.70

20. Goodness of fit – test is called a

A. Ƶ – test

B. t – test

C. χ^2test

D. f – test

21. Given the table below

A	b	c	D
E	f	g	H
I	j	k	L
M	n	o	P

The degree of freedom (df) in the above table is ------.

A. 16
B. 12
C. 9
D. 4

22.Given the Anova table below:

Find the values of "a" and "b"?

Source	Df	Ss	Ms	F
Treatment	2	"b"	39	14.67
Error	"a"	16	2.667	-
Total	8..	94	-	-

A. 8 and 16 respectively
B. 6 and 78 respectively
C. 78 and 6 respectively
D. 16 and 8 respectively

23..In an hypothetical observations between paired (x, y), we have given that $\sum x$=20, $\sum$y =13, $\sum x^2$ = 108, $\sum xy$ = 69, n = 4. Determine the coefficient of regression?

A. 0.7

B. 0.5

C. 0.4

D. 0.2

24. The measure that explains how well the regression line fits a set of data on daily sales of a product is known as coefficient of

A. Correlation

B. Product moment correlation

C. Variation

D. Determination

25. A linear relationship for the management of an organization, when expenditure (y) in is regression on the revenue (x) in (₦'000), given by y = 42.113 + 0.499x. Obtain the estimate of expenditure when the revenue is

A. ₦18,500

B. ₦17,806

C. ₦17,500

D. ₦15,806

26. If the difference between 4 – paired of observations $x$and y are -4 -3 -5 6. Determine the coefficient of correlation between x and $y$using rank method.

A. 0.44

B. 0.23

C. 0.14

D. 0.86

27. The perfect direct correlation coefficient referred to when to when the value of r is

A. r -1

B. r = 1

C. r < 1

D. r > 0

28.The difference between a sample statistic and its corresponding population parameter is called ---------.

A. Biased

B. Unbiased

C. Sampling error

D. Sample error

29. The expectation of the sample mean is the -------.

A. Sample statistic

B. Population Mean

C.Sampling Mean

D. Population parameter

30.An interval with which we accept the true value of the parameter being estimated to be is known as ---------.

A. Confidence region
B. Significance level
C. Confidence Interval
D. Confidence Limits

31. Construct 95% Confidence interval for population given that the sample size of 100 has proportion 0.25 and standard deviation 10.?

A. $2.65 < P < 3.19$

B. $-2.65 < P < 3.19$

C. $0.75 < P < 0.75$

32. A manufacturing company employed 600 personnel in 2018. During the first quarter of the year, 60 employees left the company. Calculate the Labour turnover per annum?

(A) 80%

(B) 70%

(C) 60%

(D) 40%

REFERENCES:

Adegoke, N (2012) Research Methods in Social Sciences, Prime Target Limited.

Adeleke,J.O (2010) The Basics of Research and Evaluation, Lagos So-merest Ventures.

Akinbo, R. Y (2004) Descriptive Statistics (2004): Deceptive Statistics

Wale Prints Enterprises.

Akinjayeju, O. (2008) Research Method Workshop on Grant – Oriented Proposal and Conducting Research, Yaba College of Technology, Yaba – Lagos.

Akinjayeju, O. (2004) Fundamentals of Human and Applied Nutrition, Concept Publications Ltd., Lagos.

Akinjayeju, O. (2005) Commercialization of Research Results. Paper Presented at a Workshop on Modem Trends in Presentation of Technical Papers for Academic Staff, Yaba College of Technology, Lagos.

Akinjayeju, O. (2008a) Scientific Paper Writing and Presentation (Title, Acknowledgement and Appendices). Paper presented at a departmental Seminar, Department of Food Technology, Yaba College of Technology, Yaba. March 14, 2008.

Berenson, M.L., Levine,D.M., Krehbiel,T.C (2009) Basic Business Statistics., Concepts and Applications, 11th Edition. Prentice Hall, Upper Saddle Rover, NJ.

Famuyiwa, L and Eleyowo, S (2012) Introductory Statistics:Ebensun Ventures

Idowu, M. A. (2008) Personal Communication.

Lekan Famuyiwa and Segun Eleyowo (2012) Introductory Statistics Ebensun Ventures.

Nwachukwu, V.O (2010) Introduction to Statistics, 7th Ed. zelon Enterprises, Port Harcourt.

Oloyo, R. A. (2001) fundamentals of Research Methodology for Social and Applied Sciences ROA Education Press, Ilaro.

Oyewole, O. B. (2001) Scientific Writing and Presentation in L. O. Sanni (Ed.) Research Methods for Young Scientists Jedidah Publishers, Abeokuta.

Oseni, B. A, (2008) Applied Statistics for Managerial Studies. Revised Edition, Highland Publishers.

White, B. (2000) Dissertation Skill for Business and Management Students

Continuum Publishers Ltd., London.

Walepole, R.E. (1974), Introduction To Statistics, 2nd Ed; Macmilliam pub. Co. Inc; New York.

V.O. Nwachukwu, C.C Chime and P.N. Dekwere (2014) Introduction to Business Mathematics (1st Edition). Zelon Integrated Services Limited.

www.evaluationtools.org/plan.data collect.asp
www.petech.ac.za/robert/data.html
www.petech.ac.za/robert/Datalinterpr.html
www.psywww.com/tipsheet/labrep.htm