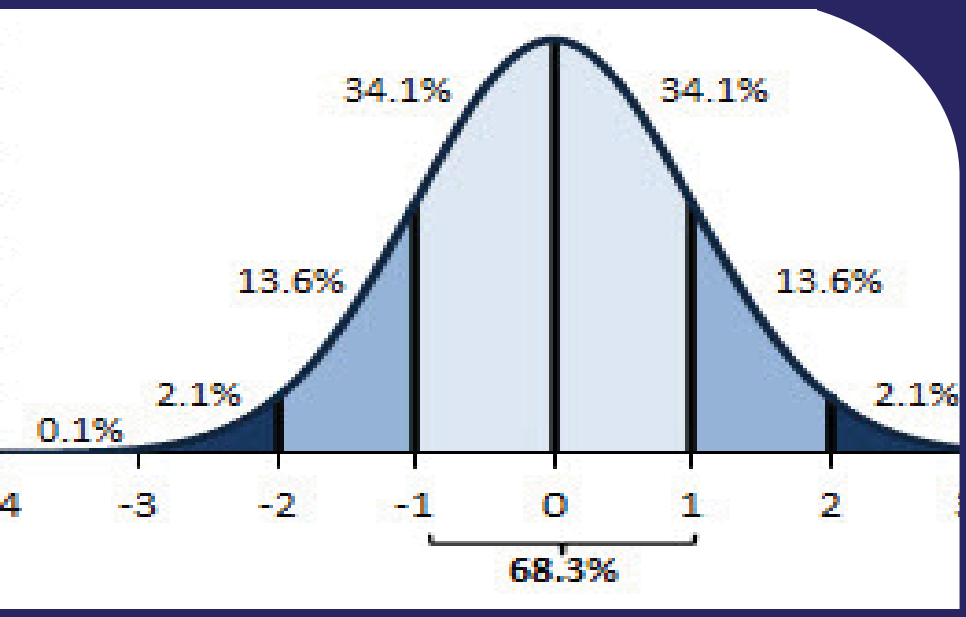
2023
EDITION



Quntitative Techniques

STUDY TEXT

CPA (U)

PAPER 3

QUANTITATIVE TECHNIQUES

STUDY TEXT

This study textbook has been designed to help students pass their CPA(U) Paper 3 exams. It has been published in close consultation with lecturers and tutors who have extensive experience in teaching this subject. Additionally, the content has been updated to reflect the new changes in the CPA(U) Syllabus that took effect on 1st January 2023. This book contains all the essential information needed to help you pass your exams with flying colours.

The study text therefore;

- Highlights the most important elements in the CPA(U) Syllabus and the key skills you need.
- Emphasizes how each chapter links to the CPA(U) Syllabus and the study guide.
- Provide a lot of exam focus points demonstrating what is expected of you in the exam.
- Emphasizes key points in regular fast forward summaries.
- Examine and Text your knowledge in our activity exercises provided at the end of each chapter.

P.O. Box 9449 Kampala

TEL: 0772690601 / 0701690601 / 0773385925

WhatsApp: 0786499326

Web: www.harvestuganda.com

Email: admin@harvestuganda.com

Sixth Edition 2023
Printed by Harvest Uganda Ltd
P.O.BOX 9449 Kampala (u)
TEL: 0701690601 / 0701690601 / 0786499326

THIS STUDY TEXT INCLUDES THE FULL CPA(U) QUANTITATIVE TECHNIQUES SYLLABUS

This book has been specially written and designed to help students preparing for their CPA(U) QUANTITATIVE TECHNIQUES Paper 3 professional examinations and those undertaking a BUSINESS STUDIES and BSC ACCOUNTING AND FINANCE course at the University. It is also useful for lecturers preparing students for those examinations.

This book covers a variety of contents, exercises, model questions, test papers, mixed exercises and answers to equip and prepare students for their examinations.

This book contains

- ✓ Progressive notes targeted to the entire Financial Accounting Syllabus and study Guide.
- ✓ Review exercises and Questions to check your understanding
- ✓ Clear layout and style designed to save your time
- ✓ Plenty of examination Questions from a variety of past papers.

We are grateful to all the examination councils for their permission to reproduce past examination questions.
The suggested solutions to the Illustrative questions have been prepared by the whole entire author Team.

TABLE OF CONTENTS

PART E: FORECASTNG TIME SERIES ANALYSIS

PART F: INDEX NUMBERS

PART G: NETWORK ANALYSIS

PART A

STATISTICAL DATA, PRESENTATION AND MEASURES

1.0
DATA COLLECTION

1.1 INTRODUCTION

Data may refer to information, statistics, facts, figures, numbers, or records. Depending on the source of information, data is called primary if it is collected afresh and for the first time (original in character) or secondary if it has already been collected by someone else and has already been passed through the statistical process. Primary data should be collected only if secondary data is not available. The techniques described below describe the different ways in which primary data can be collected.

1.2 DATA COLLECTION METHODS / TECHNIQUES

There are many methods of collecting primary data and the main methods include:

(i) Observation
(ii) Interviews
(iii) Questionnaires
(iv) Case studies
(v) Focus group interviews
(vi) Diaries
(vii) Critical incidents
(viii) Portfolios

(i) OBSERVATION

Here, data is collected through direct observation without asking respondents anything. For example, the behaviour of buyers of a certain commodity can be studied through observation.

Advantages

- If done accurately, subjective bias is eliminated
- Information on what is happening at the time of data collection is observed and attained.
- It demands less active co-operation between the person collecting data and respondents.
- It is suitable for respondents who have no verbal ability.

Disadvantages

- It is expensive
- It provides limited information
- Unforeseen factors may interfere with the observation task
- Some phenomena (fact or experience) are unobservable

(ii) INTERVIEW METHODS

These involve presentation of oral verbal stimuli (incentive, spur or motivation) seeking oral-verbal responses. There are four main types of interviews namely;

a. Structured; - These involve use of structured questionnaires.
b. Semi-structured interviews;-These involve the use of checklist of terms or topics to be discussed with respondents. As they respond to listed items, more questions emerge and are asked besides the prepared ones e.g. Focus group discussions (FGD).
c. Unstructured interviews; - This type of interview involves talking casually with respondents, usually key informants, on particular issues without prior appointment or telling them the purpose of the observation.
d. Interview schedules; - These are interviews in which questionnaires are used and filled in when interviewing respondents.

Advantages

- There is good response rate.
- The method is complete and immediate.
- Possible immediate questions.
- Interviewer is in control and can give help if there is a problem.
- One can investigate motives and feelings.
- The interviewer can use recording equipment.
- The characteristics of respondents can be assessed i.e. the tone of voice, facial expression, hesitation e.t.c.
- The interviewer can use props.
- If one interviewer is used, there is uniformity of approach.
- The interview method is used to pilot other methods.

Disadvantages:

- Need to set up interviews.
- Time consuming.
- Geographic limitations.
- Can be expensive.
- Normally need a set of questions.
- Respondent bias – tendency to please or impress, create false personal image, or end interview quickly.
- Embarrassment possible if personal questions are involved.
- Transcription and analysis can present problems – subjectivity.
- If many interviewers are to be used, then training may be required.

(iii) QUESTIONNAIRE FORMULATION

This method makes use of questionnaires which are distributed to respondents. Questionnaires can have questions that are open ended, close ended or tabular; whereby open ended questions invite free responses, close-ended questions only allow respondents to choose from alternative responses provided, and tabular questions are answered by filling in tables.

Reasons for using questionnaires

Questionnaires are a useful method to investigate: Patterns, frequency, ease and success of use. user needs, expectations, perspectives, priorities and preferences user satisfaction with collections and services shifts in user attitudes and opinions relevance of collections and services to user needs Trends (by repetition over time).

Advantages of questionnaires

The main advantages of questionnaires are:

- They are relatively easy to analyse.
- A large sample of the given population can be contacted at relatively low cost;
- They are simple to administer;
- The format is familiar to most respondents;
- Information is collected in a standardised way.
- They are usually straightforward to analyse.
- They can be used for sensitive topics which users may feel uncomfortable speaking to an interviewer about.
- Respondents have time to think about their answers; they are not usually required to reply immediately.

Disadvantages of questionnaires

The main disadvantages of questionnaires are:

- If you forget to ask a question, you cannot usually go back to respondents, especially if they are anonymous.
- It is sometimes difficult to obtain a sufficient number of responses, especially from postal questionnaires.
- Those who have an interest in the subject may be more likely to respond, skewing the sample.
- Respondents may ignore certain questions.
- Questionnaires may appear impersonal.
- Questions may be incorrectly completed.
- They are not suitable to investigate long, complex issues.
- Respondents may misunderstand questions because of poor design and ambiguous language.
- Questionnaires are unsuitable for some kinds of respondents, e.g. visually impaired persons.
- There is the danger of questionnaire fatigue if surveys are carried out too frequently.
- They may require follow up research to investigate issues in greater depth and identify ways to solve problems highlighted.

(iv) CASE-STUDIES

The term case-study usually refers to a fairly intensive examination of a single unit such as a person, a small group of people, or a single company. Case-studies involve measuring what is there and how it got there. In this sense, it is historical. It can enable the researcher to explore, unravel and understand problems, issues and relationships. It cannot, however, allow the researcher to generalise, that is, to argue that from one case-study the results, findings or theory developed apply to other similar case-studies. The case looked at may be unique and, therefore not representative of other instances. It is, of course, possible to look at several case-studies to represent certain features of management that we are interested in studying. The case-study approach is often done to make practical improvements. Contributions to general knowledge are incidental.

The case-study method has four steps:

- Determine the present situation.
- Gather background information about the past and key variables.
- Test hypotheses. The background information collected will have been analysed for possible hypotheses. In this step, specific evidence about each hypothesis can be gathered. This step aims to eliminate possibilities which conflict with the evidence collected and to gain confidence for the important hypotheses. The culmination of this step might be the development of an experimental design to test out more rigorously the hypotheses developed, or it might be to take action to remedy the problem.
- Take remedial action. The aim is to check that the hypotheses tested actually work out in practice. Some action, correction or improvement is made and a re-check carried out on the situation to see what effect the change has brought about.

The case-study enables rich information to be gathered from which potentially useful hypotheses can be generated. It can be a time-consuming process. It is also inefficient in researching situations which are already well structured and where the important variables have been identified. They lack utility when attempting to reach rigorous conclusions or determining precise relationships between variables.

(v) DIARIES

A diary is a way of gathering information about the way individuals spend their time on professional activities. They are not about records of engagements or personal journals of thought! Diaries can record either quantitative or qualitative data, and in management research can provide information about work patterns and activities.

Advantages:

- Useful for collecting information from employees.
- Different writers compared and contrasted simultaneously.
- Allows the person collecting data freedom to move from one organisation to another.
- The person collecting data is not personally involved.
- Diaries can be used as a preliminary or basis for intensive interviewing.
- Used as an alternative to direct observation or where resources are limited.

Disadvantages:

- Subjects need to be clear about what they are being asked to do, why and what you plan to do with the data.
- Diarists need to be of a certain educational level.
- Some structure is necessary to give the diarist focus, for example, a list of headings.
- Encouragement and reassurance are needed as completing a diary is time-consuming and can be irritating after a while.
- Progress needs checking from time-to-time.
- Confidentiality is required as content may be critical.
- Analyses problems, so you need to consider how responses will be coded before the subjects start filling in diaries.

(vi) CRITICAL INCIDENTS

The critical incident technique is an attempt to identify the more 'noteworthy' aspects of job behaviour and is based on the assumption that jobs are composed of critical and non-critical tasks. For example, a critical task might be defined as one that makes the difference between success and failure in carrying out important parts of the job. The idea is to collect reports about what people that do is particularly effective in contributing to good performance. The incidents are scaled in order of difficulty, frequency and importance to the job as a whole. The technique scores over the use of diaries as it is centred on specific happenings and on what is judged as effective behaviour. However, it is laborious and does not lend itself to objective quantification.

(vii) PORTFOLIOS

A measure of a manager's ability may be expressed in terms of the number and duration of 'issues' or problems being tackled at any one time. The compilation of problem portfolios is recording information about how each problem arose, methods used to solve it, difficulties encountered, etc. This analysis also raises questions about the person's use of time. What proportion of time is occupied in checking; in handling problems given by others; on self-generated problems; on 'top-priority' problems; on minor issues, etc? The main problem with this method and the use of diaries is getting people to agree to record everything in sufficient detail for you to analyse. It is very time-consuming!

SOURCES OF DATA:

Sources of data are categorized into two types; namely

a. Primary data sources

Primary data is originated and collected specifically for the problem under investigation. In research, primary data is the major source and is directly collected from the study organisation using the above named tools.

b. Secondary data sources

Secondary data usually compliments primary data. This type of data is collected from other sources other than the primary source. The secondary data material sources include newspapers, other research materials, the internet and publications among others.

2.0 SAMPLING

2.1: DEFINITION

Sampling is the process of selecting a sufficient number of elements from the population so that by studying the sample and understanding the properties or characteristics of the sample subjects, it would be possible to generate the properties or characteristics of the population elements.

2.2: OTHER KEY DEFINITIONS

(i) SAMPLE

It is the subset of a population. It comprises some members selected from the population e.g. if 200 members are drawn from a population of 1000, these 200 form the sample of the study. By studying the sample, one would be able to draw conclusions that are general to the entire population.

(ii) POPULATION

This refers to the entire group of people, events or things of interest that one wishes to investigate.

(iii) AN ELEMENT

An element is a single member of the population e.g. if 500 pieces of machinery are to be approved after inspection, we would say there are 500 elements in the machinery population.

(iv) CENSUS

This is a count of all elements in the population.

(v) POPULATION FRAME / SURVEY POPULATION

It is a listing of all elements in a population from which the sample is drawn. E.g. the payroll of an organisation would serve as a population frame if its members are to be studied. Since the population frame is a list, it has to be updated regularly.

(vi) SUBJECT

It's a single member of a sample just like an element is a single member of the population.

(vii) PARAMETERS

Parameters of the population are the characteristics that are general to the entire population and these include; the population mean, μ, the population standard deviation ,σ, and the population variance, σ^2 ,

2.3 REASONS FOR SAMPLING

(i) Resources in particular time and money are often not adequate to cover the entire population. Sampling saves time and money.

(ii) Sampling is also justified in situations where the results of the inquiry are required immediately. Time saved enables one to;
- Carry out more checks and counter checks in the interest of accuracy.
- Collect more elaborate information
- Give more time to analysis and interpretation of results.

(iii) Sometimes it's not possible to inquire from the entire population e.g. when data collected is in regard to safety of a product.

(iv) Sampling achieves better response rates than an attempt at complete coverage.
(v) The use of sampling permits employment of highly qualified personnel which translates into greater degree of accuracy.
(vi) A well designed scientifically drawn sample coupled with a well conducted survey will produce results significantly close to the population characteristics with no statistical significant difference.

2.4 SAMPLING METHODS /TECHNIQUES /DESIGNS

Considerations when selecting a method of sampling

There are two overriding considerations when selecting a method for selecting a sample.

- Avoiding bias in the selection of a sample.
- Achieving maximum precision in the results for a given cost outlay.

Bias is introduced in a sample under the following circumstances;

- If a non random method is used in selecting methods of the sample.
- If the sampling frame used is incomplete or inadequate.
- If members selected into the sample decline to be interviewed or cannot be traced.

There are 2 broad types of sampling designs or methods:

a) Probability sampling
b) Non-probability sampling

2.4.1 Probability sampling.

This is used when the elements in the population have some known chance or probability of being selected as sample subjects. Probability sampling can either be unrestricted (simple random sampling) or restricted (complex probability sampling) in nature. The choice of these depends on the nature of research problem, the availability of good sampling frame, money, time, desired level of accuracy in the sample and data collection methods. Each has its merits and demerits.

Probability sampling techniques therefore includes;

(i) Simple random sampling techniques (Unrestricted)

Under this technique, each person has some chance as any other of being selected into the sample and forms a standard against which other methods are evaluated. This technique is suitable where the population is relatively small and where the sampling frame is complete and up-to date. It may be done with or without replacement whereby:

- Random sampling with replacement; Is a situation where population members once selected into the sample are put back into the pool such that there is chance that they may be re-selected.

- Random sampling without replacement; Is a situation whereby once members have been selected they cannot be selected again.

There are two basic procedures that are used to select a random sample.

Procedure of sampling

- Obtain a complete sampling frame.
- Assign different numbers to members of the population ranging from **1 – N** then put these members in a container, mix them thoroughly and pick **n** items for the sample; where **N** is the population size and **n** is the sample size.

Point to note

Unrestricted (Simple Random) sampling has the least bias and offers the maximum accuracy of choosing members of the sample from the population.

However this sampling technique may be cumbersome and expensive. Besides, the entity on which sampling is being carried out must have an updated listing of its population.

(ii) Restricted (Complex) Probability Sampling

This offers various and sometimes more efficient alternatives to the unrestricted design. These techniques include;

a) Systematic random sampling techniques
b) Stratified random sampling
c) Random route sampling techniques
d) Multi-stage cluster sampling

(a) Systematic random sampling techniques

This is similar to simple random sampling, but instead of selecting random numbers from tables, you move through a list (sample frame) picking every n^{th} name. To choose the n^{th} name, one must work out the appropriate sampling fraction, by dividing the population size with the required sample size.

Example 1

Consider a population of 600 employees and a required sample of 120, the sampling fraction will be given as;

$$R = \frac{120}{600} = \frac{1}{5}$$

This implies that the person choosing the sample from the population must build up his or her sample size by selecting one person out of every five in the population. The sample fraction, **R**, is important in this case since it helps in determining the starting point when choosing a sample out of the population.

Disadvantage

The effect of periodicity may eventually lead to bias when selecting the sample size.

(b) Stratified random sampling

In this method, all people in the sampling frame are divided into groups or categories known as "*strata*". Within each stratum, a simple random sample or systematic sample is selected.

Example;

Consider a group of 50 employees comprising of 22 males and 28 females. Assuming a sample of 5 employees is required from the sample, compute the number of males and females that will be included in the sample ensuring an equitable selection of each.

Solution

Using stratified sampling,

$$\text{Number of males in the sample} = \frac{5}{50} \times 22 \ = 2.2 \text{ employees}$$

$$\text{Number of males in the sample} = \frac{5}{50} \times 28 \ = 2.8 \text{ employees}$$

Since we can't interview 0.2 and 0.8 of a person, we shall round off one figure upwards and the other downwards;

Thus;
Number of males in the sample = ***2 employees***.
Number of females in the sample = ***3 employees***

The **2** and **3** will then be selected by simple random sampling.

(c) Random route sampling techniques

This is a technique used in market research surveys used for sampling households, shops and other premises in rural and urban areas. In this method, an address is selected at random from a sampling frame usually an electoral register as a starting point. Then given instructions, the person collecting data identifies more addresses by taking alternate left and right turns at road junctions and calling at every n^{th} address (i.e. shop or premise)

Advantages;

(i) May be time saving
(ii) Bias is reduced since the person collecting data has to call at clearly defined addresses.

Disadvantage
(i) Characteristics of particular areas (e.g. rich / poor) may mean that the sample is not representative.
(ii) Open to abuse by the person collecting data because it is difficult to check that instructions have been fully carried out.

(d) Multi-stage cluster sampling

This involves drawing several samples known as sample areas. The sample areas are progressively reduced into smaller areas by choosing smaller sample areas from the larger sample areas. Eventually smaller areas end up into sample households and by using an appropriate method such as systematic or simple random sampling, individuals are selected from the households.

2.4.2 Non- Probability sampling

These are techniques applied when it becomes impossible to undertake a probability method of sampling; this may be due to reasons such as;

- Unavailability of a complete sampling frame for a certain population.
- The probability sampling method adopted has failed to produce good response rates thereby failing to give a good representation of the population being examined.

Non probability sampling methods include;

a. Purposive sampling
b. Quota sampling
c. Convenience sampling
d. Snow ball sampling
e. Self-selection
f. These are described as below;

a. Purposive sampling

A purposive sample is one which is selected subjectively.
The person collecting data in this case obtains a sample that appears to be representative of the population and will usually try to ensure that a range from one extreme to the other is included.

b. Quota sampling

This is often used in market survey research where the person collecting data is required to find cases with particular characteristics. The person collecting data is given a quota of particular types of people to select from and the quota is organised so that the final sample should be representative of the population.

c. Convenience sampling

In this method, the sample comprises subjects who are simply available in a convenient way to the person collecting data. There is no randomness and the likelihood of bias is high. One cannot draw meaningful conclusions from the results you obtain. However, this method is often the only feasible one, particularly in situations where time and resources are constrained.

d. Snow ball sampling

With this approach, you initially contact a few potential respondents and then ask whether they know of anybody with the same characteristics that you are looking for in your study.

e. Self-selection

In this approach, respondents themselves decide that they would like to take part in your study.

Advantages of non-probability methods

(i) They are cheaper.
(ii) Used when sampling frame is not available.
(iii) Useful when the population is widely dispersed that cluster sampling would not be efficient.
(iv) Often used in explorer studies e.g. for hypothesis generation.

2.5 THE ROLE OF SAMPLING IN AUDITING

The main role of an external auditor is to ensure that the accounts of a company are fair and comply with the law. In the early development of auditing it was not unusual for the auditor to make full examination of all of the financial dealings that many of the larger companies are involved with. Usually, this problem is overcome by checking only small portions of the accounts and then using this information to make a generalized statement concerning the entire population of accounts. This, of course, is *sampling*.

Unlike other areas where sampling has been employed, the use of judgment sampling is widespread in accountancy for reasons ranging from little research being carried out in the field of finance and lack of statistical / poor statistical background of the older accountants.

Statistical sampling on top of other advantages enables an optimum sample size to be calculated from the precision required, or, alternatively, for a given sample size it enables the degree of accuracy to be known.
This of course is important when an accountancy firm may be liable to prosecution if it is found to be negligent or if it carried out an inadequate audit. If the firm shows that its audit was carried out with a method which has a recognized scientific basis it is on a safer ground than if the method was one that was purely subjective, no matter how experienced the auditor.

Applications of statistical sampling methods of auditing include;

(a) Estimation sampling of variables;- This involves estimating the total value of a population of accounts based on a sample of items using the formula;

$$Estimate\ of\ value\ of\ population = \frac{Population\ size}{sample\ size} \times value\ of\ sample$$

Typical examples of Estimation sampling of variables in finance include estimation of;

- The value of stock
- The value of assets
- The value of errors
- The value of liabilities

(b) Estimation sampling of attributes;- This involves estimation of the number or proportion, of the population having a certain quality (or attribute). In auditing, its usual application is in the estimation of the number of errors within a set of accounts. However other applications may include;

- Estimating the proportion of debts that are currently six months overdue.
- Estimating the number of employees receiving overtime payments in a certain period.

Such estimates can be computed as;

$$Estimate\ of\ number\ in\ \ population\ with\ attribute$$

$$= \frac{Population\ size}{sample\ size} \times value\ of\ sample$$

(c) Discovery sampling;- This involves the examination of a population for a single case of serious error such as misappropriation of money The type of situation that come under such heading include;

- Misappropriation of money
- Fraud
- Breakdown of internal controls

If such an error is found sampling is immediately stopped and the error reported.

(d) Acceptance sampling

An internal auditor is employed by a company to monitor its accounting system and to check that it is functioning as it should. To aid him in this role he may take samples of items at regular intervals to check that the error rate does not rise above a specified level. This is acceptance sampling and is merely an application of quality control to accounting. In acceptance sampling, we are only concerned with errors that are not as serious as fraud or deception, from example;

- Wrong date
- Incorrect additions
- Not rubber-stamped
- Not signed by recommended person

In general, acceptance sampling is not of use to the external auditor who is only concerned with the taking of one sample.

(e) Monetary unit sampling

This is a recent method of sampling which is useful for estimating the value of error in the population from a sample of items. By using the result;

$$Estimate\ of\ total\ value\ of\ errors\ in\ population$$

$$= \frac{Total\ value\ of\ population}{Total\ value\ of\ sample} \times total\ value\ of\ errors\ in\ sample$$

It can be noted that a more accurate estimate of the total error value is obtained if the sample value are large rather than small. We could obtain this by stratifying according to some criterion of value and then selecting a high proportion of large sample values. Detailed computations and illustrations have been integrated in later topics within this text.

EXERCISE 1

1.1 *Differentiate between*

(i) Sample and Population.
(ii) Element and Subject.
(iii) Population frame and Population parameter.

1.2

(i) Outline any four methods that can be employed in the collection of primary data.
(ii) Explain any three demerits and merits of each technique.

1.3

(i) Differentiate between primary and secondary data.
(ii) Outline four sources of each type of data

1.4

(i) Give two considerations that one must look at before selecting a method of sampling.
(ii) Why should one prefer to employ sampling than make an attempt on the entire population while collecting data?

1.5

(a) Define probability random sampling.
(b) Differentiate between;
- (i) Stratified and systematic random sampling
- (ii) Purposive and convenience sampling
- (iii) Unrestricted and restricted sampling techniques
- (iv) Explain four advantages of probability sampling techniques.

1.6

Briefly describe the following ways of selecting a sample;
- (i) Quota
- (ii) Block (Cluster)
- (iii) Area sampling methods.

3.0 CLASSIFICATION AND PRESENTATION OF DATA

3.1: CLASSIFICATION OF DATA

Classification of data refers to the process of grouping data (usually raw data) into classes of common characteristics and attributes. Therefore different classes created will comprise of data that has similar attributes and characteristics which are of interest to the investigator or researcher.

3.1.1: IMPORTANCE OF DATA CLASSIFICATION

Data classification is of importance as it helps in meeting the following objectives;

(i) *Rearranging of data*; When faced with a mass of data, data classification would help in reducing such data to some kind of order that is manageable.

(ii) *Data processing*; Data classification is one of the first steps towards data processing which is a process of converting raw data into a more meaningful form.

(iii) *Data features and parameters*; Data classification helps application of statistical models which are employed to bring out the different parameters (or features) of data.

3.1.2: Qualitative classification of data

Data that cannot be measured quantitatively is classified according to its attributes such as colour, sex and age. Such data is usually descriptive in nature and only its presence or absence can be noticed. Qualitative classification of data can be simple (dichotomy) or manifold.

3.1.2: Quantitative classification of data

Data that can be measured mathematically or statistically can be classified quantitatively using *class-intervals*, *tables* or *graphs*. Such data can be assigned statistical units that are measurable.

The chapters that follow bring out the quantitative presentation of such data;

3.2: PRESENTATION OF DATA

3.2.1 RAW DATA

When we are faced with a mass of data to be analysed, the first step is to reduce the data to some kind of order. As an example, suppose students sit a mathematics examination which is marked out of 100 marks. Table 3.1 below is a presentation of the marks awarded according to the alphabetical order of the names.

TABLE 3.1 Marks obtained by 100 students in a mathematics examination

34	68	47	68	55	35	52	34	64	55
59	20	52	44	76	30	58	55	41	74
58	46	58	18	57	41	28	61	36	53
41	68	41	53	43	61	45	64	37	64
32	52	81	48	53	15	59	37	53	57
58	40	52	46	30	55	25	58	53	60
11	60	31	57	42	69	50	55	39	57
50	45	66	49	39	59	44	60	48	64
78	45	23	50	45	58	29	53	34	72
38	67	49	65	38	69	38	57	63	57

Table 3.1 above is an example of "raw data", i.e. data which has not been organised into some manageable form. The first step in dealing with such data is to group them into suitable classes. The normal procedure is to find the range of the data.

The ***range*** is the difference between the highest and lowest scores in the data. From the table, the highest score is 81 and the lowest is 11. Hence

Range = 81 – 11 = **70**

From the range we can see that it will be convenient to choose about 7 classes starting from 10, and using a class of 10 marks. We choose the groups, 10-19, 20-29, ….., 80-89. Thus ending up with 8 groups of the single mark 81 which falls into its own group and cannot be included in the 70-79 group.

The factors influencing the choice of the grouping interval are two:

- Their size should be such as to permit a convenient number of groups; not too small, and not too long, i.e. 5 to 12 groups.
- It should promote tabulating convenience.

3.2.2 The Frequency Distribution

Grouping data into suitable classes is a way of constructing a *Frequency Distribution*. This is done as follows. Proceeding down each column in Table 3.1 we put a small stroke across each score and then put an equivalent stroke in the column labeled *Tally* in Table 3.2 as the counting progresses through the raw data. Starting with the score 34, we put a small stroke against the class 30-39; the second score is 59, therefore we put a small stroke against the class 50-59; and so on until we go through all the data. Every fifth stroke is drawn across the preceding four strokes to form small 'bundles' of five. This enables one to tell the total number of scores in each class more easily. The total number of scores in each class is entered in the column labeled *Frequency*, as shown in Table 3.2.

TABLE 3.2 Frequency distribution for the marks of 100 students

Class	Tally	Frequency
10 - 19	𝍸	3
20 - 29	𝍸	5
30 - 39	𝍸 𝍸 𝍸	16
40 - 49	𝍸 𝍸 𝍸 𝍸	20
50 - 59	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	33
60 - 69	𝍸 𝍸 𝍸	18
70 - 79	////	4
80 - 89	/	1

The following points should be considered when forming frequency distributions:

- First determine the largest and smallest numbers in the raw data and hence find the range
- Divide the range into a convenient number of class intervals having the same classes.
- Avoid forming a frequency distribution in which there are too many empty classes.
- Avoid forming classes where much information is lost at the lower end or upper end of the frequency distribution.

3.2.3 Class Limits

A study of Table 3.2 shows that the class 10-19 includes only those marks which fall in the interval 10 to 19 inclusive. Similarly the class 30-39 includes only those marks which fall in the interval 30-39 inclusive, and so on. We say that 10 is the *Lower class limit* of the class 10-19, while 19 is the *Upper class limit* of the class 10-19, while 39 is the *Upper class limit* of the class 30-39.

3.2.4 Class Boundaries

The raw data in Table 3.1 suggests that the marks were recorded to the nearest whole mark, which means that the marks in the class 10-19 for example, should be only those marks which are greater than $9\frac{1}{2}$ or 9.5 and less than $19\frac{1}{2}$ or 19.5. We say that 9.5 is the *Lower class boundary* of the class 10-19, while 19.5 is its *Upper class boundary*. Hence class boundaries should be carried to one more decimal place than the recorded observations.

3.2.5 Class Width

The *Class width* is the difference between the *lower* and *upper* class boundaries. We may also define class width as *Class size*. Thus in our example the class width (or class size) is

19.5 – 9.5 = **10**

3.2.6 Class Mark

The *Class mark* is the *midpoint* of a class interval and is obtained by adding the *lower* and *Upper* class boundaries and dividing by two. Thus the class mark of the interval 50-59 is

$$\frac{49.5 + 59.5}{2} = \frac{109}{2} = \mathbf{54.5}$$

Table 3.3 below is a frequency distribution which includes the class boundaries and the class marks but excludes the "Tally" column. This is the normal way of arranging the frequency distribution table for calculations.

TABLE 3.3 Frequency distribution table for 100 students

Class	Class boundary	Class mark	Frequency
10-19	9.5 - 19.5	14.5	3
20-29	19.5 - 29.5	24.5	5
30-39	29.5 - 39.5	34.5	16
40-49	39.5 - 49.5	44.5	20
50-59	49.5 - 59.5	54.5	33
60-69	59.5 - 69.5	64.5	18
70-79	69.5 - 79.5	74.5	4
80-89	79.5 - 89.5	84.5	1
			100

In Table 3.3 we have included the number 100 at the end of the column of frequency. As you will notice this number is the total obtained when all the frequencies are added up. This number is exactly equal to the total number of students who did the examination. This serves as a check on whether you may have made an error in tallying the data. The sum of the frequencies must be exactly equal to the total number of observations. We shall talk more about this in Chapter 4.

3.2.7 Histograms

In order to illustrate the information contained in a frequency distribution in a clearer way it is instructive to use some kind of diagram, since a pictorial representation usually makes a clearer impression on the mind than columns of figures. A *histogram* is one such representation. Fig.3.1 shows a histogram drawn using Table 3.3.
On the horizontal axis we plot the midpoint of a class while on the vertical axis we plot its respective frequency. A quick study of the histogram above shows that the *area* of each column is proportional to the frequency of the corresponding class. This is because in our example, the classes have equal width. It therefore follows that the *heights* are also proportional to the corresponding frequencies

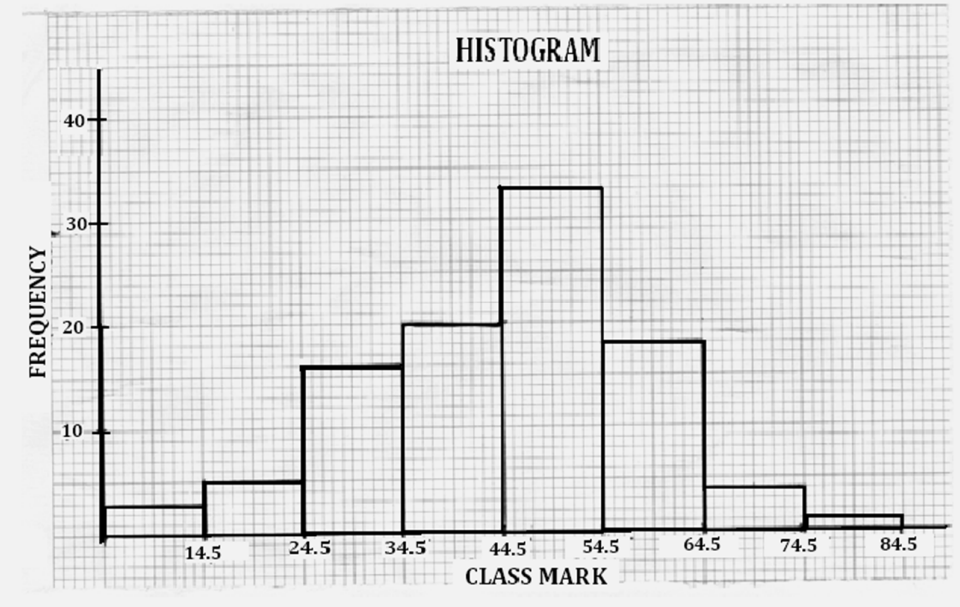


Fig.3.1 Histogram

3.2.8 Frequency Polygon

If we plot the frequency against the midpoint of the corresponding class, and then join the points with straight lines we obtain a *Frequency Polygon*, Fig.3.2.

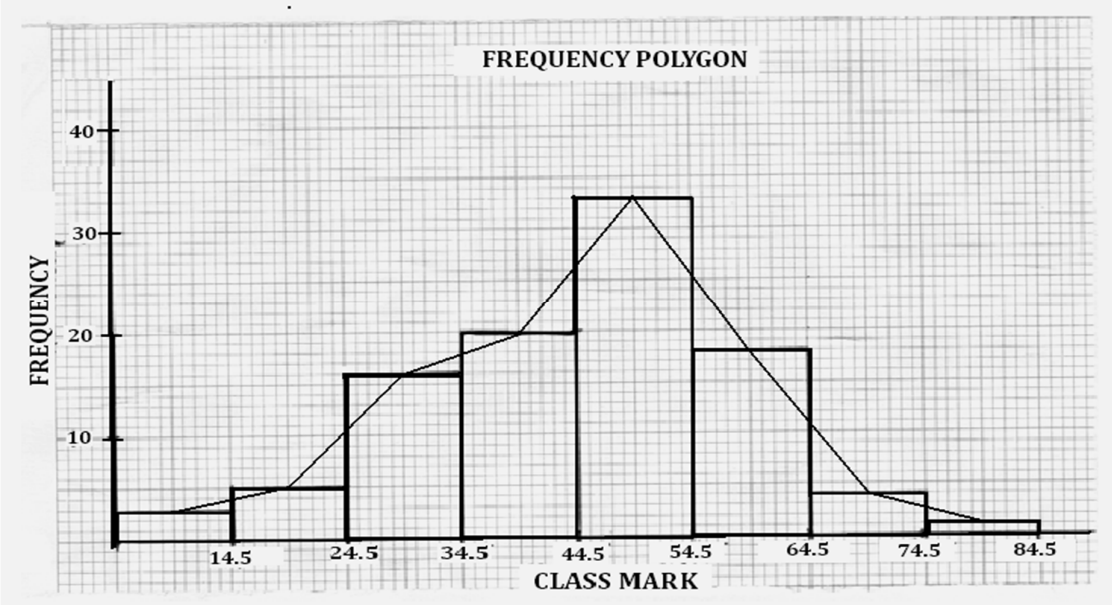


Fig.3.2 Frequency Polygon

Figure 3.2 clearly shows the position of the histogram in relation to the frequency polygon.

3.2.9 Cumulative Frequency Distribution

Using a frequency distribution we can construct a *Cumulative Frequency Distribution*. The total frequency of all values less than the upper class boundary of a given class interval is called the *Cumulative Frequency* up to and including that class. A table containing the cumulative frequencies is called a cumulative frequency distribution, and is shown below, Table 3.4.

TABLE 3.4 A cumulative frequency distribution table.

Class	Class boundaries	Class mark	Frequency	Cumulative Frequency
10-19	9.5 - 19.5	14.5	3	3
20-29	19.5 - 29.5	24.5	5	8
30-39	29.5 - 39.5	34.5	16	24
40-49	39.5 - 49.5	44.5	20	44
50-59	49.5 - 59.5	54.5	33	77
60-69	59.5 - 69.5	64.5	18	95
70-79	69.5 - 79.5	74.5	4	99
80-89	79.5 - 89.5	84.5	1	100

3.2.10 Cumulative Frequency Polygon

Using a cumulative frequency table we can plot a *Cumulative Frequency Polygon*. A cumulative frequency polygon is obtained by plotting the Upper class boundary against the cumulative frequency of that class interval and then joining all the consecutive points by straight lines. A cumulative frequency polygon is shown in Fig.3.3. A cumulative frequency polygon is sometimes also called an *Ogive*.

3.2.11 Percentage Ogive (Smoothed Ogive)

Another form of frequency polygon is the percentage Ogive or otherwise called a smoothed Ogive. In this case we first calculate the relative frequencies as shown in Table 3.5 and plot these against the *upper class boundaries*.

TABLE 3.5 Percentage Cumulative Distribution of Marks

Class boundaries	Cumulative Percentage
Less than 19.5	3%
Less than 29.5	8%
Less than 39.5	24%
Less than 49.5	44%
Less than 59.5	77%
Less than 69.5	95%
Less than 79.5	99%
Less than 89.5	100%

Note that the cumulative percentage is identical to the cumulative frequencies in Table 3.4 in this case because the total number of students who sat the examination was exactly 100. Normally this is not the case.

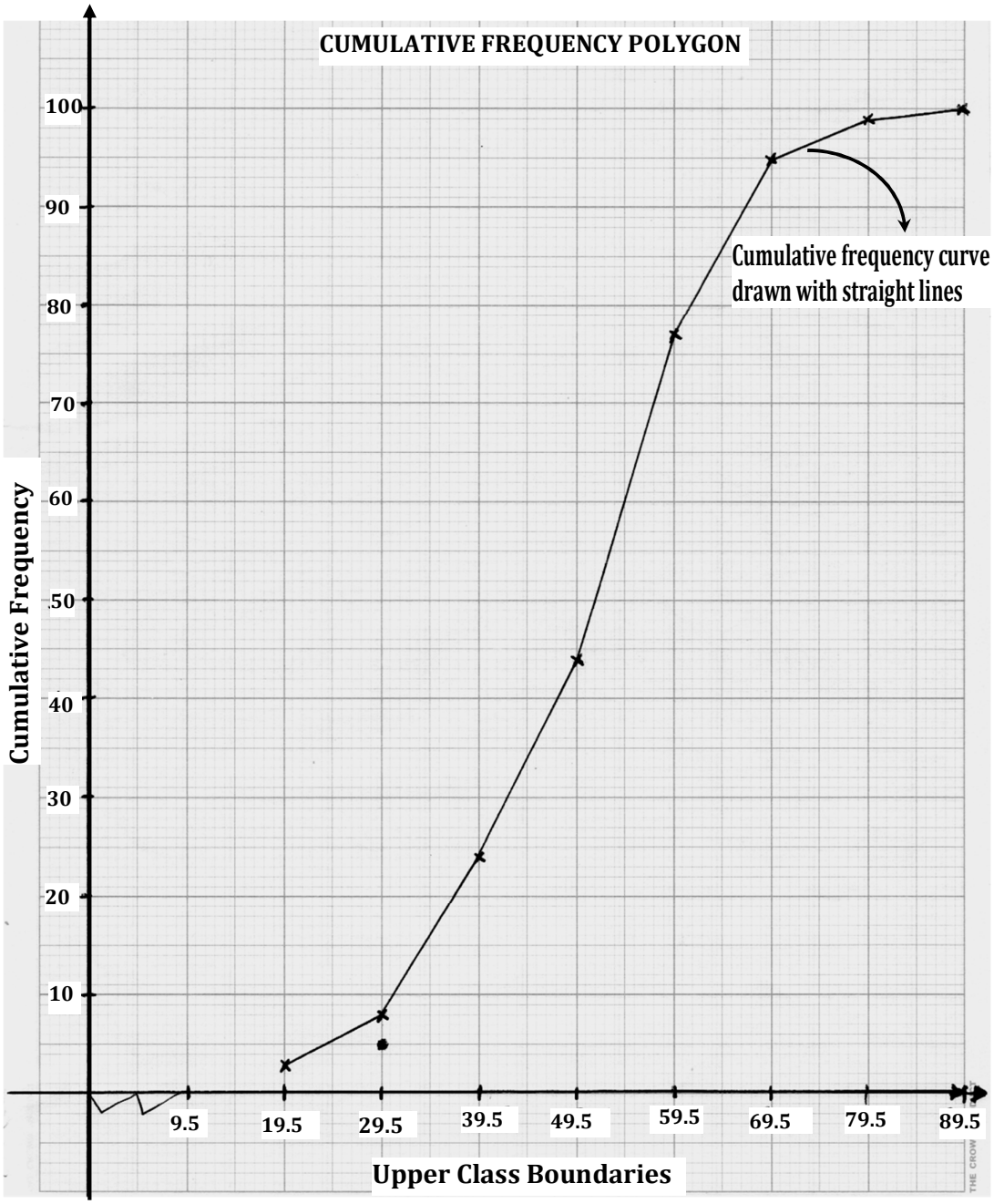


FIG.3.3 A cumulative frequency polygon

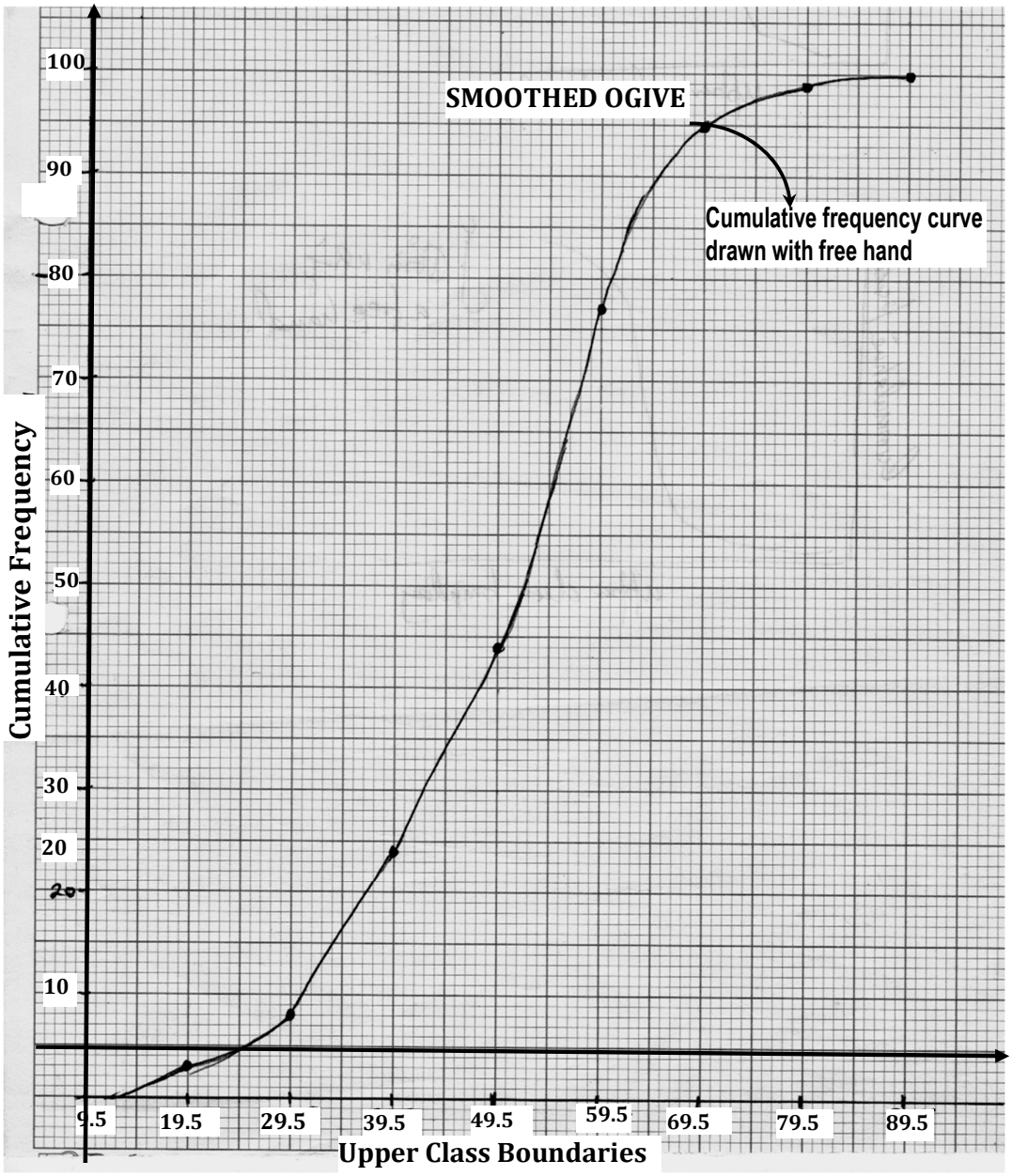


Fig.3.4 The Smoothed Ogive

The advantage of the percentage cumulative distribution is that we are able to read off the percentage of observations falling below certain specified values. This will be illustrated in Chapter 4.

3.2.12 Types of Frequency Curves

Besides the cumulative percentage curve shown in Fig.3.4, that is the smoothed Ogive, we have other types of frequency curves which can occur in statistics.

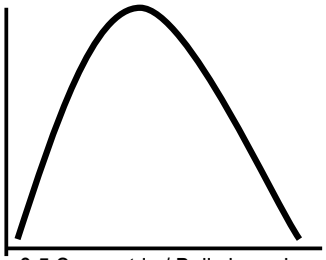


Fig 3.5 Symmetric / Bell shaped
e.g Normal curve

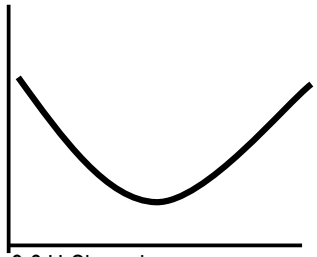


Fig. 3.6 U-Shaped

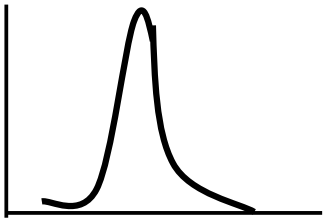


Fig 3.7 Positively Skewed

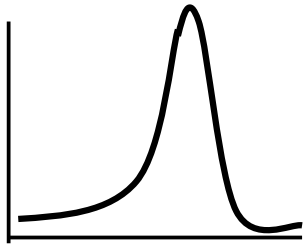


Fig 3.8 Negatively skewed

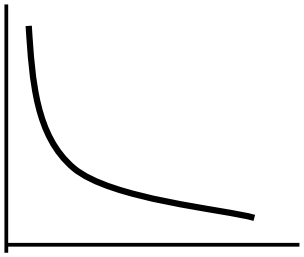


Fig 3.9 Reverse J-shape

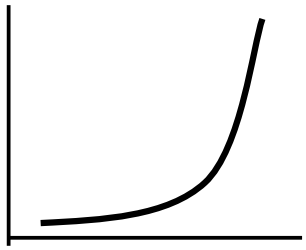


Fig 3.10 J-Shape

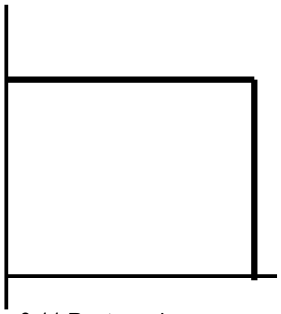


Fig 3.11 Rectangular

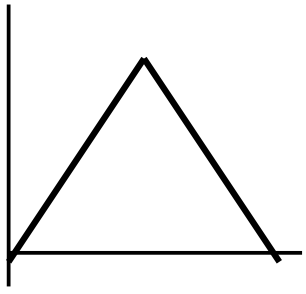


Fig 3.12 Triangular

EXERCISE 2

2.1

(a) Define "data classification" as applied in Quantitative techniques.
(b) Explain 3 main objectives for classifying data.
(c) Differentiate between Qualitative and Quantitative data.
(d) What are the 3 main ways of classifying Qualitative data?

2.2 The following are the highest marks scored in 45 different examinations.

74	76	82	76	59	67	73	80	90
63	48	73	91	66	88	63	74	86
67	71	51	49	96	71	72	85	87
86	71	82	83	66	71	76	48	51
92	90	46	72	81	86	82	76	64

(i) How can such data be classified?
(ii) Make a frequency table starting with less the class 40-49.
(iii) Draw a histogram for the table.

2.3 The table below shows a list of marks obtained by candidates in an Accounting test.

74	77	73	77	69	64	79	60
77	71	70	73	66	51	64	63
45	75	66	62	60	77	82	70
80	78	72	76	81	67	59	73
81	79	73	68	63	52	83	72

(i) Make a frequency distribution table of the marks, grouping them in class intervals of five, starting with the class interval 45-49.
(ii) Draw a histogram for the data.

2.4 The following table shows average temperatures in degrees Celsius for some weather stations for a certain month.

12	16	24	14	16
19	12	24	21	25
21	24	16	16	25
26	29	31	31	30
28	6	4	13	16

(i) Construct a frequency table starting with the class interval 4-8.
(ii) Draw a histogram of the distribution.

2.5 The data below shows the heights in centimetres of pupils in a certain school.

132	125	117	124	108	112	100
130	122	118	114	103	119	106
125	128	106	111	116	132	129
136	92	115	118	121	137	123
119	115	101	129	87	108	110
127	103	110	126	118	82	104
146	126	119	119	105	132	126

(i) From a frequency distribution table for the data having equal interval size starting with the class 80-89.
(ii) Convert the distribution in (i) above to a cumulative distribution.
(iii) Draw an Ogive of the cumulative distribution.

2.6 The table below shows the distribution of marks obtained in an Economics examination.

Marks	30-30	35-39	40-44	45-49	50-54	55-59	60-64	65-69
Frequency	10	13	21	32	29	11	3	1

(i) Construct a smoothed cumulative frequency polygon
(ii) Estimate the number of students who scored more that 48 marks.

2.7 The following table gives the number of deaths occurring in each age group during 1952 for a certain village of population 6200.

Age group	0-5	5-15	15-25	25-35	35-45	45-55	55-65	65-
Deaths	13	1	2	3	3	8	12	20

(i) Construct a histogram of the data
(ii) What type of frequency distribution does the data represent?

2.8 A company which manufactures tubes for television receivers conducted a test of a sample batch of 1000 tubes and recorded the number of faults in each tube in the following table:

No. of faults	0	1	2	3	4	5	6
Frequency	620	260	88	20	8	2	2

Plot the frequency polygon and determine whether it is a reverse-J shape type of distribution.

2.9 The following figures give the estimated diameters of forty-two nylon threads in units of thousandths of an inch.

1.10	0.96	0.86	1.02	1.08	0.92	0.88
1.06	1.04	0.88	0.92	1.10	1.16	1.06
1.10	1.00	1.04	1.04	1.06	0.98	1.06
1.10	1.00	1.06	0.98	1.04	1.06	1.08
1.00	0.92	0.86	1.02	0.90	1.04	0.92
0.88	1.06	0.92	1.02	1.04	1.04	1.16

Form a frequency table from these figures with about six groups of equal width.

2.10 The table below shows the results of throwing a single die 572 times.

Score	1	2	3	4	5	6
Frequency	94	96	96	95	95	96

Plot the frequency polygon and you will find that it is approximately rectangular.

4.0
MEASURES OF LOCATION
(CENTRAL TENDENCY)

Chapter 3 outlined how data can be reduced to more manageable proportions by means of tables, and the important features were portrayed by means of diagrams (mainly graphs). The next logical step now is to examine how the data can be summarised by calculating a few representative values which enable comparisons between one set of data and another to be made. The first quantity commonly employed is the *arithmetic average* or *mean* of the observations. Typically, the mean of a set of data tends to lie centrally within the data, and therefore we call it a "measure of central tendency". Several other measures of central tendency can be defined, for example, the *mode*, the *median*, the *geometric mean*, the *harmonic mean*, and the *quadratic mean*.

4.1 The Mode

The mode of a distribution is the most frequent value of the variable in consideration. The meaning of this statement may be illustrated by a few examples.

UNGROUPED DATA:

Example 4.1

Suppose we are given the following set of numbers

1 2 2 3 5 5 5 6 7 9

Solution

Clearly the most frequent value is 5. Hence the mode of the above data is 5.

Example 4.2

The heights of 50 people are given in the table measured to the nearest one inch.

Height	62	63	64	65	66	67	68
No. of persons	1	5	12	14	10	6	2

Find the mode.

Solution

The mode is 65 inches.

Notice that the "number of persons" is also the frequency of the data.

The above examples are quite easy and the mode can be found easily. However when the data is grouped it may be so easy to tell the mode just by looking at the data. In such cases we must actually calculate the mode.

GROUPED DATA:

For grouped data the mode is given by

$$\boldsymbol{Mode} = \boldsymbol{L_b} + \left(\frac{\boldsymbol{d_b}}{\boldsymbol{d_b} + \boldsymbol{d_a}}\right) \times \boldsymbol{C_w} \qquad (4.0)$$

Where
L_b = Lower boundary of the modal class.
d_b = the difference between the frequencies of the modal class and the class before it.
d_a = the difference between frequencies of the modal class and the class after it.
c_w = the class width.

By definition the *modal class* is the class with the highest frequency.

Example 4.3

Determine the mode of the frequency distribution depicted below.

Class	5-9	10-14	15-19	20-24	25-29
Frequency	2	9	15	20	4

Solution

A good piece of advice in doing calculations in statistics is to arrange the information you are given in a vertical table.

Class	Class boundaries	Frequency f
5-9	4.5 - 9.5	2
10-14	9.5 - 14.5	9
15-19	14.5 - 19.5	15
20-24	19.5 - 24.5	20
25-29	24.5 - 29.5	4

The modal class is **20-24**
L_b = 19.5
$d_b = f_m - f_{m-1} = 20 - 15 = 5$
$d_a = f_m - f_{m+1} = 20 - 4 = 16$
c_w = 5

$$\boldsymbol{Mode} = 19.5 + \left(\frac{5}{5+16}\right) \times 5$$

$$= 19.5 + \left(\frac{5}{21}\right) \times 5$$

$$= 19.5 + 1.2$$

$$= \mathbf{20.7}$$

While the mode provides a very easy and accurate measure for locating distributions which are simple in form and closely similar, it is of little use if the distributions are not in any way similar or if they have two or more maxima. A distribution may be found to have two modes. In this case we call the distribution a bi-modal distribution. More than three modes means that the distribution is multimodal. As a practical measure of position the mode is therefore of limited use, but in suitable cases it is useful because of its simplicity.

4.2 The Median

Definition:
The median is defined as that central value of the variable for which there are equal numbers of frequencies of greater and smaller values on both sides.

If the observations are arranged in order of magnitude, the median is simply the middle observation. If the distribution consists of an even number of observations it is not possible to apply the definition precisely as there are two middle terms. It is usual in such cases to take as the median the value of the variable half-way between the two middle classes.

UNGROUPED DATA

Example 4.4

Determine the median of the numbers given below.

13 3 4 8 9 9

Solution

The data is already arranged in order of magnitude and clearly the median is 4.

Example 4.5

Determine the median of the numbers:
40 44 48 50 51 57 60 63 64 65

Solution

The median $= \dfrac{51 + 57}{2}$

$= \dfrac{108}{2} = \underline{\mathbf{54}}$

GROUPED DATA

As we saw with the mode of grouped data, the median of grouped data is found by calculation.

For grouped data the median is given by:

$$\boldsymbol{Median} = \boldsymbol{L_b} + \left(\frac{\frac{\boldsymbol{n}}{2} - \boldsymbol{cf_b}}{\boldsymbol{f_m}}\right) \times \boldsymbol{C_w} \qquad (4.1)$$

Where L_b = Lower boundary of the median class
m = median class
f_m = frequency of the median class
cf_b = Cumulative frequency of a class before the median class
n = number of observations
c_w = class width

NOTE:

In this case, a column for cumulative frequency must be created on the frequency distribution table.

Example 4.6

Find the median of the distribution of Example 4.3

Solution There are a total of 50 observations. Since by definition the median is the middle value, that is, the 25th observation, the median falls in the class interval 15–19. Hence the Median class is 15 -19.

Thus
L_b = 14.5
m = ??

$f_m = 15$
$cf_b = 11$ $c_w = 5$
$n = 50$

$$\boldsymbol{Median} = 14.5 + \left(\frac{\frac{50}{2} - 11}{15}\right) \times 5$$

$$= 14.5 + \left(\frac{25 - 11}{15}\right) \times 5$$

$$= 14.5 + 4.7$$

$$\boldsymbol{Mode} = \mathbf{19.2}$$

4.3 The SIGMA notation

In Quantitative techniques, we find that it is convenient to use the Greek capital letter Σ, sigma to denote "the sum of all like terms like……". Let the symbol X_i (read "X sub i") denote any of the N values $X_1, X_2, X_3, ..., X_N$ assumed by a variable X. The letter i which can stand for any of the numbers 1, 2, 3, …, N is called a subscript. Combining the symbol Σ with the symbol X_i we write the symbol $\sum_{i=1}^{N} \boldsymbol{X_i}$ to denote the sum of all the X_i's from i=1 to i=N.
In other words we are saying that;

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + X_3 + \ldots. + X_N \qquad (4.2)$$

In Particular

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5 \qquad (4.3)$$

When no confusion can result, we shall often denote this simply by $\Sigma_i X_i$ or
ΣX_i or even Σ X. clearly any letter other than i, such as *k, p, q,* etc. can be used.

Example 4.7

$$\sum_{i=1}^{N} X_i Y_i = X_1 Y_1 + X_2 Y_2 + X_3 Y_3 + \ldots. + X_N Y_N \qquad (4.4)$$

Example 4.8

$$\sum_{i=1}^{N} aX_i = aX_1 + aX_2 + aX_3 + \ldots. + aX_N \qquad (4.5)$$

$= a (X_1 + X_2 + X_3 + \ldots\ldots\ldots\ldots + X_N) = \boldsymbol{a \sum_{i=1}^{N} X_i}$

Example 4.9

$$\sum_{i=1}^{N} a = a + a + a + \ldots. + a = Na \qquad (N\ times) \qquad (4.6)$$

4.4 The Mean

If the values of a variable are represented by $x_1, x_2, x_3, \ldots, x_n$, the arithmetic mean, or average denoted by $\overline{X}$ (read as "$\overline{X}$ bar") is defined by the relation;

$$\overline{X} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{N} x_i}{n} = \frac{\sum x}{n} \qquad (4.7)$$

Example 4.12

Find the mean of the numbers: 4 8 3 5 12 10

Solution

$$\overline{X} = \frac{4 + 8 + 3 + 5 + 12 + 10}{6} = \frac{42}{6} = \mathbf{7.0}$$

If the value x_1 occurs f_1 times, and x_2 occurs f_2 times, and so on, then

$$\overline{X} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \qquad (4.8)$$

$$= \frac{\sum fx}{\sum f} = \frac{\sum fx}{n}$$

Where $n = \sum f$ is the *total frequency*

Example 4.13

If the numbers, 5, 8, 6, and 2 occur with frequencies 3, 2, 4, and 1 respectively, find the mean of the numbers.

Solution

The above statement means that we have the numbers:

5 5 5 8 8 6 6 6 6 6 2

The mean of these numbers is therefore

$$\overline{X} = \frac{3(5) + 2(8) + 4(6) + 1(1)}{3 + 2 + 4 + 1} = \frac{15 + 16 + 24 + 2}{10} = \frac{57}{10} = \mathbf{5.7}$$

4.5 The Weighted Mean

Sometimes the term "frequency" can be given a wider interpretation by calling it the "weighted" mean. In this case we define the *weighted mean* as

$$\overline{X} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_n x_n}{w_1 + w_2 + w_3 \ldots + w_n} = \frac{\sum wx}{\sum w} \qquad (4.9)$$

Example 4.14

A car manufacturer produces 3 models which he sells at \$ 600, \$ 900, and \$ 1200. If he makes 10, 4, and 1 cars of each model respectively, find the average price of the cars.

Solution

The average price = $\frac{10(600) + 4(900) + 1(1200)}{15}$

$$= \frac{6000 + 3600 + 1200}{15} = \frac{10{,}800}{15} = \mathbf{\$720}$$

Example 4.15

A candidate obtains the following percentages in an examination: French 70, Mathematics 80, Commerce 50, English 75, Biology 55, History 45, and Geography 60. It is agreed to give double weight to the marks in English, Mathematics and French. What is his weighted mean?

Solution

$$\text{Mean} = \frac{2(70) + 2(80) + 1(50) + 2(75) + 1(55) + 1(45) + 1(60)}{2 + 2 + 1 + 2 + 1 + 1 + 1}$$

$$= \frac{140 + 160 + 50 + 150 + 55 + 45 + 60}{10}$$

$$= \frac{660}{10} = \mathbf{66}$$

$$\text{The Weighted mean} = \frac{435}{7} = \mathbf{62.1}$$

4.6 The mean for grouped data

When data are represented in a frequency distribution, all values falling in a given class interval are considered to be coincident with the class mark or midpoint of the interval. The mean is then calculated using an assumed mean, and deviations from this mean. The shortest way to calculate the mean is to use the *coding method*. However we shall first calculate the mean using its definition directly (Example 4.16) and then we shall use the *coding method* (Example 4.17) to illustrate how this method makes it easier to handle the arithmetic.

Example 4.16

The "recovery time" of an aircraft which is the time that elapses between its arrival at an airport and its being ready to take off again is given below for several aircrafts.

Recovery time (min)	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44
No. of Aircrafts	3	11	13	21	26	12	7	2

Calculate the mean recovery time.

Solution

By definition, the mean,

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

This may be written in the form

$$\bar{X} = A + \frac{\sum fd}{\sum f} \qquad (4.10)$$

Where; **A** is an assumed mean, and

d are the deviations from this mean.

Let **A** = 27 be the assumed mean.

We put all the information we are given in a vertical table.

Recovery time (minutes)	Class Mark X	No. of aircrafts (Frequency, f)	Deviations,d, from A=27	fd
5-9	7	3	-20	-60
10-14	12	11	-15	-165
15-19	17	13	-10	-130
20-24	22	21	-5	-105
25-29	27	26	0	0
30-34	32	12	5	60
35-39	37	7	10	70
40-44	42	2	15	30
		$\Sigma f = 95$		**$\Sigma fd = -300$**

The Mean = $A + \frac{\Sigma fd}{\Sigma f}$ = $27 + \frac{-300}{95}$ = 27 - 3.16 = 23.84

In the following example, Example 4.17 we do the same problem illustrating what is meant by "coding". Coding in this case means that we factorise out the common multiple from the deviations column. You will have noticed that the numbers under the column headed "Deviations, d, from A = 27" are all multiples of 5. Hence we can factorise out 5 thus greatly simplifying the arithmetic. The new formula to use now becomes.

$$\bar{X} = \boldsymbol{A} + \left(\frac{\sum \boldsymbol{fu}}{\sum \boldsymbol{f}}\right) \times \mathbf{C} \qquad (4.11)$$

Where **c**, the number factorised out from the deviations column, is essentially the class width!, and

u is an integer positive or negative, or zero. That is u = 0, ±1, ±2, ±3, ….

Example 4.17

Use the coding method to find the recovery time for Example 4.16.

Solution

Recovery time (minutes)	Class Mark X	No. of aircrafts (Frequency, f)	Deviations from A=27	u	fu
5-9	7	3	-20	-4	-12
10-14	12	11	-15	-3	-33
15-19	17	13	-10	-2	-26
20-24	22	21	-5	-1	-21
25-29	27	26	0	0	0
30-34	32	12	5	1	12
35-39	37	7	10	2	14
40-44	42	2	15	3	6
		$\sum f = 95$			$\sum fu = -60$

The Mean = $A + \frac{\sum fu}{\sum f}$ = $27 + 5\left(\frac{60}{95}\right)$ = $27 + 5(0.63)$ = **23.85**

4.7 Relation between Mean, Mode, and the Median

For unimodal frequency curves which are moderately skewed (asymmetrical) we have the relation (*see chapter 6 "skewness"*)

Mean – Mode = 3 (Mean – Median) (4.12)

4.8 The Geometric Mean

The geometric mean of **n** is the nth root of their product. Thus the geometric mean of the five numbers 1 2 3 4 5 is

$$= \sqrt[5]{1 \times 2 \times 3 \times 5} = \sqrt[5]{120} = \mathbf{2.605}$$

Note that the geometric mean is less than the arithmetic mean. It is not influenced by abnormal individuals to the same extent. It is a useful measure for averaging percentage relatives.

4.9 The Harmonic Mean

The *Harmonic mean* is defined in terms of its reciprocal. The *reciprocal* of the *harmonic mean* of *n* numbers is the *arithmetic mean* of their *reciprocals*. Thus the harmonic mean of the numbers 1 2 3 4 5 is given by;

$$\frac{1}{harmonic\ mean} = \frac{1}{5}\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}\right)$$

$$= \frac{1}{5}(1 + 0.5 + 0.3333 + 0.25 + 0.2) = \underline{\mathbf{0.4567}}$$

Hence the harmonic mean =**2.19**

We note that the harmonic mean is less than the geometric mean and that it might be alternately defined as the *reciprocal of the arithmetic mean of the reciprocals*.

EXERCISE 3

3.1 Find the mode, median and arithmetic mean of the distribution of the marks in a general knowledge test.

Mark	1	2	3	4	5	6	7	8	9	10
Frequency	2	7	17	29	38	41	40	30	17	6

3.2 The weight of a crop of shallots to the nearest 0.1 is given in the table below.

Weight	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Frequency	52	81	63	54	30	22	14	10	6	3	1	2	1	2

Find the mode, median, and mean of the distribution.

3.3 The following is the frequency distribution of the weights of 100 men:

Weight(lb)	Frequency	Wt. (lb)	Frequency
100-109	1	160-169	17
110-119	2	170-179	10
120-129	5	180-189	6
130-139	11	190-199	4
140-149	21	200-209	2
150-159	20	210-219	1

Calculate for the distribution

(i) the mode
(ii) the median
(iii) the mean

3.4 In estimating the value of a plantation of fir trees, the girths of the trees in a sample area of 500 trees were measured and the results tabulated as follows:

Girth (in.)	15-24	25-34	35-44	45-54	55-64	65-74	75-84
No. of trees	25	30	135	160	100	40	10

Calculate for the distribution

(i) the mode
(ii) the median
(iii) the mean

3.5 The annual salaries of four men were \$5000, \$6000, \$7000, and \$10,000. Find the arithmetic mean of their salaries.

3.6 A student's final grades in Mathematics, Physics, English and Biology are respectively 82, 86, 90, and 70. If the respective credits received for these courses are 3, 5, 3, and 1. Determine an appropriate average grade.

3.7 In a company with 80 employees, 60 earn a mean salary of \$3.00 per hour and 20 earn a mean salary of \$2.00 per hour. What is the mean salary for all 80 employees?

3.8: A set of numbers consists of six 6's, seven 7's, eight 8's, nine 9's, and ten 10's. What is the arithmetic mean of the numbers?

5.0
MEASURES OF DISPERSION

Besides the arithmetic mean which measures the central location of data, another characteristic of a distribution is the extent of spread or variation of the variables about the central value such as the median or the mean. The degree of scatter or variation is termed *dispersion*. As with measures of central tendency there are several measures of dispersion available, the most common being the *range, the semi-interquartile range, the 10-90 percentile range, the mean deviation from the median, and the standard deviation*.

5.1 The Range

The range has already been defined as the difference between the extreme values of the variable, and therefore is the simplest measure of dispersion. The major disadvantage of using the range is that it depends only on the extreme values of the variable and therefore may fail to discriminate between different types of distributions which happen to have the same end values.

5.2 Quartiles

Quartiles divide a distribution into four equal parts. The quartiles are denoted by Q_1, Q_2 and Q_3, where Q_1 is the *lower quartile* and Q_3 is the *upper quartile*. The mid-quartile Q_2 coincides with the median of the distribution. The *interquartile* range is defined by $Q_3 - Q_1$. Half the interquartile range is given by;

$$\frac{Q_3 - Q_1}{2} \tag{5.1}$$

which is called the semi-interquartile range. The semi-interquartile range is a useful measure of dispersion because it is not upset by any observations that could not be called representative values. Quartiles may be obtained both by graphical methods and of course by calculation. A few examples will best illustrate the meaning of the quantities which have been introduced above.

Example 5.1

The distribution of marks in percentage obtained in an examination was as follows:

Class	20-24	25-29	30-34	35-39	40-44	45-49	50-54
Frequency	9	17	21	18	15	9	11

Determine *(i) the upper quartile, Q_3*
(ii) the lower quartile, Q_1
(iii) the semi-interquartile range

Solution:

GRAPHICAL METHOD

This method computes the lower and the upper quartiles by use of a smoothed cumulative frequency polygon(the Ogive).

As always, arrange the information you are given in a vertical table, and in your table include a column for cumulative frequency. See Table 5.1

TABLE 5.1 Frequency distribution of marks in an examination paper.

Class	Class boundaries	Frequency	C.F
20-24	19.5-24.5	9	9
25-29	24.5-29.5	17	26
30-34	29.5-34.5	21	47
35-39	34.5-39.5	18	65
40-44	39.5-44.5	15	80
45-49	44.5-49.5	9	89
50-54	49.5-54.5	11	100

Fig.5.1 is a smoothed cumulative frequency polygon for the above data. Note that in order to deduce the quantities graphically it is best to use smoothed cumulative frequency polygon.

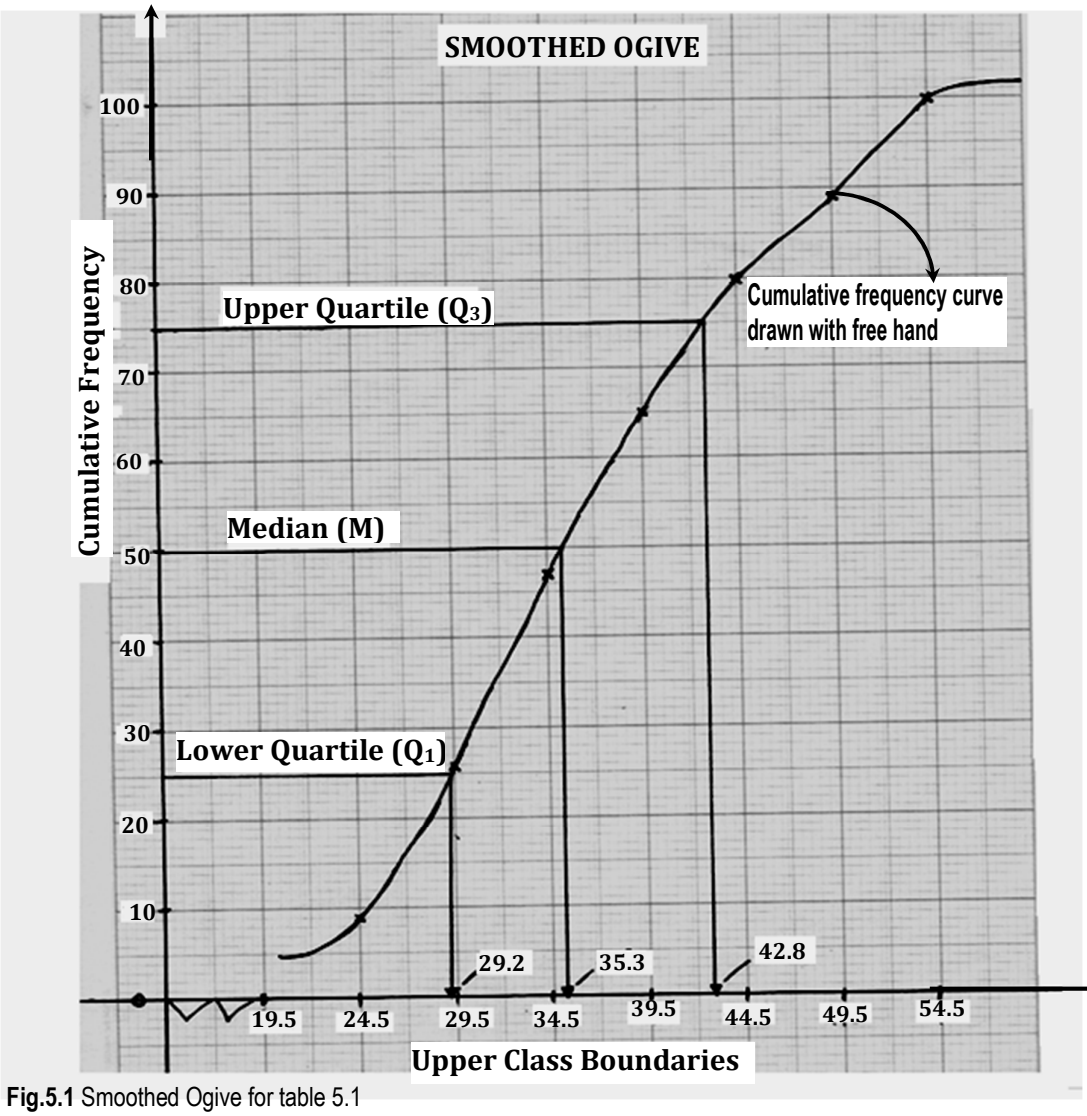


Fig.5.1 Smoothed Ogive for table 5.1

Q_3 = 42.5%, Q_1 = **29.2%,** $Q_3 - Q_3$ = 42.8 – 29.5 = **13.3%**

Hence the semi-quartile range = ½ ($Q_3 - Q_1$) = ½ x 13.3 = 6.65%

From the graph we can of course deduce the median which is the mid-quartile, Q_2, which is in this case **35.5%**.

Example 5.2

The table below shows the ages of employees in years who belong to a certain scheme.

Class	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
Frequency	3	5	16	20	33	18	4	1

Using the graphical method determine the semi-quartile range for the distribution.

Solution

Arranging the data in a vertical table including a column for the cumulative frequency, we have the following table, Table 5.2. Fig.5.2 below shows the smoothed cumulative frequency polygon for Table 5.2 above.

Class	Class boundaries	Frequency	Cumulative Frequency
10-19	9.5 - 19.5	3	3
20-29	19.5 - 29.5	5	8
30-39	29.5 - 39.5	16	24
40-49	39.5 - 49.5	20	44
50-59	49.5 - 59.5	33	77
60-69	59.5 - 69.5	18	95
70-79	69.5 - 79.5	4	99
80-89	79.5 - 89.5	1	100

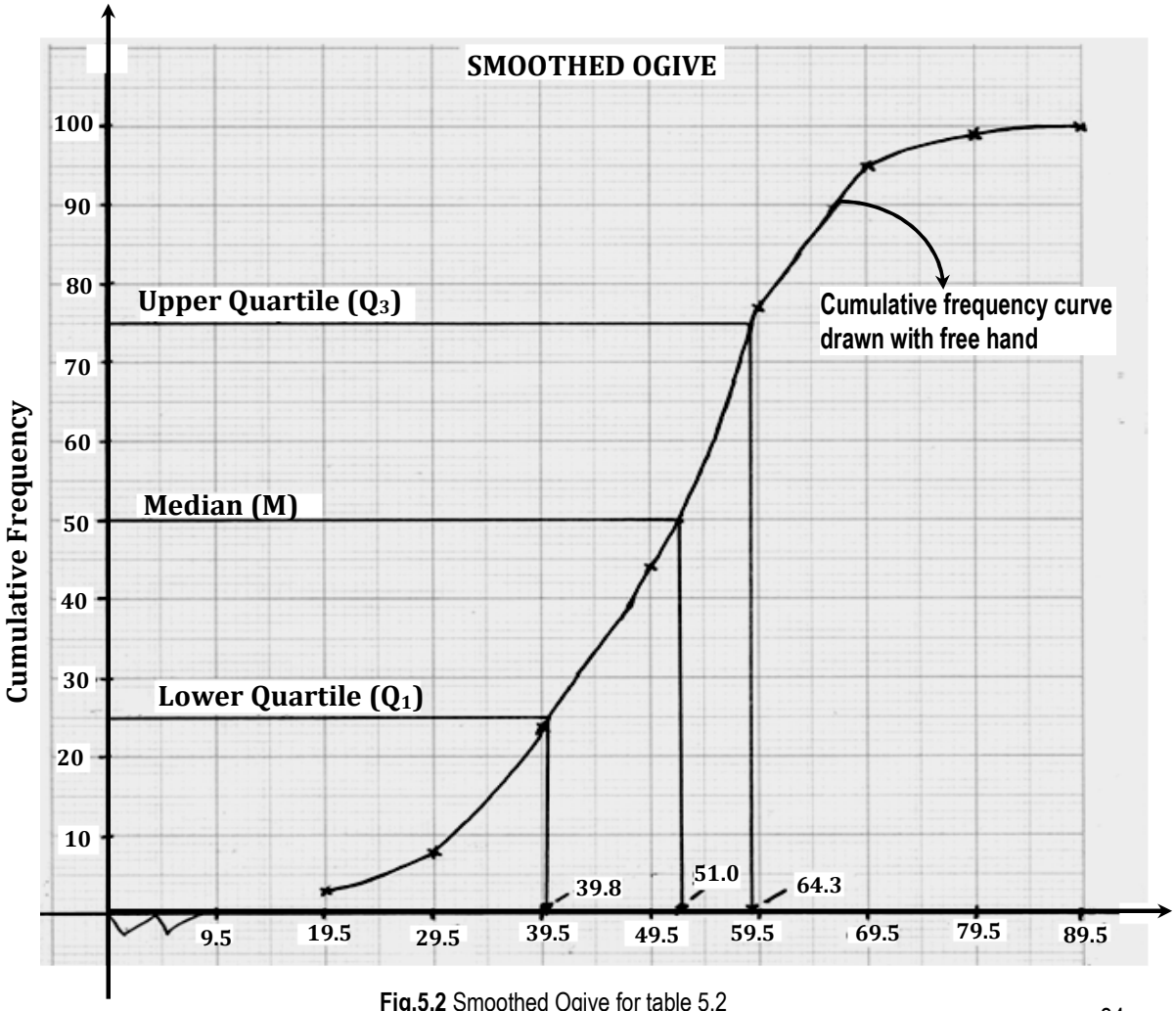


Fig.5.2 Smoothed Ogive for table 5.2

From the graph we deduce that

Q_3 = 64.3 years

Q_1 = 39.8 years

$Q_3 - Q_1$ = 64.3 – 39.8 = **24.5 years**

Hence the semi-inter quartile range = $\frac{1}{2}(Q_3 - Q_1) = \frac{1}{2} \times 24.5 = \mathbf{12.25}$

The median is **51.0years**

NUMERICAL METHOD

In this method, we determine the median, **M**, the upper quartile, $\mathbf{Q_3}$, and the lower quartile, $\mathbf{Q_1}$, by calculation.

By calculation we proceed by computing the median and then build up formulas for the upper and lower quartiles as below;

(i) Median (M)
We have already seen that the median for grouped data is calculated from the formula

$$\boldsymbol{Median(M) = L_b + \left(\frac{\frac{n}{2} - cf_b}{f_m}\right) \times C_w} \tag{5.2}$$

Since $\mathbf{M \equiv Q_2}$, then Q_2 can be calculated using the same formula. Similarly the upper and lower quartiles can, respectively, be calculated from

(ii) Upper Quartile (Q_3)

$$\boldsymbol{Q_3 = L_u + \left(\frac{\frac{3n}{4} - cf_u}{f_u}\right) \times C_w} \tag{5.3}$$

Where L_u = Lower boundary of the upper-quartile class
cf_u = cumulative frequency of a class before the upper quartile class
f_u = frequency of the upper-quartile class
n = the number of observations
c_w = class width

(iii) Lower Quartile (Q_2)

$$\boldsymbol{Q_1 = L_l + \left(\frac{\frac{n}{4} - cf_l}{f_l}\right) \times C_w} \tag{5.4}$$

Where $\boldsymbol{L_l}$ = Lower boundary of the lower quartile-class
$\boldsymbol{cf_l}$ = cumulative frequency of a class before the lower quartile-class
$\boldsymbol{f_l}$ = frequency of the lower-quartile class

Example 5.3
Using the data of Example 5.1, determine by calculation,

(i) the median
(ii) the upper quartile
(iii) the lower quartile, and
(iv) the semi-interquartile range.

Solution

We redraw Table 5.1 but including the class boundaries because we need these for the calculation.

TABLE 5.3 Frequency distribution of marks in an examination

Class	Class Boundaries	Frequency	Cumulative Frequency	
20-24	19.5 - 24.5	9	9	Lower quartile class
25-29	24.5 - 29.5	17	26	
30-34	29.5 - 34.5	21	47	
35-39	34.5 - 39.5	18	65	Median class
40-44	39.5 - 44.5	15	80	Upper quartile class
45-49	44.5 - 49.5	9	89	
50-54	49.5 - 54.5	11	100	

(i) The median class is **35 – 39**. This is the class width which contains the middle observation, i.e. the student who occupies the 50th position, since there were a total of 100 students. In addition we note that

$$n = 100,\ cf_b = 47,\ f_m = 18,\ c_w = 5,\ \text{and}\ L_b = 34.5$$

Hence we have

$$\boldsymbol{Median}\ (\boldsymbol{M}) = 34.5 + \left(\frac{\frac{100}{2} - 47}{18}\right) \times 5 \quad = 34.5 + \left(\frac{50 - 47}{18}\right) \times 5 \quad 34.5 + \left(\frac{3}{18}\right) \times 5 = \mathbf{35.33}$$

(ii) In order to calculate the upper quartile we note that
$n = 100,\ cf_u = 65,\ f_u = 15,\ c_w = 5,\ L_u = 44.5$

hence we have that

$$Q_3 = 39.5 + \left(\frac{\frac{3}{4} \times 100 - 65}{15}\right) \times 15$$

$$= 39.5 + \left(\frac{75 - 65}{15}\right) \times 15$$

$$= 39.5 + \left(\frac{10}{15}\right) \times 15 = 39.5 + 3.33 = \mathbf{42.33}\ (appr)$$

(iii) In order to calculate the lower quartile we note that
$n = 100,\ cf_l = 9, f_l = 17,\ c_w = 5,\ \text{and}\ L_l = 24.5$

$$Q_1 = 24.5 + \left(\frac{\frac{1}{4} \times 100 - 9}{17}\right) \times 5$$

$$= 24.5 + \left(\frac{25 - 9}{17}\right) \times 5$$

$$= 24.5 + \left(\frac{16}{17}\right) \times 5 = 24.5 + 4.71 = \mathbf{29.21}\ (\boldsymbol{appr})$$

The Semi-interquartile range $= \frac{42.33 - 29.21}{2} = \frac{13.12}{2} =$ **6.56**

These results agree very well with those obtained using the graphical method. Note that the results obtained by calculation are carried to two decimal places whereas those obtained by graphical method are given to the nearest whole number. This is because the results obtained by calculation are more accurate than those read off from the graph. Therefore we are justified to give our answers to two decimal places. Needless to say that even results given to one decimal place would be accurate enough. However we prefer two decimal places because the class boundaries are already given to one decimal place.

5.3 Percentiles

Percentiles divide the distribution into 100 equal parts. The percentiles are denoted by P_1, P_2, P_2, … P_{98}, P_{99}. This means that the total frequency is divided into 100 equal parts. The median, the upper quartile and the lower quartile coincide with the 50-, 75-, and 25- percentiles respectively. Percentiles can be calculated in the same way as the upper and lower quartiles have been calculated.

Example 5.3

Determine the 56-percentile for the data of Example 5.1.

Solution

The 56-percentile is denoted by P_{56}. Note that in order to calculate the upper quartile we could have replaced the fraction **3n/4** by the fraction **75n/100** in equation 5.3. Therefore in order to calculate the 46-percentile we use the same equation i.e. Equation 5.3 by replacing the fraction **3n/4** by the fraction **56n/100**.

The formula to use is thus

$$\boldsymbol{P_{56} = L_{56} + \left(\frac{\frac{56n}{100} - cf_b}{f_{56}}\right) \times C_w} \qquad (5.5)$$

Where
L_{56} = 50.5, cf_b = 296, f_{56}= 107, c_w = 10.

Hence we have

$$P_{56} = 50.5 + \left(\frac{\frac{56 \text{ x } 600}{100} - 296}{107}\right) \times 10$$

$$= 50.5 + \left(\frac{336 - 296}{107}\right) \times 10 = 50.5 + \left(\frac{40}{107}\right) \times 10$$

= 50.5 + 3.783 = **54.28**

All the other percentiles can be calculated in the same way.

5.4 Deciles

Deciles divide a distribution into 10 equal parts. The deciles are denoted by D_1, D_2, …., D_9. This means that the total frequency is divided into 10 equal parts. The median coincides with the decile**D_5**. Furthermore D_5 coincides with P_{50}. This means that we are able to calculate all the deciles by noting that

$D_1 \equiv P_{10}$, $D_2 \equiv P_{20}$, $D_3 \equiv P_{30}$, $D_4 \equiv P_{40}$, $D_5 \equiv P_{50} \equiv Q_2 \equiv$ Median, $D_6 \equiv P_{60}$, $D_7 \equiv P_{70}$, $D_8 \equiv P_{80}$, $D_9 \equiv P_{90}$

5.5 The Mean Deviation

Suppose $x_1, x_2, \ldots, x_n$ are a set of values of a variable. Let us denote the mean by x. Then the deviations of these values from the mean are;

$$\boldsymbol{d_1} = x_1 - \bar{x}$$

$$\boldsymbol{d_2} = x_2 - \bar{x}$$

$$\vdots$$

$$\boldsymbol{d_n} = x_n - \bar{x}$$

If we take the absolute values of the deviations, $|d_1|, |d_2|, \ldots, |d_n|$, then the *mean deviation* is defined by

$$Mean\ deviation = \frac{|\boldsymbol{d_1}| + |\boldsymbol{d_2}| + \ldots + |\boldsymbol{d_n}|}{\boldsymbol{n}} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}|\boldsymbol{d_i}| \tag{5.6}$$

Example 5.4

Calculate the mean deviation of the numbers
9, 11, 17, 18, 20, 21

Solution

We first calculate the mean

$$Mean = \frac{9 + 11 + 17 + 18 + 20 + 21}{6} = \frac{96}{6} = \mathbf{16}$$

We then find the deviations from the mean and their absolute values

		Deviation	Absolute Value
d_1 =	9 - 16 =	-7	7
d_2 =	11 - 16 =	-5	5
d_3 =	17 - 16 =	1	1
d_4 =	18 - 16 =	2	2
d_5 =	20 - 16 =	4	4
d_6 =	21 - 16 =	5	5

$$\text{Therefore mean deviation} = \frac{7 + 15 + 1 + 2 + 4 + 5}{6} = \frac{24}{6} = \mathbf{4}$$

Now consider a distribution whose frequencies $f_1, f_2, \ldots, f_n$ correspond to the class intervals of which $x_1, x_2, \ldots, x_n$ are the mid-values. The mean deviation of such a distribution is given by

$$Mean\ deviation = \frac{\boldsymbol{f_1}|\boldsymbol{d_1}| + \boldsymbol{f_2}|\boldsymbol{d_2}| + \ldots + \boldsymbol{f_n}|\boldsymbol{d_n}|}{\boldsymbol{f_1} + \boldsymbol{f_2} + \ldots + \boldsymbol{f_n}} = \frac{\sum \boldsymbol{f_i}|\boldsymbol{x_i} - \bar{\boldsymbol{x}}|}{\sum \boldsymbol{f_i}} \tag{5.7}$$

Example 5.5

The table below shows the weights (in Kg) of 100 sacks of beans packed at random by a certain farm. Calculate the mean deviation of the weights of the sacks.

Weight (kg)	60-62	63-65	66-68	69-71	72-74
No. of sacks	5	18	42	27	8

Solution

Representing the given data in tabular form we have

Weight (kg)	Class Mark x	Frequency f	Fx	Deviation, d from the mean	$\|d\|$	$f\|d\|$
60-62	61	5	305	-6.45	6.45	32.25
63-65	64	18	1152	-3.45	3.45	62.1
66-68	67	42	2814	-3.45	0.45	18.9
69-71	70	27	1890	2.55	2.55	68.85
72-74	73	8	584	5.55	5.55	44.4
		$\sum f = 100$	$\sum fx = 6745$			$\sum f\|d\| = 226.5$

$$The\ mean\ = \frac{\sum fx}{\sum f} = \frac{6745}{100} = \mathbf{67.45}$$

$$The\ mean\ deviation\ = \frac{\sum f|d|}{\sum f} = \frac{226.5}{100} = \mathbf{2.265}$$

5.6 The Standard Deviation

The most important measure of dispersion is the *standard deviation*, usually denoted by $\boldsymbol{\sigma}$. The standard deviation is obtained by first finding the *variance*. Suppose we square the deviations obtained in the previous problem, Example 5.5, and then find the mean of the results. The mean of the squared deviations is what is called the *variance*. The variance as it stands is not suitable as a measure of dispersion. The major disadvantage of using the variance as a measure of the dispersion is that it does not have the dimensions of the variable. Instead if we take the square root of the variance we obtain a statistical quantity possessing the units of the variable concerned. By definition, the positive square root of the variance is what is called the *standard deviation*.

Ungrouped data

Symbolically, the *standard deviation* for ungrouped data is given by

$$\sigma\ = \sqrt{\frac{\sum(\boldsymbol{x} - \bar{\boldsymbol{x}})^{\boldsymbol{2}}}{\boldsymbol{n}}} = \sqrt{\frac{\sum \boldsymbol{d^2}}{\boldsymbol{n}}} \qquad (5.8)$$

Example 5.6

Find the standard deviation of the numbers
9 11 17 18 20 21

Solution

As already seen in Example 5.4, the mean of the numbers is 16.

Deviations from the mean are

d_1 = 9 - 16 = -7 that is d_1^2 = 49
d_2 = 11 - 16 = -5 " d_2^2 = 25
d_3 = 17 - 16 = 1 " d_3^2 = 1
d_4 = 18 - 16 = 2 " d_4^2 = 4
d_5 = 20 - 16 = 4 " d_5^2 = 16
d_6 = 21 - 16 = 5 " d_6^2 = 25

The variance is

$$\sigma^2 = \frac{49 + 25 + 1 + 4 + 16 + 25}{6} = \frac{120}{6} = \mathbf{20}$$

Hence the standard deviation is

$$\sigma = \sqrt{20} = \mathbf{4.47}$$

NOTE:
The standard deviation can also be calculated from the formula,

$$\boldsymbol{\sigma = \sqrt{\overline{x^2} - \bar{x}^2}} \tag{5.9}$$

Grouped data

The standard deviation of a distribution whose frequencies f_1, f_2, …, f_n correspond to the class intervals of which x_1, x_2, …, x_n are the mid-values (class marks) is given by

$$\sigma = \sqrt{\frac{\sum \boldsymbol{f_i(x - \bar{x})^2}}{\sum \boldsymbol{f_i}}} = \sqrt{\frac{\sum \boldsymbol{fd^2}}{\sum \boldsymbol{f}}} \tag{5.10}$$

Example 5.7
Consider a distribution whose class marks and the corresponding frequencies are given below.

Class Mark	21	24	27	30	33
Frequency	2	9	20	14	5

Find the standard deviation of the distribution.

Solution

Class Mark x	Frequency f	Fx	Deviation d	d^2	fd^2
21	2	42	-6.66	44.36	88.72
24	9	216	-3.66	13.4	120.6
27	20	540	-0.066	0.44	8.8
30	14	420	2.34	5.48	76.72
33	5	165	5.34	28.52	142.6
	$\sum f = \mathbf{50}$	$\sum fx = \mathbf{1383}$			$\sum fd^2 = \mathbf{437.44}$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1383}{50} = 27.66$$

$$The\ variance = \frac{\sum fd^2}{\sum f} = \frac{437.44}{50} = 8.75$$

$$The\ standard\ deviation = \sqrt{8.75} = \mathbf{2.96}$$

Determining the standard deviation by applying the definition directly, that is, taking deviations from the mean, can lead to unnecessarily cumbersome arithmetic as you will have noticed from the examples above on the determination of the mean deviation. Again as you may have noticed, the arithmetic is easily handled when the mean is an integer. This can be easily seen by comparing Examples 5.4 and 5.5. We can therefore simplify calculations on the standard deviation greatly by using an "Assumed mean" in place of the actual mean. In Example 5.8 below, we illustrate the use of an assumed mean for a simple case and then we follow an example of a more difficult case, Example 5.9

Example 5.8

Find the standard deviation of the numbers

11 12 13 14 15 17

Solution

Using the real mean.

The real mean is **13.67** to 2 decimal places)

Number	Deviation, d	d^2
11	11 – 13.67 = -2.67	7.1289
12	12 – 13.67 = -1.67	2.7889
13	13 – 13.67 = -0.67	0.4489
14	14 – 13.67 = 0.33	0.1089
15	15 – 13.67 = 1.33	1.768
17	17 – 13.67 = 3.33	11.0889

$$\sigma = \sqrt{\frac{\sum d^2}{n}} = \sqrt{\frac{23.3334}{6}} = \sqrt{3.8889} = \mathbf{1.97}\ (to\ 2\ decimal\ places)$$

Using an assumed mean. For example let us assume the mean is 13. Then we have;

Number	Deviation, d	d^2
11	11 – 13 = -2	4
12	12 – 13 = -1	1
13	13 – 13 = 0	0
14	14 – 13 = 1	1
15	15 – 13 = 2	4
17	17 – 13 = 4	16

Note that we cannot use Equation 5.6 because the deviations are not taken from the "real" mean. Hence we have to use a modified formula. If $\boldsymbol{d_i = x_i - A}$ are the deviations of $\boldsymbol{x}$ from an assumed mean, A, and f_i are the corresponding frequencies, then the standard deviation is calculated from the formula;

$$\sigma = \sqrt{\frac{\sum \boldsymbol{fd^2}}{\boldsymbol{n}} - \left(\frac{\sum \boldsymbol{fd}}{\boldsymbol{n}}\right)^{\mathbf{2}}} \qquad (5.11)$$

Where; n = Σ f

In this example each number occurs once hence*f = 1* for each number. Hence we have

$$\frac{\sum fd^2}{n} = \frac{26}{6}$$

$$\left(\frac{\sum fd}{n}\right)^2 = \left(\frac{4}{6}\right)^2$$

That is,

$$\sigma = \sqrt{\mathbf{\frac{26}{6}}\frac{26}{6} - \left(\frac{4}{6}\right)^{\mathbf{2}^2}} = \sqrt{\frac{26 \times 6}{36} - \left(\frac{16}{36}\right)^{\mathbf{2}^2}} = \sqrt{\frac{140}{36}} = \mathbf{1}1.97\ (2dps)$$

We get exactly the same answer but with less labour.

Example 5.9

The table below shows the distribution of marks obtained in an Economics examination.

Marks	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69
Frequency	10	13	21	32	29	11	3	1

Calculate *(i) the mean*
(ii) the standard deviation

Solution

Class	Class Boundary	Class Mark	Frequency f	Deviation from A=47	(x-A)/c c=50	fu	fu^2
30-34	29.5-34.5	32	10	-15	-3	-30	90
35-39	34.5-39.5	37	13	-10	-2	-26	52
40-44	39.5-44.5	42	21	-5	-1	-21	21
45-49	44.5-49.5	47	32	0	0	0	0
50-54	49.5-54.5	52	29	5	1	29	29
55-59	54.5-59.5	57	11	10	2	22	44
60-64	59.5-64.5	62	3	15	3	9	27
65-69	64.5-69.5	67	1	20	4	4	16
			$\Sigma f = 120$			$\Sigma fu = -13$	$\Sigma fu^2 = 279$

$$TThe\ meane\ mean = \mathrm{A} + \mathrm{C}\frac{\sum \mathrm{fu}}{\sum \mathrm{f}} = 47 + 5 \times \left(\frac{-13}{120}\right)$$

$$= 47 - \frac{65}{120} = 47 - 0.54$$

$$= \mathbf{4646.46}$$
$$\mathbf{46}$$

$$TThe\ standard\ deviatione\ standard\ deviation = \mathrm{C}\sqrt{\frac{\sum \mathrm{fu}^2}{\sum \mathrm{f}} - \left(\frac{\sum \mathrm{fu}}{\sum \mathrm{f}}\right)^2}$$

$$= 5 \times \sqrt{\frac{297}{120} - \left(\frac{-13}{120}\right)^2} = 5 \times \sqrt{2.325 - 0.012} = 5 \times 1.52 = \mathbf{7.6060}$$

Note that Equation 5.10 is the equation you are going to use most often to calculate the standard deviation for grouped data.

EXERCISE 4

4.1 The following distribution of marks (out of 100) was obtained from a certain examination paper.

Mark	20-29	30-39	40-49	50-59	60-69	70-79	80-89
Frequency	2	8	14	26	28	10	4

Find P_{14}, Q_1, and D_7 for this data.

4.2 The frequency distribution below shows the ages at puberty (in months, to the nearest month) of a group of boys and girls.

Age group	**Number**	**Age group**	**Number**
120-129	4	160-169	32
130-139	12	170-179	15
140-149	28	180-189	11
150-159	22	190-199	4
200-209	2		

For this data, determine P_{66}, Q_3, and D_9**.**

4.3 The birth rate per thousand of the population in 60 towns is given in the following table, where the frequency is the number of towns whose birth rate lies in each interval

Birth rate (Centre of interval)	8	9	10	11	12	13	14	15	16
Frequency	2	4	5	10	15	9	8	5	2

Calculate the standard deviation

4.4 In the following table ***f*** is the frequency of an observation **x.**

x	2.7	2.9	3.1	3.3	3.5	3.7
f	2	7	15	21	12	3

Calculate the mean and standard deviation of x.

4.5 The following table gives the distribution by seating capacity of new bus and coach registrations during a certain year.

Seating capacity	15-20	21-26	27-32	33-40	41-48	49-56
No. of vehicles	3	2	69	390	1823	25

Calculate the mean and standard deviation of the distribution.

4.6 The frequency distribution below shows the marks obtained by 50 boys in an English test.

Marks	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
Frequency	8	4	8	7	10	6	5	2

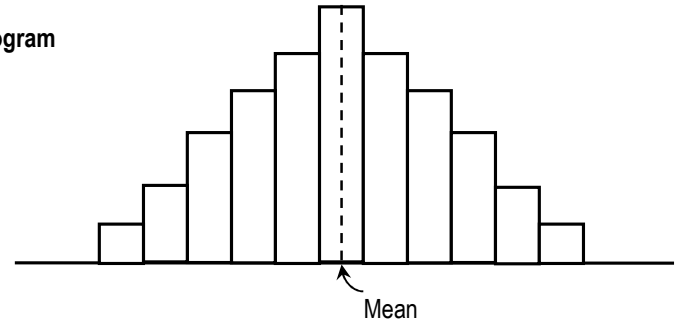
(i) Calculate the mean
(ii) Calculate the standard deviation.

6.0
MEASURES OF SKEWNESS

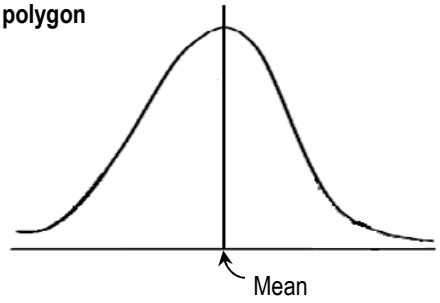
6.1 SYMMETRY

Symmetry refers to a situation where there is regularity, balance, equilibrium, evenness and proportionality within the distribution. In a symmetrical distribution, the mean, mode and median are equal as illustrated in the figure below.

Histogram



Frequency polygon



6.2: ASYMMETRY (SKEWNESS)

Where there is lack of symmetry in a distribution, skewness exists. The reason for this is that the mode, mean and median have differing values. There are two categories of skewness exhibited by distributions and these include:

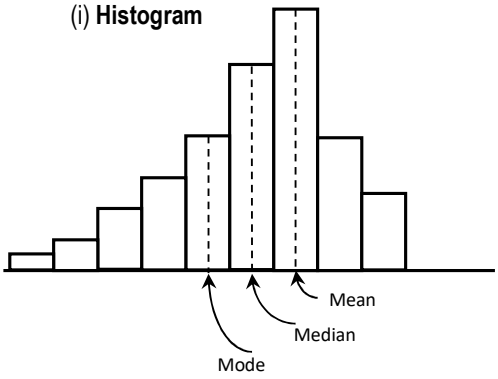
(i) Negative skewness
(ii) Positive Skewness

These are explained as below;

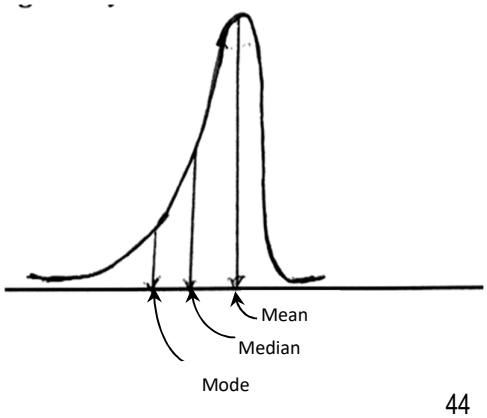
(i) Negative skewness

In a negatively skewed distribution, the tail of the distribution is pulled in the negative direction.
In such a situation, Mean < Median < Mode; Where < is a symbol for *"less than*"

(i) **Histogram**



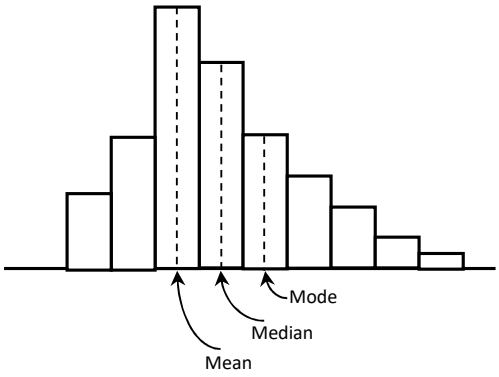
(ii) **Frequency polygon**



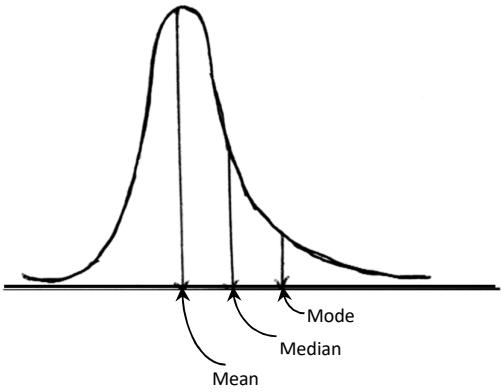
(ii) **Positive Skewness**

In a positively skewed distribution, the tail of the distribution is pulled in the positive direction.
In such a situation, mode < median < mean.

(i) Histogram



(ii) Frequency polygon



When distributions are skew, the median generally lies between the mode and the mean, and the following relation is satisfied;

$$\textbf{Mean} - \textbf{Mode} = 3\,(\textbf{Mean} - \textbf{Median}) \qquad (6.1)$$

6.3: MEASURES OF SKEWNESS

Measures of skewness study the degree of departure from symmetry of any given distribution. There are basically three relative mathematical measures of skewness and these include;

(a) Karl Pearson's coefficient of skewness

This measures skewness on the basis of measures of central tendency that include the mode, the mean and median plus the standard deviation (measure of dispersion) and the formula is given by;

$$Sk = \frac{\text{Mean - mode}}{\text{Standard deviation}} \qquad (6.2)$$

If mean > mode, the skew is positive.
If Mean < mode, the skew is negative.
If Mean = mode, the skew is zero and the distribution is symmetrical.

Alternatively;

$$\text{Pearson's coefficient of skewness} = \frac{\textbf{3 (Mean – Median)}}{\textbf{Standard deviation}} \qquad (6.3)$$

The above equation implies that skewness can take any value between 3 and -3.

Example 6.1;

The table below shows marks scored by 125 employees in an aptitude test.

Marks	25-28	28-29	29-30	30-31	31-32	32-33	33-34	34-35	35-40
Frequency	6	12	27	30	18	14	9	4	5

Draw a histogram to represent this data and compute the following:

(i) The median
(ii) The mean
(iii) Standard deviation
(iv) Pearson's coefficient of skewness
(v) Comment on the skewness of this distribution.

Solution
Median = **30.6,** Mean = **30.892,** Standard deviation = **2.219**

Skewness = $\frac{3\ (\text{Mean} - \text{Median})}{\text{Standard deviation}}$ = $\frac{(30.892 - 30.58)}{2.219}$ = **0.42 (2dps)**

Since skewness > 0, the distribution is positively skewed.

Now illustrate using the graphical method by constructing a histogram and frequency polygon on a graph paper.
From your respective graphs, you should be able to observe that the histogram and the frequency polygon further confirm that the distribution is skewed to the right i.e. positively skewed.

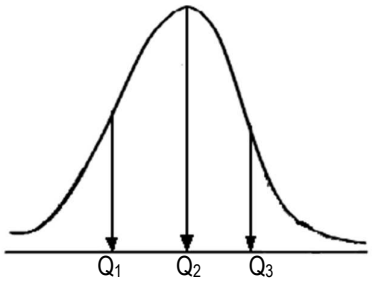
(b) The Quartile coefficient of skewness

Another measure of skewness is defined in terms of the quartiles. This measure of skewness was forward by Bowel and is also referred to as the Bowel's coefficient of skewness. It measures skewness of a distribution basing on the relative positions of the median, the upper plus the lower quartiles denoted by Q_3 and Q_1respectively. In this regard therefore, the coefficient of skewness is given by;

$$Sk = \frac{(Q_3 - Q_1) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - Q_2}{Q_3 - Q_1}$$

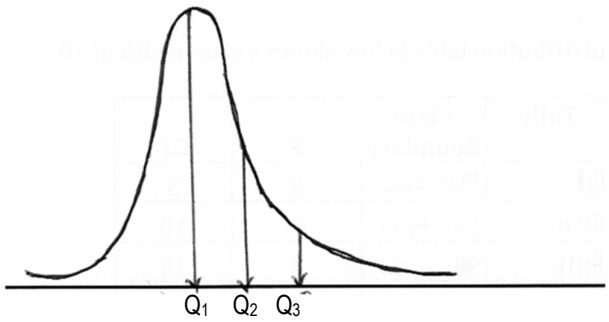
Illustrations

- **Symmetry Distributions**



$Q_3 - Q_2 = Q_2 - Q_1$
Quartile skewness = 0

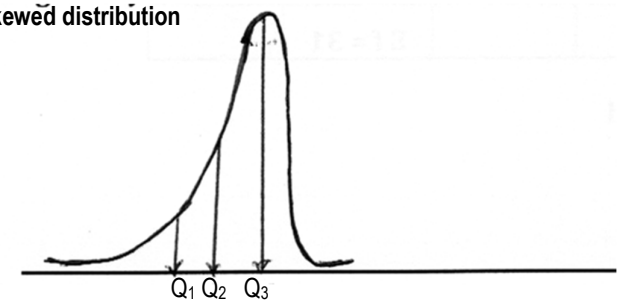
- **Positively skewed skewness**



$Q_3 - Q_2 > Q_2 - Q_1$
skewness> 0

- **Negatively skewed distribution**

$Q_3 - Q_2 < Q_2 - Q_1$
Quartile skewness< 0



Example 6.2
31 students tried to estimate the length of a line. The line was actually 60mm long. These are their results in millimeters.

61	70	46	44	26	23	30	83	52	44	61
37	49	59	58	63	31	29	37	48	76	
46	31	38	41	49	52	56	75	61	38	

(i) Find the median and the quartiles of this distribution and use the quartiles to estimate the skewness
(ii) Draw a histogram with equal class intervals to represent the data.

Solution
The frequency distribution table below shows a class width of 10.

Length /mm	Tally	Class	F	C.F
20-30	𝍸	19.5-30.5	3	3
30-40	𝍸 ||	29.5-40.5	7	10
40-50	𝍸 |||	39.5-50.5	8	18
50-60	𝍸	49.5-60.5	5	23
60-70	𝍸	59.5-70.5	4	27
70-80	𝍸	69.5-80.5	3	30
80-90	|	79.5-90.5	1	31
			Σ f = 31	

(a) Numerical Method
$Q_1 = 37$, $Q_3 = 61$
$Q_3 - Q_1 = 61 – 48 = 13$
$Q_2 - Q_1 = 48 – 37 = 11$

Since $Q_3 - Q_1 > Q_2 - Q_1$, the distribution is positively skewed.

Furthermore,
Quartile coefficient of skewness = $\frac{(Q_3-Q_1) – (Q_2-Q_1)}{Q_3-Q_1} = \frac{13 – 11}{61 – 37} =$ **0.083** (3dps)

This further confirms the positive skew.
Graphically, the shape of the histogram further confirms the positive skew!

(c) Kelly's coefficient of skewness

Kelly's coefficient of skewness is also referred to as the Percentile coefficient of skewness and it measures skewness basing on the percentiles of the distribution, i.e. the 10th and 90th percentiles and the median or 50th percentile of the distribution. Since percentiles are also positional averages like quartiles, Kelly's measure of skewness is also applicable in situations as those of Bowel's measure.

It is given by;

$$Sk = \frac{P_{90} + P_{10} – 2P_{50}}{P_{90} – P_{10}} \qquad (6.5)$$

Where;
P_{90} is the ninetieth (90th) percentile
P_{10} is the tenth (10th) percentile
P_{50} is fiftieth (50th) percentile

EXERCISE 5

5.1 The following table shows the time to the nearest minute spent reading during a particular day by a group of school children.

Time	10-19	20-24	25-29	30-34	30-39	40-44	45-49
No. of children	2	8	14	26	28	10	4

(a) Represent this data on a histogram.
(b) Comment on the shape of the distribution

5.2: The frequency distribution below shows a given set of data.

X	0	1	2	3	4	5	6	7	8	9
f	5	22	46	38	31	23	16	11	6	2

(i) Calculate Pearson's coefficient of measure of skewness for the above data.
(ii) Construct a frequency polygon for the above data and state how the above skewness is reflected in the shape of your polygon.

5.3: For a skewed distribution, the mean is 86, the mode is 78 and the variance is 16. Calculate Pearson's coefficient of skewness and sketch the curve.

5.4: For a skewed distribution, the mean is 5. Calculate Pearson's coefficient of skewness and sketch the curve.

5.5: The following table gives the blood pressure of 60 students

Blood Pressure	95-	105-	110-	115-	120-	125-	130-	135-	140-	150-
frequency	2	5	6	9	14	3	6	5	4	6

Find;
(i) Pearson's coefficient of skewness.
(ii) The quartile coefficient of skewness
(iii) Draw the histogram for the above data

5.6 The following table gives the distribution of 1200 families according to the number of children.

No. of children in a family	0	1	2	3	4	5	6	7	8
No. of families	30	200	310	310	200	80	40	20	10

Draw the frequency polygon and use it to comment on the nature of skewness.

PART B

PROBABILITY & DISTRIBUTIONS

7.0 PROBABILITY

7.1 Introduction

Statistical inference and the testing of hypothesis involve the concept of chance or probability. The word probability has been defined in several ways but the simplest definition is to state that probability is *the proportional frequency of occasions on which some stated event occurs.*
The probability of a tossed coin giving heads is the proportion of times that heads occur in repeated tossing of the coin. After a thousand tossing this proportion would be expected to be quite close to 0.50 and in an infinite set of tossing of which the thousand are apart, it seems reasonable to assume that the proportion would be exactly 0.5.
If we had a box containing 70 white and 30 black balls, well mixed, and were to draw 1 ball at random, the chance of the ball being black is said to be 30 out of 100, and the chance of its being white would be 70 out of 100 or 0.7. This can be interpreted to mean that if we made 1000 random draws, each time replacing the drawn ball and remixing the contents of the box, the percentage of black balls drawn would be about 30, and of white draws about 70.

Definition:

If an event can happen in A ways and fail in B ways, all possible ways being equally likely, the *probability* of its occurring is $\frac{A}{A+B}$ and of its failing is $\frac{B}{A+B}$.

Thus a probability number is the ratio of the number of favourite events to the total number of events, and it is therefore necessary that we are able to enumerate events in order to arrive at a probability number.

Event: Tossing a coin, throwing a die or picking a ball from a number of balls are all *events*.

Outcome: The result of tossing a coin, throwing a die, or picking a ball is said to be the outcome.

An example would be, if I toss a coin and it falls heads up. In that case "the outcome is a head" similarly if I pick a red ball from a number of balls, the "the outcome is a red ball".

7.2 Casting a die

When an ordinary die is cast, each of the numbers *one, two, three, four, five* and *six* has an equal chance of falling uppermost. These six outcomes of the trial are said to be exhaustive and mutually exclusive because one of the six numbers must occur and only one can occur in addition there are only six numbers available since a normal die has only six faces. There is, therefore, 1 chance out of 6 obtaining any particular number and we say that;

the probability of throwing a six = $\frac{1}{6}$
This is often abbreviated to;

$$P(six) = \frac{1}{6} \qquad or \qquad P(6) = \frac{1}{6}$$

7.3 Playing cards

An ordinary pack of 52 cards contains 4 aces and hence, if we select at random 1 card from a well-shuffled pack, there are 4 chances out of 52 of it being an ace. Thus,

$$the\ probability\ of\ selecting\ an\ ace = \frac{4}{52} = \frac{1}{13}\ or\ P(any\ ace) = \frac{1}{13}$$

As you well know, there is only one ace of spades in a pack of 52 playing cards and therefore, the probability of selecting the ace of spades is

$$p(ace\ of\ spades) = \frac{1}{52}$$

7.4 Tossing a coin

When a coin is tossed it may fall either as a *head* or a *tail*. Denoting the outcome "head" as **H** and the outcome "tail" as **T**, we say

$$the\ probability\ of\ a\ head = P(H) = \frac{1}{2}$$

$$the\ probability\ of\ a\ tail = P(T) = \frac{1}{2}$$

7.5 Probability

The probability of an event is a number between 0 and 1. A probability cannot be less than zero nor take on a value greater than 1. if you are certain that the event must occur then its probability is equal to 1. On the other hand if you are extremely sure that the event cannot occur, then its probability is 0. Any other probability figure must lie between 0 and 1.

$$The\ probability\ of\ an\ event\ happening\ will\ be\ represented\ by = \boldsymbol{P(Event)}$$

$$The\ probability\ of\ an\ event\ not\ happening\ will\ be\ represented\ by = \overline{\boldsymbol{Event}}$$

Example 7.1

Suppose we toss a die. What is the probability

(i) Of obtaining a 4?
(ii) Of obtaining a 4 or a 6?
(iii) Of obtaining an odd number?
(iv) Of 3 not appearing?

Solution

Total number of outcomes = 6

The outcomes are: 1, 2, 3, 4, 5,6.

(i) $P(4) = \frac{1}{6}\ \mathbf{\frac{1}{6}}$

(ii) $PP(4 \text{ or } or\ 6) = \text{P}P(4) + \text{P}P(6) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \mathbf{\frac{1}{3}}$

(iii) $\text{P}P(odd\ number odd\ number) = PP(1 \text{ or} or\ 3 \text{ or } or\ 5) = P(1) + P(3) + P(5)$
$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \mathbf{\frac{1}{2}}$

(iv) $P(3\ not\ appearing) = P(\bar{3}) = 1 - P(3) = 1 - \frac{1}{6} = \mathbf{\frac{5}{6}\frac{5}{6}}$

This is also obtained by P(getting any other number except 3); i.e.

$PP(3 \text{ not appearing} not\ appearing) = \text{P}P(1 \text{ or } or\ 2 \text{ or } or\ 4 \text{ or } or\ 5 \text{ or } or\ 6)$
$= P(1) + P(2) + P(4) + P(5) + P(6)$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \mathbf{\frac{5}{6}\frac{5}{6}}$$

7.6 Statistical Experiments

Suppose we toss two dice simultaneously and find the sum of the numbers which appear on the top faces.
We shall call the process of tossing the two dice an experiment. In statistics an experiment is any process that generates raw data.
In the above experiment all the possible outcomes may be set out in a table called the *sample space* (also called the *possibility space*). Table 7.1 below shows the possibility space.

TABLE 7.1

Possible outcomes of tossing *two* dice simultaneously.

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

Definitions:

The following definitions should be noted.

1. A sample space is a set whose elements represent all the possible outcomes of an experiment.
2. An element of a sample space is called a sample point.
3. An event is a subject of a sample space.

Example 7.2
(a) What is the probability of obtaining a sum of 7 when two dice are tossed simultaneously?
(b) What is the probability of obtaining an odd sum?

Solution

From the probability space, **Table 7.1** above, there are 36 outcomes. The sample space is the set

S = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

$$(aa) \quad \mathrm{P}P(7) = \frac{6}{36} = \frac{1}{6}$$

$$(b) \quad P(odd\ Sum) = P(3\ or\ 5\ or\ 7\ or\ 9\ or\ 11)$$

$$= P(3) + P(5) + P(7) + P(9) + P(11)$$

$$= \frac{3}{36} + \frac{4}{36} + \frac{6}{36} + \frac{4}{36} + \frac{2}{36} = \frac{18}{36} = \mathbf{\frac{1}{2}} = \frac{18}{36} = \mathbf{\frac{1}{2}}$$

Example 7.3
Write out the probability space when three coins are tossed once.

(a) What is the probability of obtaining heads only?
(b) What is the probability of obtaining at least one head?

(c) What is the probability of obtaining at least two heads?

Solution

The set of outcomes may be arranged in a table as follows:

TABLE 7.2
Possible outcomes of tossing *three* coins once.

COIN 1	COIN 2	COIN 3
H	H	H
H	H	T
H	T	T
H	T	H
T	H	H
T	H	T
T	T	H
T	T	T

We have a set of 8 outcomes

$(a)\ \ P(Heads\ only) = P(H\ H\ H) = \frac{1}{8}$

$(b)\ \ P(at\ least\ one\ head) = P(HHH, HHT, HTH, THH, HTT, THT, TTH) = \mathbf{\frac{7}{8}}$

This is also given by;

$$1 - P(no\ head) = 1 - P(TTT) = 1 - \frac{1}{8} = \mathbf{\frac{7}{8}}$$

$(c)\ \ P(at\ least\ two\ heads) = P(HHH, HHT, HTH, THH) = \frac{4}{8} = \mathbf{\frac{1}{2}}$

We can now formalise the idea of probability already acquired by stating the following definition:

Definition: If an experiment can result in any one of **N** different equally likely outcomes, and if exactly ***n*** of these outcomes correspond to event **A**, then the probability of event **A** is

$$\boldsymbol{P(A) = \frac{n}{N}} \qquad (7.1)$$

Example 7.4

Suppose we have 3 red and 5 black balls (all identical in size and material) in a bag, what is the probability that when you pick one ball at random from the bag it will be a red ball?

Solution

There are 8 possible outcomes. The event is picking a red ball. This event consists of only 3 outcomes.

$Hence\ P(Red) = \mathbf{\frac{3}{8}}$

7.7 Theorems of Probability

So far we have been dealing with a single event at a time. However most problems in probability involve two or more events which leads us to two useful theorems. Therefore we state the theorems to distinguish between two sets of events, *mutually exclusive events* and *independent events*.

Definition: Two events A and B are mutually exclusive if they cannot occur at the same time. This simply means that one event occurring excludes the other from occurring. A simple example is when one tosses a single coin. A single coin can only show one face. The event "falling heads uppermost" excludes the event "falling tails uppermost".

Definition: Two events A and B are independent if the occurrence of one event does not affect the occurrence of the other event. If one event is a coin falling heads uppermost, and the other is drawing a seven of hearts from a pack of playing cards, then the two events are clearly independent. The fact that your coin has fallen heads uppermost does not mean that therefore you cannot draw a seven of hearts from a pack of playing cards.

The theorem we are going to state, Theorem 1, generally, applies to any two events.

Theorem 1: If A and B are any two events, then

P(AUB) = P(A) + P(B) – P(A∩B) (7.2)

In particular for mutually exclusive events,

P(A∩B) = Φ (7.3)

Hence for mutually exclusive events, Theorem 1 reduces to

P(AUB) = P(A) + P(B) (7.4)

Equation 7.3 may be represented in a Venn diagram as in Fig 7.1 below.

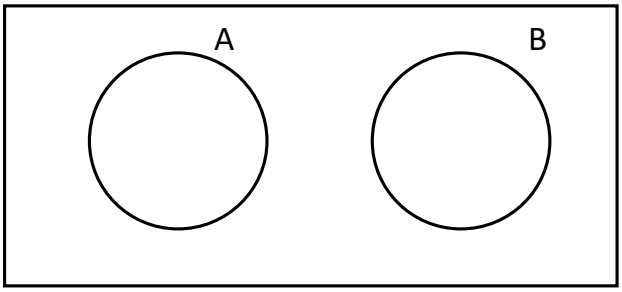


FIG.7.1

The two events do not intersect. From the Venn diagram we can conclude that

$\bar{A} \cap B = B$ (7.5)

And $\mathbf{A \cap B = A}$ (7.6)

Example 7.5

Suppose two dice are tossed, let A be the event that the sum is an odd number and B is the event that the sum is a prime number. What is P(A U B)?

Solution

These events are *not* mutually exclusive since some odd numbers are also prime numbers. Hence we use equation (7.2).

P(A U B) = P(A) + P(B) – P(A ∩ B)

Using Table 7.1, p. 78, we have

$$P(A) = P(Sum\ is\ odd) = \frac{18}{36}$$

$P(A) = P(Sum\ is\ prime) = \frac{15}{36}$

$P(A \cap B) = \frac{14}{36}$

$Hence;\ P(A \cap B) = \frac{18}{36} + \frac{15}{36} - \frac{14}{36} = \mathbf{\frac{19}{36}}$

This can be verified by actual counting as follows:

The prime numbers in the sample space are:

2, 3, 3, 5, 5, 5, 5, 7, 7, 7, 7, 7, 7,11, 11.

These are 15 elements in the sample space. The odd numbers which are not prime are: 9, 9, 9, 9

These are 4 elements in the sample space.

Total = 15 + 4 = 19 elements.

$Thus;\ P(A \cap B) = \frac{19}{36}\ \ as\ the\ formula\ shows.$

Example 7.6

When two dice are tossed, what is the probability of getting a sum of 3 or a sum of 5?

Solution

These events are mutually exclusive since a sum of 3 and a sum of 5 cannot both occur on the same toss.

$P(Sum\ is\ 3) = \frac{2}{36}$

$P(Sum\ is\ 5) = \frac{4}{36}$

Using equation (7.6) we have

$P(Sum\ is\ 3 \cup\ Sum\ is\ 5) = \frac{2}{36} + \frac{4}{36} = \frac{6}{36} = \mathbf{\frac{1}{6}}$

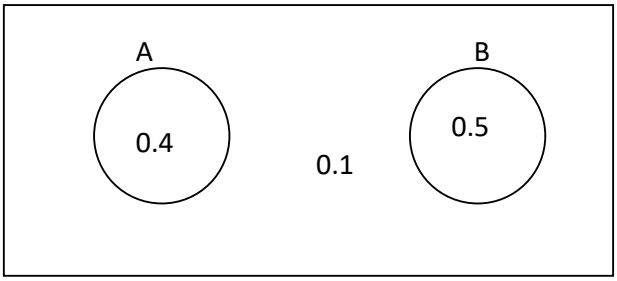
Example 7.7

If A and B are mutually exclusive events and P(A) = 0.4 and P(B) = 0.5, find

(i) P(A U B)
(ii) P(Ā)
(iii) P(Ā ∩ B)

Solution

Since A and B are mutually exclusive events, then A ∩ B = Φ. Drawing a Venn diagram for the problem above we have



(i) P(A U B) = P(A) + P(B) = 0.4 + 0.5 = **0.9**
(ii) P($\bar{A}$) = 1 – P(A) = 1 – 0.4 = **0.6**
(iii) P($\bar{A}$ ∩ B) = P(B) = **0.5**

Theorem 2 which we are going to state below refers particularly to independent events and states as follows:

Theorem 2: Two events A and B are independent if only and only if

$$\mathbf{P(A \cap B) = P(A) \,.\, P(B)} \qquad \mathbf{(7.7)}$$

We shall follow with some examples to illustrate this theorem.

Example 7.8

Suppose a boy tosses a die, and then tosses a coin, what is the probability of obtaining a six on the die and a head on the coin?

Solution

Let A be the event of obtaining a *six* on the die, and B a head on the coin. These events are clearly independent since obtaining a *six* on the die does not influence a *head* on the coin.

Hence using Equation 7.4 we have

$$\mathrm{P(A \cap B)} = \mathrm{P(A)} \times \mathrm{P(B)} = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

Example 7.9

If the probability that it will rain on the 1st of January 1999 is 0.7 and the probability that it will rain on the 1st of January 2000 is 0.5, what is the probability that it will rain on both days?

Solution

A particular place getting rain on the 1st of January 1999 and again getting rain on the 1st of January the following year are purely two independent events under normal circumstances.

Let A be the event "it will rain on the 1st January 1999"
Let B be the event "it will rain on the 1st January 2000"

Then using the Equation 5.7 we have

P(A and B) = P(A) × P(B) = 0.7 × 0.5 = 0.35

7.8 Conditional Probability

The probability that an event A has already occurred is called a *conditional probability*.

We denote this by **P(B/A)** (7.8)

which reads "the probability that B will occur given that A has already occurred".

Or simply

"the probability of B given that A".

To make the concept of conditional probability clear, let us illustrate with examples.

Example 7.10

A box contains 5 black and 3 white balls. Find the probability that when two balls are drawn *without* replacement, the second ball will be black.

Solution

The probability that the second ball picked is black depends on what colour the ball picked first was. Hence this is a conditional probability problem.

"Let B be the event "picking a black ball"
"Let W be the event "picking a white ball"

There are two possibilities: A second ball picked can be black when the first was white or black. The probability that the first ball picked is white is

$$P(W) = \frac{3}{5+3} = \frac{3}{8}$$

This would leave only 2 white balls and 5 black balls since the ball picked is not replaced.

The probability that the second ball picked is black given that the first was white is

$$P(B/W) = \frac{5}{2+5} = \frac{5}{7}$$

The probability that the first ball picked is black is

$$P(B) = \frac{5}{5+3} = \frac{5}{8}$$

This would leave only 4 black balls and 3 white balls. That is a total of 7 balls.

The probability that the second ball picked is black given that the first was also black is

$$P(B/B) = \frac{4}{3+4} = \frac{4}{7}$$

Therefore,

The probability that the second ball picked is black is given by

P(picking *White* and the Black) + P(picking *Black* and the *Black*)

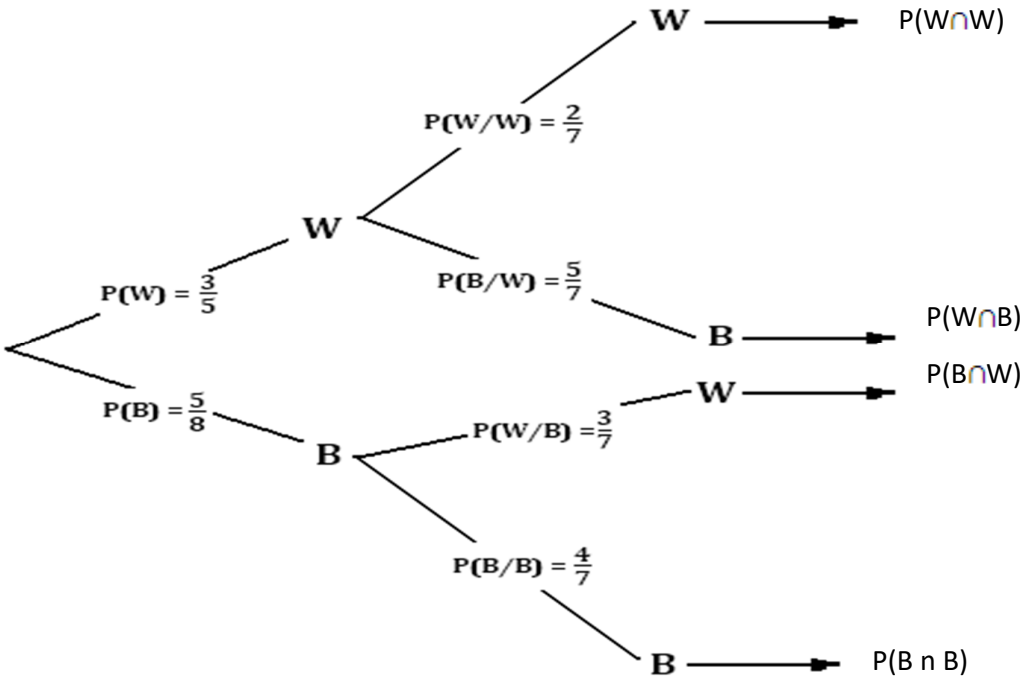
= P(W ∩ B) + P(B ∩ B)
= P(W) . P(B/W) + P(B) . P(B/B)

$$= \frac{3}{8} \times \frac{5}{7} + \frac{5}{8} \times \frac{4}{7}$$

$$= \frac{15}{56} + \frac{20}{56} = \frac{35}{56} = \frac{5}{8}$$

Reasoning out the solution to a problem of this nature according to the outcome given above can often lead to uncalled for errors because one can get lost in the whole argument. The best way to tackle the problem is to use a ***TREE-DIAGRAM***.

A tree-diagram for the above problem is shown below.



From the-diagram we observe the following:

$$P(W \cap W) = P(W/W).\ P(W) = \frac{2}{7} \times \frac{3}{8} = \frac{6}{56}$$

$$P(W \cap B) = P(B/W).\ P(W) = \frac{5}{7} \times \frac{3}{8} = \frac{15}{56}$$

$$P(B \cap W) = P(W/B).\ P(B) = \frac{3}{7} \times \frac{5}{8} = \frac{15}{56}$$

$$P(B \cap B) = P(B/B).\ P(B) = \frac{4}{7} \times \frac{5}{8} = \frac{50}{56}$$

When we add all the possible probabilities we obtain

$$Total\ probability = \frac{6}{56} + \frac{15}{56} + \frac{15}{56} + \frac{20}{56} = \frac{56}{56} = 1$$

This exhausts all the possibilities!

An alternative method of illustrating the above possibilities is to draw what is called a *CONTINGENCY TABLE* below.

∩	FIRST DRAW W B	
SECOND DRAWW B	6/56 15/56 15/56 20/26	21/56 35/56
	21/56 35/56	56/56

From the foregoing example we can state the *MULTIPLICATIVE RULE* of Conditional probability, thus:

If **A** and **B** are events, then the conditional probability of **A** given **B** is denoted by **P(A/B)** and is defined by

$$\boldsymbol{P(A/B)} = \frac{\boldsymbol{P(A \cap B)}}{\boldsymbol{P(B)}} \quad (7.9)$$

Provided P(B) ≠ 0 or B ≠ Ø.

We can also write Equation 7.9 in the form

P(A ∩ B) = P(A/B) . P(B) (7.10)

For independent events we can state the following rules:

Two events A and B are independent if

$$(i) \qquad P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A).P(B)}{P(B)} = P(A)$$

$$(ii) \qquad P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B).P(A)}{P(A)} = P(B)$$

Example 7.11

An urn contains 20 mangoes of which 5 are bad. If 2 mangoes are selected at random from the basket in succession without replacing the first, what is the probability that both mangoes are bad?

Solution

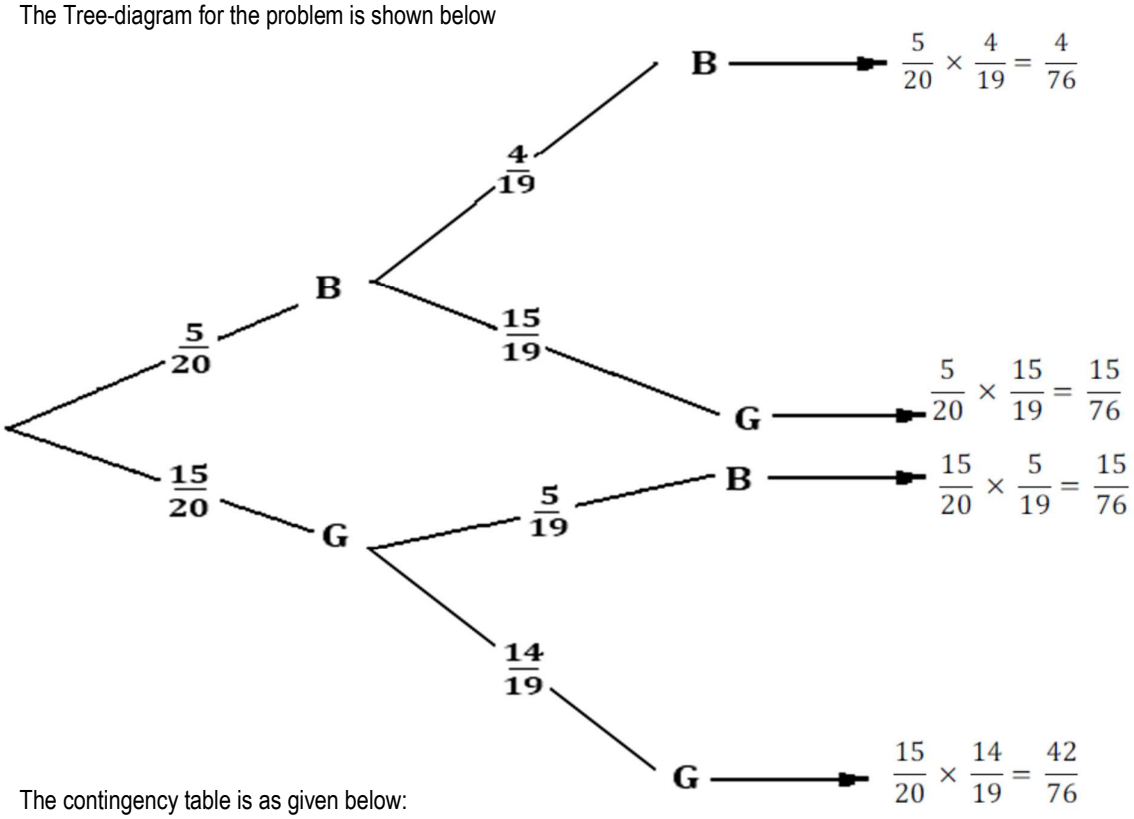
Let A represent the event 'first mango is bad"
Let B represent the event "second mango is bad".

$$Then \quad P(A) = \frac{5}{20} = \frac{1}{4}$$

$$P(B/A) = \frac{4}{19}$$

$$P(A \cap B) = P(B/A).P(A) = \frac{4}{19} \times \frac{1}{4} = \frac{1}{19}$$

The Tree-diagram for the problem is shown below



The contingency table is as given below:

∩	FIRST DRAW B	G	
SECOND DRAW B	4/76	15/76	19/76
G	15/76	42/76	57/76
	19/76	57/76	76/76

Example 7.12

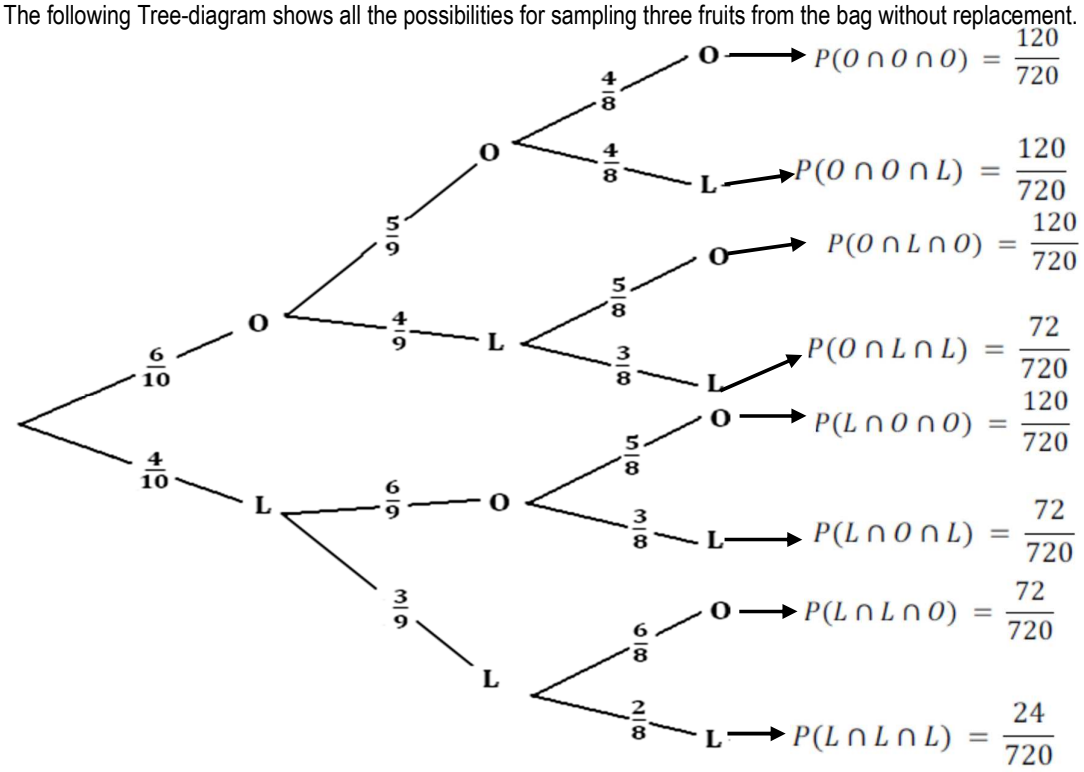
A bag contains 6 oranges and 4 lemons; sampling is random without replacement. Obtain the probabilities of the following events:

(i) the first drawn is an orange.
(ii) the first two fruits drawn are oranges.
(iii) the second fruit is a lemon, given that the first was an orange.
(iv) the third fruit is an orange.

Solution

Let O represent the event "picking an orange".
Let L represent the event "picking a lemon".

The following Tree-diagram shows all the possibilities for sampling three fruits from the bag without replacement.



The answers are easily deduced from the Tree-diagram as follows:

First fruit is an orange:

$$P(O) = \mathbf{\frac{6}{10}}$$

The first two fruits drawn are oranges:

$$P(O \cap O) = \frac{6}{10} \times \frac{5}{9} = \frac{30}{90} = \mathbf{\frac{1}{3}}$$

The second fruit drawn is a lemon given that the first was an orange:

$$P(L/O) = \mathbf{\frac{4}{9}}$$

The third drawn is an orange:

$$= P(O \cap O \cap O) + P(O \cap L \cap O) + P(L \cap O \cap O) + P(L \cap L \cap O)$$

$$= \frac{120}{720} + \frac{120}{720} + \frac{120}{720} + \frac{72}{720} = \frac{432}{720} = \mathbf{\frac{3}{5}}$$

Example 7.13

One bag contains 5 white and 7 black balls, and another contains 3 white and 9 balls. One ball is taken from the first bag and placed in the second. After thorough mixing, one ball is taken from the second bag and placed in the first bag. What is the probability that there are 5 white balls in the first bag at the end?

Solution

Bag 1

5W 7B

Bag 2

3W 9B

If the ball picked is white we have after picking

Bag 1

4W 7B

Bag 2

4W 9B

If the ball picked is black we have after picking

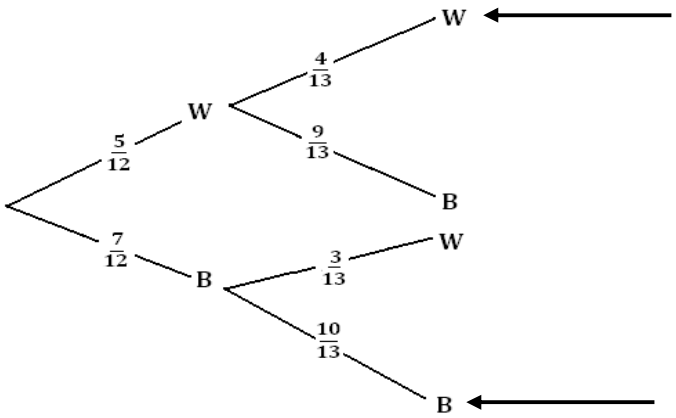
Bag 1

5W 6B

Bag 2

3W 10B

Hence we can draw a Tree-diagram for the problem



In order to have 5 White balls in the first bag again then we have either

to pick a White ball first and then pick another White ball a second time which implies

$$\boldsymbol{P(W \cap W)} \ or;$$

to pick a Black ball first and then pick another Black ball a second time which implies:

$$\boldsymbol{P(B \cap B)}.$$

Both instances are indicated with arrows on the Tree-diagram.
Hence the required probability is

$$\boldsymbol{P(W \cap W) + P(B \cap B)} = \left(\frac{5}{12} \times \frac{4}{13}\right) + \left(\frac{7}{12} \times \frac{10}{13}\right) = \frac{20}{156} + \frac{70}{156} = \frac{90}{156} = \mathbf{\frac{45}{78}}$$

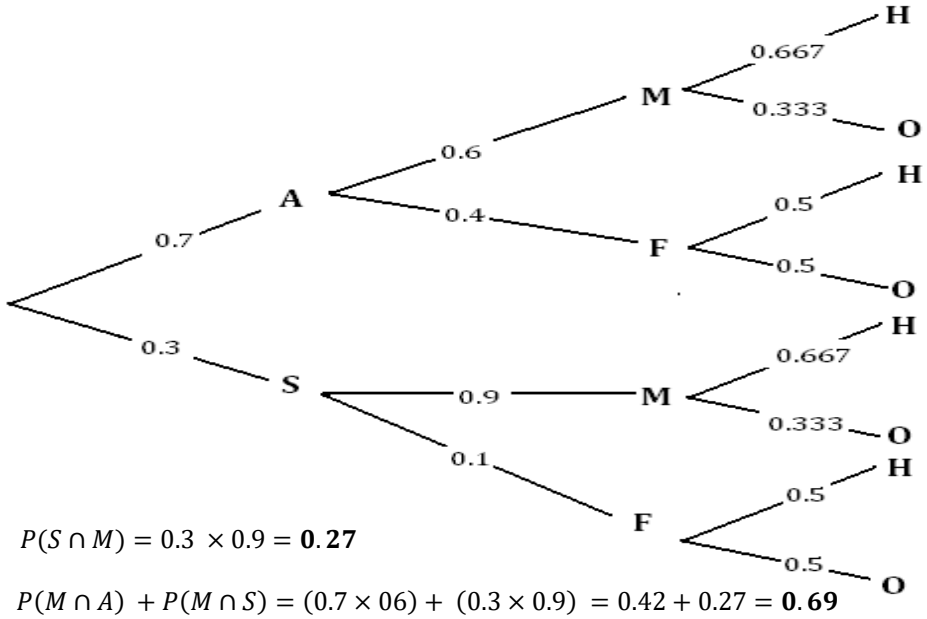
Example 7.14

At a particular University 70% of the students are reading Arts subjects and 30% are reading Science. Of the Arts students 40% are female, while of the Science students 10% are female. Of the men, 331/3% are in halls of residence, the corresponding figure for women being 50%. Residence in a hall is independent of whether a student studies Arts or Science. Find the probability that;

(i) a student chosen at random is a male Science student.
(ii) A student chosen at random is male.
(iii) A male student chosen at random is a Science student.
(iv) A Science student chosen at random is female and lives in a hall.
(v) A female student, living in a hall, chosen at random, is a Science student.

Solution

A Tree-diagram for the above is as follows:
A = Arts, **S** = Science, **M** = Male, **F** = Female, **H** = Resident in a Hall, **O** = Not resident in a Hall.



(i) $P(S \cap M) = 0.3 \times 0.9 = \mathbf{0.27}$

(ii) $P(M \cap A) + P(M \cap S) = (0.7 \times 06) + (0.3 \times 0.9) = 0.42 + 0.27 = \mathbf{0.69}$

$(iii)\ \frac{P(M \cap S)}{P(M)} = \frac{0.3 \times 0.9}{(0.3 \times 0.9) + (0.7 \times 0.6)} = \frac{0.27}{0.27 + 0.42} = \frac{0.27}{0.69} = \mathbf{\frac{9}{23}}$

$(iv)\ \frac{P(S \cap F \cap H)}{P(S)} = \frac{0.3 \times 0.1 \times 0.5}{0.3} = \mathbf{0.05}$

$(v)\ \frac{P(F \cap H \cap S)}{P(F \cap H \cap S) + P(F \cap H \cap A)} = \frac{0.3 \times 0.1 \times 0.5}{(0.3 \times 0.1 \times 0.5) + (0.7 \times 0.4 \times 0.5)}$

$$= \frac{0.03}{(0.03) + (0.285)} = \mathbf{\frac{3}{31}}$$

EXERCISE 6

6.1 In a shooting competition, the probability that participants P, Q and H hit the target are ½, $1/3$, and ¼ respectively. Given that all three participants open fire at the same time each releasing one bullet, find the probability that:

(i) one and only one bullet hits the target.
(ii) at least one bullet hits the target.

6.2 A bag contains ten oranges and eight lemons. A fruit is drawn at random from the bag one after the other without replacement. Find the probability that:

(i) the first is an orange
(ii) the first two are lemons.

6.3 A basket contains 20 mangoes of which 5 are bad. If 2 mangoes are selected at random from the basket in succession without replacing the first, what is the probability that both mangoes are bad?

6.4 In a certain large company, 70% of the employees are men and 30% are women. It has been determined that 80% of the men smoke and 60 % of the women. What is the probability that a person who smokes is a woman?

6.5 In a container there are 6 black balls and 4 white balls. Two balls are selected in succession without replacement. What is the probability that:

(i) the two balls are black?
(ii) the two balls are of the same colour?
(iii) the two balls are of different colour?

6.6 When three marksmen take part in a shooting contest their chances of hitting the target are ½, $1/3$ and ¼. Calculate the chance that one (and only one) bullet will hit the target if all the three men fire at it simultaneously.

6.7 The independent probabilities that three components of a television set will need replacing within a year are $\frac{1}{10}$, $\frac{1}{12}$, and $\frac{1}{15}$. Calculate the probability that:

(i) at least one component will need replacing.
(ii) one and only one component will need replacing.

6.8 A student's school bag contains 8 red books, 4 blue books, and 1 white book. He pulls out one book at random. Find the probability that he pulls out:

(i) a red book.
(ii) a blue book
(iii) a red or blue book.

6.9 One bag contains **4** white and **3** black balls, and a second bag contains 3 white balls and 5 black balls. One ball is drawn from the bag and is placed unseen in the second bag. What is the probability that a ball now drawn from the second bag is black?

6.10 A box contains 3 red, 2 green and 5 blue crayons. Two crayons are randomly selected from the box without replacement. Find the probability that:

(i) the crayons are of the same colour.
(ii) at least one red crayon is selected.

6.11 In an experiment, a box contains 2 green and 5 black balls. A second box contains 5 green and 3 blue balls. One ball is drawn at random from the second box and placed into the first box. What is the probability that a ball now drawn at random from the first box is green?

6.12 A box contains 36 beads, 28 of which are white and the rest yellow. One bead is picked from the box at random and not replaced. A second bead is then picked at random. Find the probability that:

(i) the first bead picked is white
(ii) the second bead picked is yellow
(iii) both beads picked are white.
(iv) both beads picked ate yellow.

6.13 A box P contains 4 white and 6 red balls and a box Q contains 6 white and 2 red balls. A box is chosen at random and a ball is drawn at random from it.

(i) Find the probability that the ball drawn is red.
(ii) Given that the ball selected is white, find the probability that the box P was selected.

6.14 Given that A and B are mutually exclusive events, such that P(A) = 0.5, P(A ∪ B) = 0.9, find

(i) $P(\bar{A} \cup B)$
(ii) $P(\bar{A} \cup \bar{B})$

6.15 If A and B are events and P(B) = $\frac{1}{6}$, P(A and B) = $\frac{1}{12}$, P(B/ A) = $\frac{1}{3}$. Calculate:

(i) P(A)
(ii) P(A / B)
(iii) P(A/$\bar{B}$), where $\bar{B}$ is the event "B does not occur".

6.16 Given that A and B are independent events in a sample space such that P(A ∩ B) = $\frac{2}{5}$ and P(A U B) = $\frac{4}{5}$. Find

(i) P(B)
(ii) $P(\bar{A} \cup \bar{B})$

6.17 A fair die is tossed 5 times. Calculate the probability that:

(i) a 3 or 5 appears on the first throw.
(ii) a 5 appears on each throw.
(iii) four 6's appear in the five throws.

6.18 The probability that three soldiers A, B and C hit a target when each of them fires once at the target is $\frac{1}{3}$,$\frac{1}{4}$, and $\frac{1}{5}$ respectively. The three soldiers together fire at a monkey each releasing one bullet. Determine:

(i) the probability that the monkey is hit by two bullets.
(ii) the probability that at least the monkey will be hit.

6.19 A bag contains ten Pepsi and eight Mirinda bottle tops. A bottle top is picked at random from the bag after the other without replacement. Find the probability that the first three bottle tops picked are of the same type.

6.20 A restaurant has four times as many male customers as female; 40% of men and 70% of women take the set lunch, the remainder choosing from the optional items of the menu. Of the men choosing the set lunch, 10% drink wine, 50% beer, and the remainder soft drinks, while of those choosing optional items the corresponding proportions are 60% and 30%. Among the women customers the corresponding proportions are 30% and 10% of those taking the set lunch, and 40% and 20% of those choosing optional items.

(i) What proportion of customers takes the set lunch?
(ii) What proportion of this drink wine?
(iii) What proportion of women customers drink beer?
(iv) What proportion of wine-drinking customers are men?

8.0
DISCRETE DISTRIBUTIONS

8.1 Discrete Distributions

Consider an experiment in which a die is tossed and the number **X** appears on the top face of the die. **X** can only take on the values 1, 2, 3, 4, 5, 6. In this case **X** is called a discrete random variable since it assumes only a finite number of values.
The small letter x is used to denote only one of the values of the random variable **X**. It is clear that the *discrete* random value **X** assumes one of its values with a certain probability. In the case of the die being tossed once, each of the values 1, 2, 3, 4, 5, 6 has a probability $\frac{1}{6}$ of occurring given that the die being tossed is a fair die.

For instance, when X = 4, then P(X = 4) = $\frac{1}{6}$

Making the table of the values of x and P(X = x) we obtain the following table.

TABLE 8.1 Discrete distribution

X	1	2	3	4	5	6
P(X = x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Table 8.1 above is called a *discrete probability distribution function* abbreviated as p.d.f.

Instead of giving the table we can also write out the information contained in the table in form of a formula. We can write

$$p(x) = \begin{cases} \frac{1}{6}, & x = 1, 2, 3, 4, 5, 6 \\ 0, & elsewhere \end{cases} \qquad (8.1)$$

We note that all the possibilities have been exhausted and hence the probabilities add up to 1.

By definition a table or a formula listing all the possible values that a discrete variable takes on, together with the associated probabilities is called a *discrete probability distribution*.

It is given by

$$\boldsymbol{p(x) = P(X = x)} \qquad (8.2)$$

The probability that the random variable **X** will take on a value between two specified outcomes ***a*** and ***b*** both values inclusive is given by

$$\boldsymbol{P(a \leq X \leq b) = \sum_{x=a}^{b} p(x)} \qquad (8.3)$$

In the above formula, if a = 2 and b = 5, then we have

$$\boldsymbol{P(2 \le X \le 5)} = \sum_{\mathbf{2}}^{\mathbf{5}} \boldsymbol{p(x)} = P(2) + P(3) + P(4) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \mathbf{\frac{2}{3}}$$

If both values 2 and 5 are not inclusive then we have

$$\boldsymbol{P(2 < X < 5)} = \sum_{\mathbf{3}}^{\mathbf{4}} \boldsymbol{p(x)} = P(3) + P(4) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \mathbf{\frac{1}{3}}$$

If only one of the two values is inclusive, e.g. 2 then we have

$$\boldsymbol{P(2 \le X < 5)} = \sum_{\mathbf{2}}^{\mathbf{4}} \boldsymbol{p(x)} = P(2) + P(3) + P(4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \mathbf{\frac{1}{2}}$$

Example 8.1

Two dice are tossed and the sum X of the numbers which appear on top is found. Find the probability distribution of X.

Solution

The sample space is the set of numbers

S = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13}

The probability distribution function is given in Table 8.2 below.

TABLE 8.2 Probability distribution function

x	2	3	4	5	6	7	8	9	10	11	12
P(X = x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

We plot a bar chart for the p.d.f. as follows:

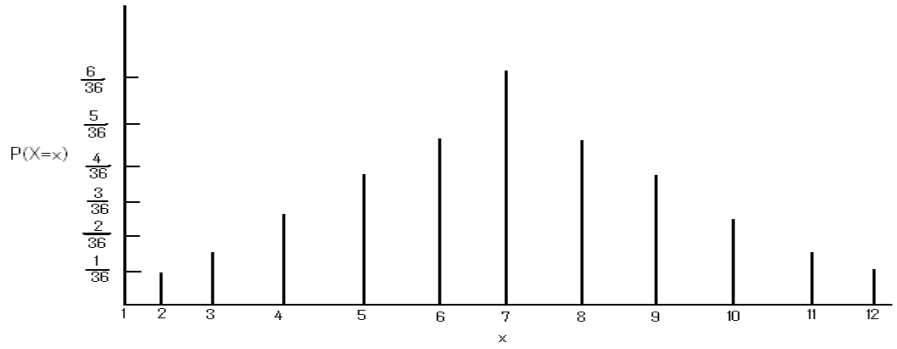


FIG.8.1 Bar chart for Table 8.2

We can plot a cumulative distribution using the table below.

TABLE 8.3: Cumulative distribution function

X	2	3	4	5	6	7	8	9	10	11	12
P(X = x)	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
F(X = x)	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

The bar chart for the cumulative distribution given in Table 8.3 is given below in Fig. 8.2

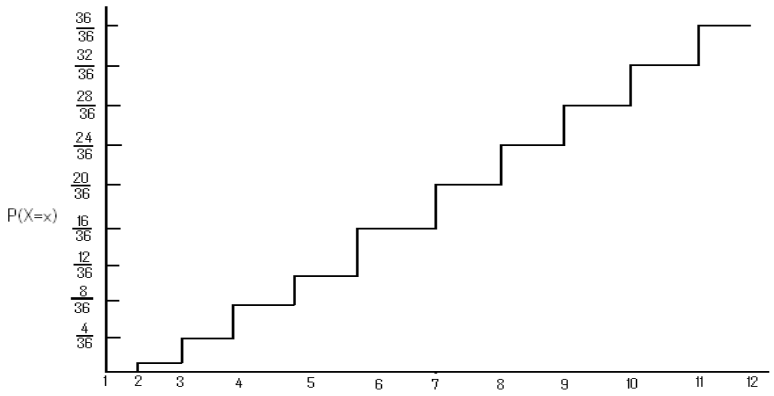


FIG.8.2 Bar chart for the cumulative distribution Table 8.3

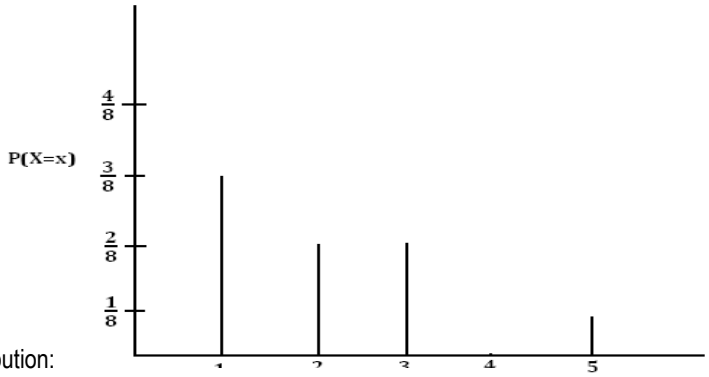
Example 8.2: Consider the discrete function

X	1	2	3	5
P(X = x)	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$
F(X = x)	$\frac{3}{8}$	$\frac{5}{8}$	$\frac{7}{8}$	$\frac{8}{8}$

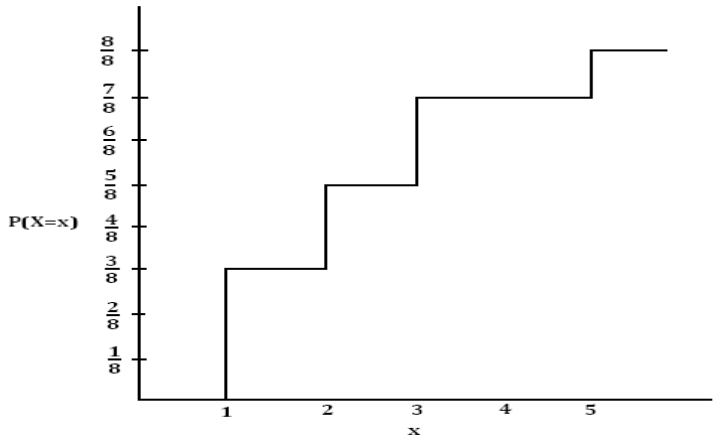
(i) Draw a bar chart for the function
(ii) Draw a cumulative distribution for the function
(iii) Find P(X < 3)

Solution

Bar chart:



The cumulative distribution:



$$P(X < 3) = P(X = 1) + P(X = 2) = \frac{3}{8} + \frac{1}{4} = \mathbf{\frac{5}{8}}$$

EXERCISE 7

7.1 A random variable X had the probability function:

$$f(x) = \begin{cases} k2^x, & x = 0,1,2,\dots,6 \\ 0, & elsewhere \end{cases}$$

Determine: (i) *the value of* k., *(ii) E(X) and (iii) P(x< 4/ x>1)*

7.2 A discrete random variable X has a probability function

$$f(x) = \begin{cases} \frac{x}{k}, & x = 1,2,3\dots,n \\ 0, & elsewhere \end{cases}$$

Where k is a constant. Given that the expectation of X is 3, find:

(i) the value of n and the constant k,
(ii) the median and the variance of X,
(iii) $P(X = 2/x \geq 2)$

7.3 The discrete rectangular distribution is defined by

$P(X = r) = \frac{1}{n}$, r = 0, 1, 2, …, n-1

Find the mean of the distribution and show that the variance is $\frac{n^2-1}{12}$.

7.4 In 4 fair tosses of a coin, let
X = number of heads.
Y = number of changes in sequence.
Tabulate the probability function of the random variable by first constructing sample space of the experimental outcomes.

7.5 Two boxes each contain 6 slips of paper numbered 1 to 6. Two slips of paper are drawn, one from each of the boxes. Let X be the difference between the numbers drawn (absolute value).
Tabulate the probability function of the random variable X.

7.6 A discrete random variable X has a probability mass function

$$f(x) = \begin{cases} \frac{a}{2x}, & x = 1,2,\dots,4 \\ 0, & elsewhere \end{cases}$$

Determine:

(i) the value of ***a.***
(ii) the probability of $p(x \leq 3)$.
(iii) the expectation of X.
(iv) the variance and standard deviation of X.

9.0 EXPECTATION

9.1 Introduction

Imagine that you have a chance (probability) of $\frac{1}{10}$ of winning a prize of Shs. 100,000. Your *mathematical expectation* is given by

the amount × probability (your chance)

That is,

100,000 $\times \frac{1}{10}$ = Shs. 10,000.

This means that mathematically you expect to get Shs. 10,000 only.

Thus if $p(x)$ is the probability that a person will receive a sum of money x, the *mathematical expectation* is defined as $xp(x)$. That is,

$$\boldsymbol{Expectation = xp(x)} \qquad (9.1)$$

The following two examples will clarify further the meaning of mathematical expectation.

Example 9.1

In a given business venture a person can make a profit of Shs. 300,000 with a probability of 0.6 or take a loss of Shs. 100,000 with a probability 0.4. Find his expectation.

Solution

Expectation = $xp(x)$

= (300,000 x 0.6) + (-100,000 x 0.4)

= 180,000 – 40,000

= Shs. 140,000

Example 9.2

A box contains 12 cards. Two cards are labelled 1, three cards are labelled 2, four cards are labelled 3, two cards are labelled 4, and one card is labelled 5. A person draws one card at random. Determine his expectation.

Solution

We draw a table showing the probabilities of drawing each card. We then calculate the quantity $xp(x)$for each value of x and include this as well in the table.

x	1	2	3	4	5
$p(x)$	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{4}{12}$	$\frac{2}{12}$	$\frac{1}{12}$
$xp(x)$	$\frac{2}{12}$	$\frac{6}{12}$	$\frac{12}{12}$	$\frac{8}{12}$	$\frac{5}{12}$

The total expectation is given by

$$\sum xp(x) = \frac{2}{12} + \frac{6}{12} + \frac{12}{12} + \frac{8}{12} + \frac{5}{12} = \frac{33}{12} = \mathbf{2.27}$$

On the other hand the probabilities can be replaced by the relative frequencies. Then we can form a table as follows:

x	f	fx
1	2	2
2	3	6
3	4	12
4	2	8
5	1	5
	∑f = 12	∑fx = 33

From which we obtain

$$\frac{\sum fx}{\sum f} = \frac{33}{12} = \mathbf{2.75}$$

We note that the quantity $\sum fx$ is identical to the quantity$xp(x)$. But we already know that the quantity $\frac{\sum fx}{\sum f}$ is the mean, μ. This means that if the probabilities are replaced by the relative frequencies, then the expectation reduces to the arithmetic mean. We can interpret this by saying that expectation represents the mean of the population from which the sample is drawn.

9.2 Mathematical Expectation

Definition: Let X be a discrete random variable with the following probability distribution

x	x_1	x_2	x_3			x_n
P(X = x)	$p(x_1)$	$p(x_2)$	$p(x_3)$			$p(x_n)$

The expected value of X or the mathematical expectation of X is

$$\boldsymbol{x_1p(x_1) + x_2p(x_2) + x_3p(x_3) + \cdots + x_np(x_n) = \sum_{i=1}^{n} x_ip(x_i)} \tag{9.2}$$

Equation (9.2) is a more general expression of Equation (9.1).

NOTE: The mathematical expectation E(X) of a random variable X is also its mean μ.

That is, we can also write:

$$\boldsymbol{Mean = \mu = E(X) = \sum_{x} xp(x)} \tag{9.3}$$

Example 9.3
A die is tossed once. Let *X* be the number which appears on top. Determine the expectation of *X*.

Solution

Each of the numbers 1, 2, 3, 5 and 6 occurs with a probability of $\frac{1}{6}$. Hence we have the following probability distribution.

x	1	2	3	4	5	6
$p(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The mathematical expectation of *X* is therefore

$$\boldsymbol{E(X)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{6}} \boldsymbol{x_i p(x_i)} \qquad = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right)$$

$$= \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6}$$

$$= \frac{21}{6} = \mathbf{3.5}$$

This means that on average when a person tosses a die very many times, his mathematical expectation is 3.5, which is exactly the mean of the numbers which occur.

Example 9.4

Three coins are tossed once. Let ***X*** denote the possible number of heads. Determine the expectation of ***X***.

Solution

The possible number of heads is the set {1, 2, 3}. The pdf. is therefore

x	0	1	2	3
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The mathematical expectation of ***X*** is therefore

$$E(X) = \sum_{x} xp(x) \; = 0\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) \; = 0 + \frac{3}{6} + \frac{6}{8} + \frac{3}{8} = = \frac{12}{8} = \mathbf{1.5}$$

9.3 Laws of Expectation

The following rules hold for calculations dealing with expectation:

1. $The\ expectation\ of\ x\ is\ E(X) \quad = \sum xp(x)$
2. $The\ expectation\ of\ x^2\ is\ E(X^2) = \sum x^2 p(x)$
3. $The\ expectation\ of\ x^3\ is\ E(X^3) = \sum x^3 p(x)$
4. $If\ a\ is\ a\ constant, then\ E(aX) \quad = aE(X)$
5. $If\ a\ and\ b\ are\ constants, then\ E(aX + b) \; = aE(X) + E(b)$
6. $If\ b\ is\ a\ constant, then\ E(b) \; = b$
7. $If\ X\ and\ Y\ are\ two\ independent\ random\ variables, \quad then \quad E(XY) \; = E(X)E(Y)$
8. $The\ expectation\ of\ f(x) = E[f(x)] = \sum f(x)p(x)$

Note: Rules 1, 2, 3 are particular cases of Rule 8.

9. $E(X^2 + X) = \sum (x^2 + x)px$

$$= \sum x^2 p(x) + \sum xp(x)$$

Example 9.5

Let X be a random variable with the following probability distribution

x	-2	1	4
$p(X = x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

Find

(i) E(X)
(ii) $E(X^2)$
(iii) E(2X – I)

Solution

$(i) \quad E(X) = \sum xp(x)$

$$= -2\left(\frac{1}{6}\right) + 1\left(\frac{1}{2}\right) + 4\left(\frac{1}{3}\right)$$

$$= \frac{-2}{6} + \frac{3}{6} + \frac{8}{6} = \frac{9}{6} = \mathbf{1.5}$$

$(ii) \quad E(X^2) = \sum x^2 p(x)$

$$= 2^2\left(\frac{1}{6}\right) + 1^2\left(\frac{1}{2}\right) + 4^2\left(\frac{1}{3}\right)$$

$$= 4\left(\frac{1}{6}\right) + 1\left(\frac{1}{2}\right) + 16\left(\frac{1}{3}\right)$$

$$= \frac{4}{6} + \frac{3}{6} + \frac{32}{6} = \frac{36}{6} = \mathbf{6.5}$$

$(iii) \quad E(2X - 1) = 2E(X) - E(1)$

$$= 2(1.5) - 1 = 3 - 1 = \mathbf{2}$$

Alternatively;

$$E(2X - 1) = \sum (2x - 1)p(x)$$

$$= (-4 - 1)\left(\frac{1}{6}\right) + (2 - 1)\left(\frac{1}{2}\right) + (8 - 1)\left(\frac{1}{3}\right)$$

$$= \frac{\mathbf{-5}}{\mathbf{6}} + \frac{\mathbf{3}}{\mathbf{6}} + \frac{\mathbf{14}}{\mathbf{6}} = \frac{\mathbf{12}}{\mathbf{6}} = \mathbf{2.0}$$

Example 9.6

In a consignment of 12 bulbs a person finds 4 defectives. Another person makes a random purchase of 6 bulbs. If *X* is the number of defective bulbs purchased, find *E(X)*.

Solution

$$P(defective) = \frac{4}{12} = \frac{1}{3}$$

Therefore;

$$E(X) = xp(x) = 6\left(\frac{1}{3}\right) = 2$$

Variance of a Random Variable

The variance of a discrete random variable X is defined by

$$Var(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

$$= \sum x^2 p(x) - [\sum xp(x)]^2 \qquad (9.4)$$

$$= \sum \boldsymbol{x^2 p(x) - \mu^2}$$

Example 9.7

Let *X* be a random variable with the following probability distribution

x	- 2	1	4
P(X = x)	0.2	0.5	0.3

Find the standard deviation of *X*.

Solution

From Equation (9.4) we see that we have to calculate the mean, μ, first.

μ = ∑xp(x) = - 2(0.2) + 1(0.5) + 4(0.3)
= -0.4 + 0.5 + 1.2
= 1.7 – 0.4
= 1.3

Hence:
σ^2 = 4(0.20 + 1(0.5) + 16(0.3) – 1.3^2
= 0.8 + 0.5 + 4.8 – 1.3^2
=6.1 – 1.69
= 4.41

The standard deviation is obtained by taking the square root of $\boldsymbol{\sigma^2}$, thus

$$\sqrt{\sigma^2} = \sqrt{4.41} = \mathbf{2.1}$$

EXERCISES 8

8.1 A discrete random variable X has the following probability:

$\boldsymbol{x}$	1	2	3	4	5
P($\boldsymbol{X = x}$)	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{5}{15}$

Calculate the variance of the distribution.

8.2 The number of bids submitted by Mukisa Construction Co. prior to winning a competitive government contract from the small-scale business administration in Uganda is given in the following distribution.

X	1	2	3	4	5
P(X=x)	0.1	0.2	0.4	0.2	0.1

Compute

(i) The mean:
(ii) The Variance
(iii) The standard deviation

8.3X is a random variable that denotes the sum of two numbers that show up when two dice are rolled. Copy and complete the following distribution table.

X	2	3	4	5	6	7	8	9	10	11	12
p(x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$					$\frac{4}{36}$			$\frac{1}{36}$

Find:

(i) the expected value of X.
(ii) the variance of X.

8.4 The table below shows the probability distribution of a random variable X.

$\boldsymbol{x}$	- 3	0	2	3
P($\mathbf{X = x}$)	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

Given that Y and Z are the random variables defined by $Y = X^2$, $Z = X(X - 1)$

(i) draw a table to show the probability distribution of Y and Z.
(ii) determine E(Y) and E(Z).
(iii) show that $E(Z) = E(X^2) - E(X)$.

8.5 Let X be a random variable with the following probability distribution

$\boldsymbol{x}$	- 3	6	9
P(X = x)	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

(i) Find $E(X)$ and $E(X^2)$
(ii) Evaluate $E(2X - 1)$.

8.6 A shipment of 6 television sets contains 2 defects. A hotel makes a random purchase of 3 of the sets. If X is the number of defective sets purchased by the hotel, find E(X).

8.7 Find the expected number of jazz records when 4 records are selected at random from a collection consisting of 5 jazz records, 2 classical records and 3 polka records.

8.8 Let X be a random variable with the following probability distribution:

x	- 2	3	5
P(X = x)	0.3	0.2	0.5

Find the standard deviation of X.

8.9 In a gambling game a man is paid \$2 if he draws a jack or queen and \$5 if he draws a king or ace from an ordinary deck of 52 playing cards. If he draws any other card, he loses. How much should he pay if the game is fair?

8.10 A race-car driver wishes to insure his car for the racing season for \$10,000. The insurance company estimates a total loss may occur with probability 0.002, a 50% loss with probability 0.01, and a 25% loss with probability 0.1. Ignoring all other partial losses, what premium should the insurance company charge each season to make a profit of \$10?

10.0
CONTINUOUS DISTRIBUTIONS

10.1 Introduction

The expression "*x* is a continuous variable in the interval (2,5) or in the interval (a,b)" means that x can take on any and all values from the value x = 2 to x = 5 or from x = a to x = b, where the more positive limit is written second. The interval could be (-3, 10), (-6, -2), (0, ∞), (-∞, ∞) and so on. This is in contrast to the discrete variable which only takes on values here and there, e.g. x = 1, 2, 3, …, ∞. However this still remains a discrete distribution because the values between 1 and 2 or between 2 and 3 e.t.c. are not included.

10.2 The Uniform Continuous Distribution

Let a and b be two given real numbers such that a < b, and consider an experiment in which a point *X* is selected from the interval ***S*** = {x : a ≤ x ≤b} in such a way that the probability that *X* will belong to any subinterval of S is proportional to the length of the subinterval. This distribution of the random variable *X* is called the uniform distribution on the interval (a, b). Since this is a probability distribution function, then we must have

$$\int_{a}^{b} \boldsymbol{f(x)dx} = \mathbf{1} \qquad (10.1)$$

The constant value of *f(x)* throughout S must be $\frac{1}{(b-a)}$, and the p.d.f. of *X* must be as follows:

$$f(x) = \begin{cases} \dfrac{1}{(b-a)}, & for a \le x \le b \\ \\ 0, & elsewhere \end{cases} \qquad (10.2)$$

We sketch the *p.d.f* as below.

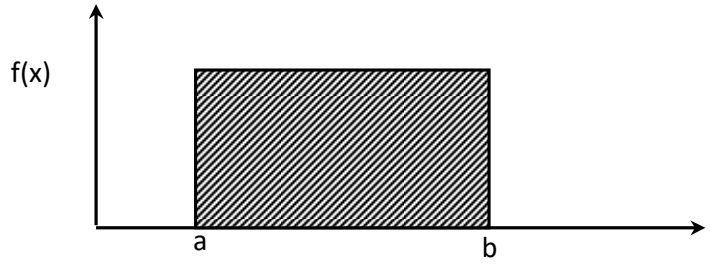


Fig 10.1

Using a simple analogy the area which is shaded is

A = *f(x)* . (b – a) = 1, for a *p.d.f.*

Which implies that

$$\boldsymbol{f(x) = \frac{1}{b-a}}$$

As in Equation (10.2).

Example 10.1

A variable ***X*** is uniformly distributed in the interval (2,5). Find P(2 ≤ X ≤ 3).

Solution

The probability density curve is shown in the diagram below:

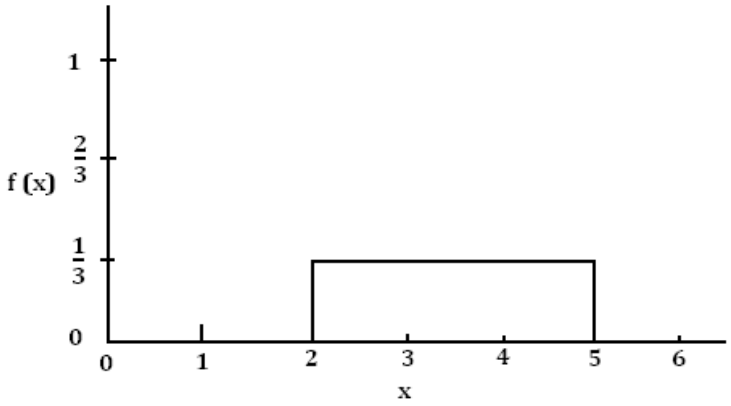


Fig 10.1

Since the distribution is uniform, the "curve" in figure.10.2 is a "horizontal" straight line whose equation, using Equation (10.2) is given by

$$f(x) = \begin{cases} \frac{1}{3}, & 2 \le x \le 5 \\ 0, & elsewhere \end{cases}$$

P(2 ≤ X ≤ 3) is shown shaded in Fig.10.3.

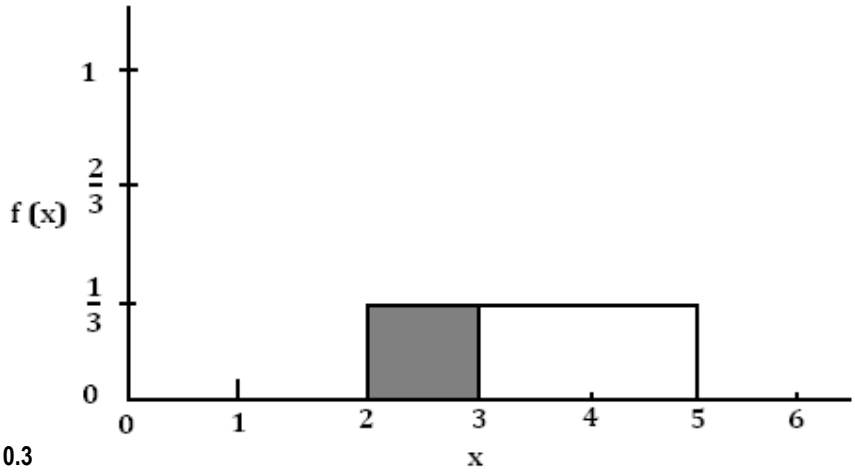


FIG. 10.3

$$P(2 \le x \le 3) = \int_2^3 f(x)dx$$

$$= \int_2^3 \frac{1}{3}dx = \left[\frac{1}{3}\right]_2^3 = \mathbf{1}$$

It is also equal to the shaded area, i.e.

$$P(2 \le X \le 3) = \frac{1}{3} \times 1 = \mathbf{1}$$

The total area under the curve is 1 and this represents the total probability which must be equal to 1.

Definition:
A *p.d.f* is constructed in such a way that the area under the curve bounded by the x-axis is 1.

The area is represented by the integral:

$$\int_{-\infty}^{\infty} \boldsymbol{f(x)dx = 1} \qquad (10.3)$$

Example 10.3
A continuous random variable X has a probability density function defined by

$$f(x) = \begin{cases} kx(4-x), & 0 \le x \le 4 \\ 0, & elsewhere \end{cases}$$

(i) Find k
(ii) Determine P(2 ≤ X ≤ 3)

Solution

By definition

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$Hence \int_0^4 kx(4-x)dx = 1$$
$$k\int_0^4 (4x - x^2)dx = 1$$
$$k\left[2x^x - \frac{x^3}{3}\right]_0^4 = 1$$
$$k\left(32 - \frac{64}{3}\right) = 1, \ \ from\ which\ k = \boldsymbol{\frac{3}{32}}$$

Hence the *p.d.f* is;

$$f(x) = \begin{cases} \frac{3}{32}x(4-x), & 0 \le x \le 4 \\ 0, & elsewhere \end{cases}$$

(ii)

$$P(2 \le X \le 3) = \int_2^3 \frac{3}{32}x(4-x)dx$$

$$= \frac{3}{32}\left[2x^2 - \frac{x^3}{3}\right]_2^3$$

$$= \frac{3}{32}\left[(18-9) - \left(8 - \frac{8}{3}\right)\right]$$

$$= \frac{3}{32}\left[\left(9 - \frac{16}{3}\right)\right] \;=\; \frac{3}{32}\left(\frac{27-16}{3}\right)$$

$$= \frac{3}{32} \times \frac{11}{3} = \frac{11}{32} = \mathbf{0.34}$$

10.3 Expectation of a continuous Random Variable

The expectation of a continuous random variable X is defined by

$$\boldsymbol{E(x) = \int_{-\infty}^{\infty} xf(x)dx = 1} \qquad (10.4)$$

= μ = MEAN

Example 10.4

A random variable X has a probability density function given by

$$f(x) = \begin{cases} kx(2-x), & 0 \le x \le 2 \\ 0, & elsewhere \end{cases}$$

Find *(i) the constant k*
(ii) the expectation (mean) of k.

Solution

By definition

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$Hence \int_0^2 k(2x - x^2)dx = 1$$

$$k\int_0^2 (2x - x^2)dx = 1$$

$$k\left[x^2 - \frac{x^3}{3}\right]_0^2 = 1$$

$$k\left(4 - \frac{8}{3}\right) = 1, \;\; from\ which\ k = \mathbf{\frac{3}{4}}$$

By definition

$$E(x) = \mu = \int_{-\infty}^{\infty} xf(x)dx = 1$$

$$= \int_0^2 \frac{3}{4}x(2-x)dx = \frac{3}{4}\int_0^2 (2x^2 - x^3)dx$$

$$= \frac{3}{4}\left[\frac{2x^3}{3} - \frac{x^4}{4}\right]_0^2 = \frac{3}{4}\left(\frac{16}{3} - \frac{16}{4}\right) = \frac{3}{4} \times \frac{4}{3} = \mathbf{1}$$

Example 10.5

Find the expectation of the *p.d.f* in Example 10.2

Solution

The *p.d.f*is given by

$$f(x) = \begin{cases} \frac{2}{12}, & 2 \le x \le 4 \\ \frac{3}{12}, & 4 \le x \le 6 \\ \frac{1}{12}, & 6 \le x \le 8 \\ 0, & elsewhere \end{cases}$$

$$E(X) = \int_2^4 \frac{2}{12} x dx + \int_4^6 \frac{3}{12} x dx + \int_6^8 \frac{1}{12} x dx$$

$$= \left[\frac{2x^2}{2 \times 12}\right]_2^4 + \left[\frac{3x^2}{2 \times 12}\right]_4^6 + \left[\frac{x^2}{2 \times 12}\right]_6^8$$

$$= \frac{1}{12}[(32 - 8) + (108 - 48) + (64 - 36)]$$

$$= \frac{1}{12}(24 + 60 + 28)$$

$$= \frac{112}{24} = \mathbf{4.67}$$

10.4 Variance of a Continuous Random Variable

The variance of a continuous random variable is defined by

Var(X) = σ^2 = E(X^2) – [E(X)]^2

$$= \int_{-\infty}^{\infty} \boldsymbol{x^2 f(x) dx} - [\boldsymbol{x f(x) dx}]^2$$

$$= \int_{-\infty}^{\infty} \boldsymbol{x^2 f(x) dx} - \mu^2 \ (10.5)$$

The Standard Deviation is then given by

$$\boldsymbol{S.D} = \sqrt{\boldsymbol{\sigma^2}} = \sqrt{\boldsymbol{Var(x)}} (10.6)$$

Example 10.6

A continuous random variable is given by

$$f(x) = \begin{cases} kx, & 0 \le x \le 3 \\ 0, & elsewhere \end{cases}$$

Find *(i) the constant k*
(ii) the mean of X
(iii) the variance of X

Solution

$$\int_2^3 kxdx = 1$$

$$k\left[\frac{x^2}{2}\right] = 1$$

$$k\left[\frac{9}{2}\right] = 1,$$
$$k = \mathbf{\frac{2}{9}}$$

Thus the *p.d.f* is

$$f(x) = \begin{cases} \frac{2}{9}x, & for\ 0 \le x \le 3 \\ 0, & elsewhere \end{cases}$$

(ii)

$$\mu = \int_0^3 x.\frac{2}{9}xdx$$

$$= \frac{2}{9}\left[\frac{x^3}{3}\right]_0^3 = \frac{2}{9} \times \frac{27}{3} = \mathbf{2}$$

(iii)

$$Var(X) = \int_0^3 x^2.\frac{2}{9}xdx - \mu^2$$

$$= \frac{2}{9}\left[\frac{x^4}{4}\right]_0^3 - 2^2$$

$$= \frac{2}{9}\left[\frac{81}{4}\right] - 4 \quad = \frac{18}{4} - 4 = \mathbf{\frac{1}{2}}$$

10.5 Median of a Continuous Random Variable

For a continuous random variable *X*, the median is defined by

$$= \int_{-\infty}^{\boldsymbol{m}} \boldsymbol{f(x)dx} = \frac{\mathbf{1}}{\mathbf{2}} (10.7)$$

Where $\boldsymbol{m}$ represents the median.

Example 10.7

A continuous random variable is given by

$$f(x) = \begin{cases} kx, & 0 \leq x \leq 1 \\ 0, & elsewhere \end{cases}$$

Find the median.

Solution

First determine k

$$\int_0^1 kxdx = k\left[\frac{x^2}{2}\right]_0^1 = k.\frac{1}{2} = \frac{k}{2} = \mathbf{1}$$

Hence ***k = 2***

Therefore the *p.d.f* is

$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & elsewhere \end{cases}$$

By definition the medium m is given by

$$\int_{-\infty}^{m} f(x)dx = \frac{1}{2}$$

Hence

$$\int_{-\infty}^{m} 2xdx = \left[\frac{x^2}{2}\right]_0^m = m^2 = \frac{1}{2}$$

$$Therefore, \qquad \boldsymbol{m = \frac{1}{\sqrt{2}}} = Median.$$

Example 10.8

Find the median of the *p.d.f* in Example **10.4**

Solution

The p.d.f is;

$$f(x) = \begin{cases} \frac{3}{4}x(2-x), & 0 \le x \le 2 \\ 0, & elsewhere \end{cases}$$

The median, *m*, is given by.

$$\int_0^m \frac{3}{4}x(2-x)dx = \frac{1}{2}$$

$$\frac{3}{4}\int_0^m (2x - x^2)dx = \frac{1}{2}\left[x^2 - \frac{x^3}{3}\right]_0^m$$

$$= \frac{3}{4}\left(m^2 - \frac{m^2}{3}\right) = \frac{3}{4}\left(\frac{3m^2 - m^3}{3}\right) = \frac{1}{2}$$

$$= 3m^2 - m^2 = 2$$

$$or \quad m^3 - 3m^2 + 2 = 0$$

This is cubic equation which can be solved by the usual methods. However, by inspection we find that the median, $m = 1$.

10.6 Quartiles of a Continuous Random Variable

The **UPPER QUARTILE**, $\boldsymbol{Q_3}$ is defined by

$$= \int_{-\infty}^{\boldsymbol{Q_3}} \boldsymbol{f(x)dx} = \frac{\mathbf{3}}{\mathbf{4}} (10.8)$$

The **LOWER QUARTILE**, $\boldsymbol{Q_1}$ is defined by

$$= \int_{-\infty}^{\boldsymbol{Q_1}} \boldsymbol{f(x)dx} = \frac{\mathbf{1}}{\mathbf{4}} (10.9)$$

Example 10.9

Find the upper and lower quartiles for the *p.d.f.* in Example10.4

The Upper Quartile

$$= \int_0^{Q_3} 2x dx = \frac{3}{4}$$

$$= [x^2]_0^{Q_3} = \frac{3}{4}$$

$$Q_3^2 = \frac{3}{4}, from\ which\ \boldsymbol{Q_3} = \frac{\sqrt{\mathbf{3}}}{\mathbf{2}}$$

The lower Quartile

$$= \int_0^{Q_1} 2x dx = \frac{1}{4}$$

$$= [x^2]_0^{Q_1} = \frac{1}{4}$$

$$Q_1^2 = \frac{1}{4}, from\ which\ \boldsymbol{Q_1} = \frac{\mathbf{1}}{\mathbf{2}}$$

10.7 The Cumulative Distribution

For a continuous random variable X the cumulative distribution function F(X) is defined by

$$\boldsymbol{F(x)} = \int_{-\infty}^{\boldsymbol{x}} \boldsymbol{f(t)dt}\ (10.10)$$

Example 10.10

A continuous random variable is given by

$$f(x) = \begin{cases} kx(1-x), & 0 \le x \le 1 \\ 0, & elsewhere \end{cases}$$

Determine (i) k
(ii) Find F(x)

Solution

(i)

$$= \int_0^1 kx(1-x)dx$$

$$= k\int_0^1 (x - x^2)dx = k\left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^1$$

$$= k\left(\frac{1}{2} - \frac{1}{3}\right) = \frac{k}{6} = 1$$

$$= \boldsymbol{k = 6}$$

Hence;

$$f(x) = \begin{cases} 6x(1-x), & 0 \le x \le 1 \\ 0, & elsewhere \end{cases}$$

(ii) The cumulative distribution F(x) is therefore given by;

$$F(X) = \int_{-\infty}^{x} f(t)dt$$

$$= \int_{0}^{x} 6t(1-t)dt = \int_{0}^{x} (6t - 6t^2)dt$$

$$= [3t^2 - 2t^3]_0^x = 3x^2 - 2x^3$$

Thus

$$\mathrm{F(x)} = \begin{cases} 0, & x < 1 \\ 3x^2 - 2x^3, & elsewhere \\ 1\,, & x \ge 1 \end{cases}$$

10.8 Mode of a Continuous Random Variable

For a function f(x) of a continuous random variable **x**, the mode of the function is such that

$$\boldsymbol{f'(x) = 0}$$
$$\boldsymbol{f''(x) < 0}\,, where exist the mode \tag{10.11}$$

Example 10.11

For

$$f(x) = \begin{cases} \frac{3}{5} - \frac{6}{5}x - \frac{3}{5}x^2, & 0 \le x \le 1 \\ 0, & elsewhere \end{cases}$$

Then

$$f'(x) = -\frac{6}{5} - \frac{6}{5}x = 0, \qquad from\ which, x = -1$$

$$f''(x) = -\frac{6}{5} < 0,$$

$$Hence\ the\ mode\ is\ x = -1$$

EXERCISE 9

9.1. The probability density function of a random variable X is given by

$$f(x) = \begin{cases} ksinx, & 0 \leq x \leq \pi \\ 0, & elsewhere \end{cases}$$

Determine:

(i) the value of k
(ii) $Pr\left(X > \frac{\pi}{3}\right)$
(iii) The median value of X

9.2. A random variable $\boldsymbol{x}$ has the probability density function

$$f(x) = \begin{cases} kx, & 0 < x < 1 \\ \left(\frac{k}{2}\right)x \quad , & 1 \leq x \leq 2 \\ 0\,, & elsewhere \end{cases}$$

Find:

(i) the value of the constant k
(ii) E(x)
(iii) The median of x
(iv) Pr(x< 1.5 : 1 ≤ x ≤ 2)
(v) Distribution function of x, and sketch it.

9.3. The outputs of 9 machines in a factory are independent random variables each with probability density function given by

$$f(x) = \begin{cases} ax, & 0 < x < 10 \\ a(20-x), & 1 \leq x \leq 20 \\ 0\,, & elsewhere \end{cases}$$

Find:

(i) the value of a
(ii) the expected value and variance of the output of each machine.
(iii) Hence or otherwise find the expected value and variance of the total output from all machines.

9.4. A continuous random variable X has the distribution function

$$f(x) = \begin{cases} 3kx\left(1 - \frac{x^2}{3}\right), & 0 \leq x \leq 1 \\ 1, & x \geq 1 \end{cases}$$

Determine:

(i) the value of k,
(ii) the probability density of X,
(iii) the median of X,
(iv) P(X> 0.5 / 0.25 ≤ X ≤ 1)

9.5. A random variable X has a probability density function given by

$$f(x) = \begin{cases} cx(2-x), & 0 \le x \le 2 \\ 0, & elsewhere \end{cases}$$

Find:

*(i) the constant **c**,*
(ii) the expectation of X,
(iii) the median of X.

11.0 BINOMIAL & POISSON DISTRIBUTIONS

11.0 PERMUTATIONS AND COMBINATIONS

Before we discuss Binomial distributions, its of importance to understand the in-depth knowledge on permutations and combinations.
Permutations and combinations are terminologies used to refer to the possible number of ways in which objects can be selected from larger sets or can be arranged among themselves.

PERMUTATIONS

A permutation is a combination arranged in a particular way.
Permutations are sometimes referred to as selections. Permutations refer to situations where all orderings of the objects to be considered are in some way different from each other i.e. they are distinguishable.

COMBINATIONS

A combination is a selection of distinct items.
They are sometimes called arrangements. They refer to situations where different orderings of the same set of objects are indistinguishable.

Illustration 1

Consider a set of three items X, Y and Z.

XY is one selection and is referred to as a *combination* of two items; XZ and YZ being the other possible combinations.
On the other hand, XYZ is a combination of three items of which there are six separate *permutations* that is XYZ, XZY, YZX, ZXY and ZYX.

Point to note:

Re-arranging items within a combination does not result into a different combination. However, re-arranging the items within a permutation gives a different permutation i.e. XYZ is the same combination of three letters X, Y and Z, while XYZ, XZY and ZYX are three different permutations of the three letters XYZ.

The factorial notation

The factorial notation is commonly used in permutations and combinations and consequently in binomial probability.
The symbol $n!$ is read as 'n *factorial*' and means n multiplied successively by every integer smaller than itself down to 1. For example;

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

Therefore we can write in general terms that;

$$n! = n \times (n-1) \times (n-2) \times \dots\dots\dots\dots \times 2 \times 1 \quad (11.1)$$

The Combination Formula

The number of different combination of *r* items from *n* distinct items is written as ${}^{n}C_{r}$ and is calculated from;

$${}^{n}C_{r} = \frac{n!}{r! \times (n-r)!} \quad (11.2)$$

The Permutations Formula

The number of different permutations of *r* items from *n* distinct items is written as

$$^{n}P_{r} = \frac{n!}{(n-r)!} \qquad (11.3)$$

Example 11.0:

Joseph the internal Auditor of Pepsi Uganda Ltd hopes to complete his Audit on four sets of accounts of a particular size within two months. He has been presented with nine such sets of accounts.

Required; Determine

(a) How many different sequences of four might he deal with in the first period?
(b) How many different sets of accounts might he deal with in the first period?

Solutions;
This part requires the number of sequences, which implies we shall need to compute the different number of Account permutations available.

Therefore, from the question; **n = 9** while **r = 4**;

Hence; from the permutations formula (Formula 11.3) we shall have

$$^{9}P_{4} = \frac{9!}{(9-4)!} = \frac{9\times8\times7\times6\times5\times4\times3\times2\times1}{5\times4\times3\times2\times1}$$

$$= 9\times8\times7\times6 \qquad = \textbf{3,024 sequences}$$

This part requires the different number of sets of accounts, which implies that we shall need to compute the different number of Combinations available.

Hence; from the combinations formula (Formula 11.2) we shall have

$$^{9}C_{4} = \frac{9!}{4!\,(9-4)!} = \frac{9\times8\times7\times6\times5\times4\times3\times2\times1}{(4\times3\times2\times1)\times(5\times4\times3\times2\times1)}$$

$$= \frac{9\times8\times7\times6}{4\times3\times2\times1} = \boldsymbol{126\ sets\ of\ accounts}$$

Pascal's Triangle

For cases where the total number of n of items is small Pascal's triangle provides an alternative method for finding numbers of combinations. It is a very simple method since it involves no arithmetic beyond addition of pairs of numbers.
However it's not recommended if the total number n of items exceeds 10, unless several $^{n}C_{r}$ figures for the same value of n are required.

11.1: The Binomial Distribution

An experiment consisting of repeated trials that have only two outcomes; a "success" and a "failure" denoted by S and F is referred to as a *Binomial* experiment or a *Bernoulli* trial. The probability of success is denoted by **p** and of failure by **q** and are such that p + q = 1 or q = 1 – p.

11.1: Properties of the Binomial Distribution

(i) The experiment consists of x repeated trials.
(ii) Each trial results in an outcome that is either a *Success* or a *Failure*.
(iii) The probability of a success p remains constant from the trial and of failure becomes **1 – p**.
(iv) The repeated trials are independent.

The number of successes **X** in **n** trials of a binomial experiment is a binomial random variable. The probability distribution of this random variable is called a *binomial distribution*. A binomial distribution therefore is the probability of a random variable X, the number of successes in **n** independent trials.

The probability of **X** successes in **n** independent trials denoted by **b(x;n, p)** with probability of successes p and of failure **q = 1 – p** is given by

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x} \qquad (11.4)$$

Where

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

Example 11.1

If an experiment is to toss a coin say 6 times, each trial results in a Head or a Tail, then its a *binomial experiment*.
If we say that a Head is a success then P(S) = ½. If we say that a Tail is a failure then P(F) = 1 – ½ = ½.
Then, what would be the probability of getting exactly 2 Heads in 6 tosses of a fair coin?

Solution

n = 6 is the number of trials
x = 2 is the number of successes
p = q = ½; and therefore we have

$$b(2; 6, ½) = \binom{6}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{6-2}$$

$$= \frac{6!}{2!\,4!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^4$$

$$= \frac{\mathbf{5.4.3.2.1}}{(\mathbf{2.1}).(\mathbf{4.3.2.1})}\left(\frac{\mathbf{1}}{\mathbf{2}}\right)^{(\mathbf{4+2})} = \frac{\mathbf{15}}{\mathbf{64}} = \mathbf{0.234}$$

Example 11.2

Find the probability of obtaining exactly three 2's if an ordinary die is tossed 5 times.

Solution

n = 5 is the number of trials
x = 3 is the number of successes
$p(2) = \frac{1}{6}$; therefore we have

$$b\left(3; 5, \frac{1}{6}\right) = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{5-3}$$

$$= \frac{5!}{3!\,2!}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = \frac{5.4.3.2.1}{3.2.1.2.1}.\frac{5.5}{6^5}$$

$$= \frac{5^3}{6^5}.2 = \frac{125 \times 2}{7776} = \mathbf{0.032}$$

Example 11.3
Find the probability that in a family of 4 children there will be

(a) exactly one boy,
(b) exactly 2 boys,
(c) at least one boy,

(Assume that the probability of a male birth is ½).

Solution
(a)
n = 4 is the number of trials
x = 1 is the number of successes
p = q = ½; and therefore we have

$$b(1;4,½) = \binom{4}{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{4-1}$$

$$= \frac{4!}{1!\,3!}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{3}$$

$$= \frac{4.3.2.1}{2.2.2.2.3.2.1}$$

$$= \mathbf{\frac{1}{4}}$$

(b)
n = 4 is the number of trials
x = 2 is the number of successes
p = q = ½; therefore we have

$$b(2;4,½) = \binom{4}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{4-2}$$

$$= \frac{4!}{2!\,2!}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{2}$$

$$= \frac{4.3.2.1}{2.1.2.1}.\frac{1}{2.2}.\frac{1}{2.2} = \mathbf{\frac{3}{8}}$$

(c)

P(at least one boy)

= P(1 boy) + P(2 boys) + P(3 boys) + P(4 boys)

= 1 – P(no boy)

$$= \binom{4}{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{3} + \binom{4}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{2} + \binom{4}{3}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{1} + \binom{4}{4}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{0}$$

$$= \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \mathbf{\frac{15}{16}}$$

11.3: Mean of the Binomial Distribution

The mean of the binomial distribution μ is defined by

$$\boldsymbol{\mu = np} \qquad (11.5)$$

11.4: Variance of the Binomial Distribution

The variance of the binomial distribution is defined by

$$\boldsymbol{\sigma^2 = npq} \tag{11.6}$$

The standard deviation is given by

$$\boldsymbol{S.D = \sqrt{npq}} \tag{11.7}$$

Example 11.4

If the probability of a defective bolt is 0.1.find (a) the mean, (b) the standard deviation for the distribution of defective bolts in a total of 400.

Solution

Mean = npq = 400 x 0.1 = 40.
That is, you would expect 40 bolts to be defective.

Variance = npq = 400 x 0.1 x 0.9 = 36

Hence the standard deviation

= √36 = 6

11.5: THE POISSON DISTRIBUTION

The discrete probability distribution

$$\boldsymbol{P(x;\mu) = \frac{\mu^x e^{-\mu}}{x!}} (x = 0,1,2 \ldots) \tag{11.8}$$

Where ***e*** = *2.71828…* and $\boldsymbol{\mu}$ is a given constant, is called the *Poisson Distribution*.

11.5.1 Relation between Binomial and Poisson Distributions

In the binomial distribution

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x}$$

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

If *n* is large while the probability *p* of occurrence of an event is close to zero so that *q = (1 – p)* is close to 1, then the event is called a rare event.

In practice an event is considered rare if the number of trials is at least 50 (n≤50) while *np* is less than 5.

That is,

np< 5
50*p*< 5
Thus *p*< 0.1

In such cases the binomial distribution is very closely approximated by the Poisson distribution with $\mu = np$.

μ is normally the *average number* of successes outcomes occurring in a given time interval.

Comparison of Properties

	Binomial	Poisson
Mean	μ = np	μ = np
Variance	σ^2 = npq	σ^2 = μ = np
Standard Deviation	σ = √npq	σ = √μ = √np

Example 11.5

The average number of days the school is closed due to snow during winter is 4. What is the probability that the school will close for 6 days during winter?

Solution

μ = 4
x = 6

$$P(x;\mu) = \frac{\mu^x e^{-\mu}}{x!} = \frac{4^6 e^{-4}}{6!}$$

$$= \frac{4.4.4.4.4.4}{6.5.4.3.2.1} \times 0.0183 = \mathbf{0.1042}$$

Example 11.6

Ten percent of the tools produced in a certain manufacturing process turn out to be defective. Find the probability that in a sample of 10 tools chosen at random, exactly two will be defective by using

(a) the binomial distribution
(b) the Poisson distribution

Solution

Probability of a defective tool = ***p = 0.1***

(a)
P(2 defective tools in 10) = $b(x;n,p)$

$$= b\left(2;10,\frac{1}{10}\right) = \binom{10}{2}\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^8$$

$$= \binom{10}{2}\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^8 = 45 \times \left(\frac{9}{10}\right)^8 = \mathbf{0.194}$$

(b)
$\mu = np = 10 \times 0.1 = 1$

$$P(2\ defective\ tools\ in\ 10) = \frac{\mu^x e^{-\mu}}{x!}$$

$$= \frac{1^2 e^{-1}}{2!} = \frac{e^{-1}}{2} = \mathbf{0.184}$$

In general good agreement is obtained when ***p ≤ 0.1***

Example 11.7

If the probability that an individual suffers a bad reaction from injection of a given serum is 0.001, determine the probability that out of 2000 individuals

(a) exactly 3,

(b) more than 2 individuals will suffer a bad reaction.

Solution

P (x individuals suffer bad reaction)

$$= \frac{\mu^x e^{-\mu}}{x!}$$

$$\left.\begin{matrix} p = 0.001 \\ n = 2000 \end{matrix}\right\} np = \mu = \frac{2000}{1000} = \mathbf{2}$$

(a)

P(3 individuals suffer a bad reaction)

$$= \frac{2^3 e^{-2}}{3!} = \frac{8}{6} \times 0.1353$$

$$= \mathbf{0.18}$$

(b)

P(more than 2 suffer) = 1 – {P(0 suffer) + P(1 suffer) + P(2 suffer)}

$$\mathrm{P}(0 \text{ suffer}) = \frac{2^0 e^{-2}}{0!} = e^{-2}$$

$$\mathrm{P}(1 \text{ suffers}) = \frac{2^1 e^{-2}}{1!} = 2e^{-2}$$

$$\mathrm{P}(2 \text{ suffers}) = \frac{2^2 e^{-2}}{2!} = 2e^{-2}$$

Hence P(more than 2 suffer) = 1 – $(e^{-2} + 2e^{-2} + 2e^{-2})$

= 1 – $5e^{-2} = 1 - 5 \times 0.1353$

= 1 – 0.6765 = $\mathbf{0.3235}$

According to the Binomial distribution (a) would be

$$\binom{2000}{3} (0.001)^3 (0.999)^{1997}$$

$$= \frac{2000}{3!\,1997!} (0.001)^3 (0.999)^{1997}$$

which is very difficult to work out.

EXERCISE 10

10.1
(a) A coin is tossed 6 times. What is the probability that two heads will show up in the 6 tosses?
(b) A binomial distribution has a mean of 40 and a standard deviation of 6. Calculate the number of observations and the frequency of occurrence of the event.

10.2 A housewife buys a dozen eggs of which three turn out to be bad. She chooses five eggs at random to prepare breakfast. Find the probability that she chooses:
(i) all good eggs,
(ii) three good and two bad eggs,
(iii) at least one bad egg.

10.3 If 3% of the electric bulbs manufactured by a company are defective, find the probability that in a sample of 100 bulbs (a) 0, (b) 1, (c) 2 bulbs are defective.

10.4 A bag contains one red and seven white marbles. A marble is drawn from the bag and its colour is observed. Then the marble is put back into the bag and the contents are thoroughly mixed. Using (a) the binomial distribution, and (b) the Poisson approximation, find the probability that in 8 such drawings a red ball is selected exactly 3 times.

10.5 The probability that a person from a certain respiratory infection is 0.002. Find the probability that fewer than 5 of the next 2000 so infected will die.

10.6 A fair die is tossed 5 times. Calculate the probability that:
(i) a 3 or 5 appears on the first throw.
(ii) A 5 appears on each throw.
(iii) Four 6's appear in the five throws.

10.7 In a certain clan the probability of a family having a boy is 0.6. If there are 5 children in a family, determine:
(i) the expected number of girls,
(ii) the probability that there are at least three girls,
(iii) the probability that they are all boys.

10.8 It is known that 20% of the items produced by a certain manufacturer are defective. If 150 items are selected at random from the products, determine:
(a) the expected number of defective items,
(b) the standard deviation.

10.9 A doctor has observed that 8 out of every 10 people who contract a certain disease recover when given the necessary treatment. Five people are admitted to a hospital.
(a) Find the probability that at least 3 will recover.
(b) What is the probability that all will recover?

12.0
NORMAL DISTRIBUTION

12.1 Introduction

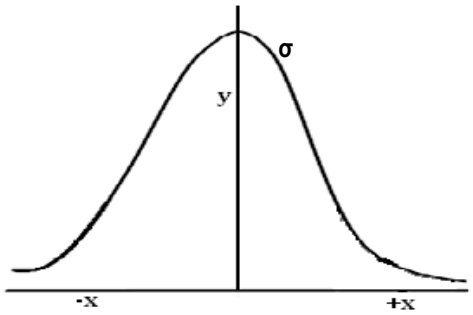
In the previous chapter we have seen that probabilities associated with binomial experiments are readily obtainable by the usual methods from the formula ***b(x;n,p)*** as long as ***n*** is small. If ***n*** is large then it becomes very cumbersome to use the binomial formula to calculate the probabilities. However in such cases it is possible to find an alternative way of calculating the probabilities. This is made possible by the fact that as ***n*** becomes large, the shape of the binomial distributions becomes very similar whatever the values of ***n*** and ***p*** are.

As ***n*** increases, the plot of probability, ***p(x)*** against the number of successes becomes more symmetric. As ***n*** increases the resulting figure is a symmetric bell-shaped type of distribution. This form of symmetrical and bell-shaped distributions is always reached provided ***n*** is sufficiently large. The value of ***n*** will of course depend on how nearly ***p*** and ***q*** equal a half. In other words the more ***p*** and ***q*** are different the bigger ***n*** will be to attain a symmetrical shape. If ***n*** becomes very large and the tops of the lines are joined by a curve, the distribution approaches the smooth form and this is the*normal curve*.

12.2 The Normal Curve

The normal curve is a bell-shaped curve which is often approximated closely by frequency distributions. Many sets of data that occur in nature, industry and research are found to be normally distributed. This curve is a very important type of continuous probability distribution.

A random variable *X* having the bell-shaped distribution is called a *normal random variable*.



The normal curve is symmetrical about its mean $\boldsymbol{\mu}$ and standard deviation σ. This means that the mean and median coincide. When x = 0, y has its maximum value, and therefore the mean and mode also coincide (see skewness).

12.2 The Equation of the Normal Curve

The theoretical relation giving the normal curve is

$$\boldsymbol{y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}} \qquad -\infty < x < +\infty \tag{12.1}$$

Where;
N = the total number of observations
μ = the mean of the distribution
σ = the standard deviation of the distribution
π = 3.14159…
e = 2.71828…

Theoretically, the curve does not reach the x-axis.

12.4 Properties of the Normal curve

(i) It is bell shaped
(ii) It is symmetrical about the mean, **μ**.
(iii) It extends from -∞ to +∞.
(iv) The total area under the curve is 1.
(v) If X is the normal variable / distribution, **μ** is the mean and **σ²**is the variance, then the normal distribution is denoted as;

$X \sim N(\mu, \sigma^2)$

The probability that *X* lies between **a** and **b** is written as$P(a \leq X \leq b)$. To find the probability, you need to find the area under the normal curve between **a** and **b**.
One way of finding the area described above is to integrate the normal curve equation or use standard **Z**-tables.

12.4 The standard normal variable, Z.

Since integration is lengthy, it is preferable to use tables in order to get probabilities for normal distributions.
However, in order to use the same set of tables for all possible values of ***μ*** and ***σ²***, the variable ***X*** is standardized so that the mean is 0 and the standard deviation is 1.
Since the standard deviation is 1, the variance is also 1.

The standard normal variable is called Z and it is denoted as below;

$Z \sim N(0,1)$

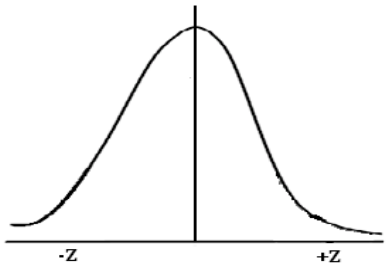
12.5 The standardisation procedure;

To standardise*X*, where $X \sim N(\mu, \sigma^2)$;

- Subtract the mean, *μ*
- Then divide by the standard deviation,*σ* to obtain the formula shown below;

$$\boldsymbol{Z = \frac{x - \mu}{\sigma}} \qquad \ldots standardisation\ formula \qquad (12.2)$$

Graphical illustration for a standardized normal variable, Z



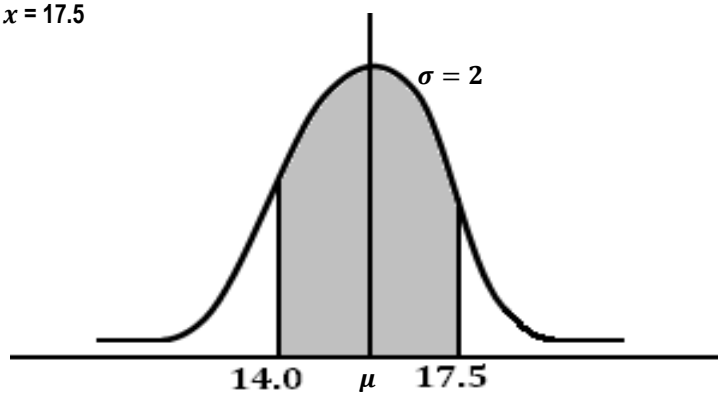
12.6 The normal approximation to the binomial

With distributions which are approximately normal we can use the areas under the normal curve to approximate binomial probabilities when ***n*** is sufficiently large, by use of the following theorem:

Theorem: If *X* is a normal random variable with mean *μ* = *np* and variance
$\sigma^2 = npq$, then the limiting form of the distribution of $Z = \frac{X-np}{\sqrt{npq}}$ as $n \to \infty$ is the standardized normal distribution *n(z; 0, 1)*. The best approximation is obtained when both ***np*** and ***nq*** are greater than 5.

Example 12.1
A Normal distribution has a mean $\mu = 16$ and a standard deviation $\sigma = 2$. Find the following areas under the curve.

(a) From x = 14 to x = 17.5



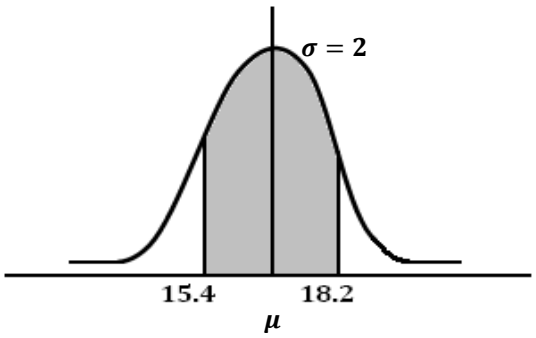
We standardize the given x values as follows:

$$z_1 = \frac{14 - 16}{2} = -\frac{2}{2} = -1.00$$

$$z_2 = \frac{17.5 - 16}{2} = \frac{1.5}{2} = 0.750$$

P(14 < x < 17.5) = P(-1.00 < z < 0.750) = 0.34134 + 0.27337 = **0.61471**

(b) From x = 15 to x = 18.2



$$z_1 = \frac{15.4 - 16}{2} = -\frac{0.6}{2} = -0.300$$

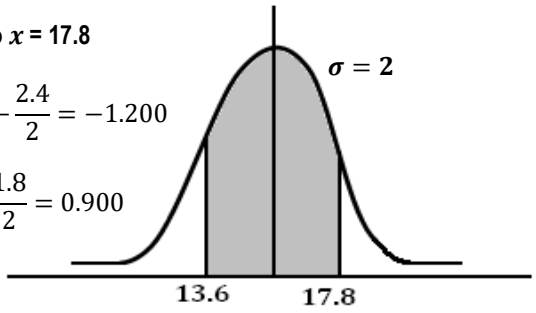
$$z_2 = \frac{18.2 - 16}{2} = \frac{2.2}{2} = 1.100$$

P(15.4 < x < 18.2) = P(-0.033 < z < 1.100) = 0.11791 + 0.36433 = **0.48224**

(c) From x = 13.6 to x = 17.8

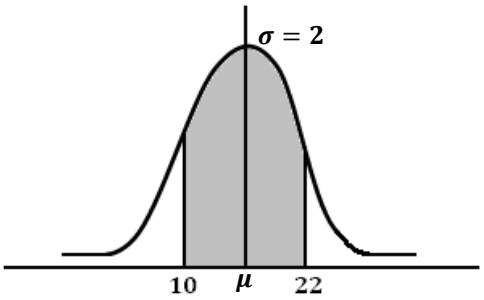
$$z_1 = \frac{13.6 - 16}{2} = -\frac{2.4}{2} = -1.200$$

$$z_2 = \frac{17.8 - 16}{2} = \frac{1.8}{2} = 0.900$$



P(13.6 < x < 17.8) = P(-1.200 < z < 0.900) = 0.38493 + 0.31594 **= 0.70087**

(d) From x = 10 to x = 22



$$z_1 = \frac{10 - 16}{2} = -\frac{6}{2} = -3.000$$

$$z_2 = \frac{22 - 16}{2} = \frac{6}{2} = 3.000$$

P(10 < x < 22) = P(-3.000 < z < 3.000) = 0.49865 + 0.49865 = **0.99730**

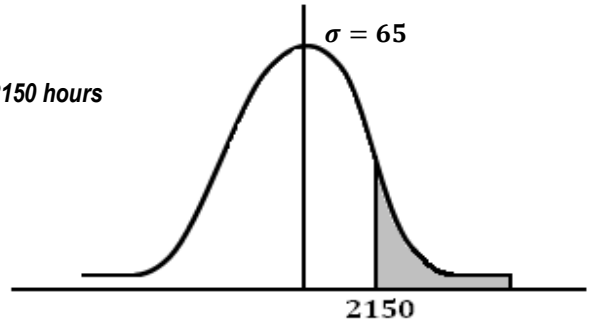
Example 12.2

Tests carried out on electric light bulbs indicated that their lifetime could be assumed to be normally distributed with a mean of 2000 hours and standard deviation of 65 hours. Estimate the proportion of bulbs which can be expected to burn:

(i) more than 2150 hours,
(ii) less than 1900 hours

Solution

(i) ***More than 2150 hours***



The proportion of bulbs which can be expected to burn more than 2150 hours is represented by the shaded area under the curve. Standardising the given information we have

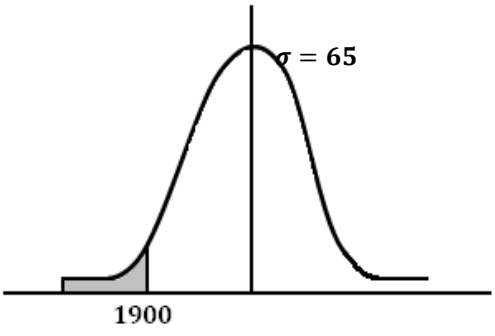
$$z_2 = \frac{2150 - 2000}{65} = \frac{150}{65} = 2.3077$$

Hence

P(X < 2150) = P(Z > 2.3077) = 1 – 0.98956 = 0.01044

Therefore the proportion of bulbs which can be estimated to burn more than 2150 hours is 1.044%

(ii) ***Less than 1900 hours***



$$z_2 = \frac{1900 - 2000}{65} = \frac{-100}{65} = \mathbf{-1.538}$$

Hence

P(X > 1900) = P(Z > -1.538) = 0.5000 – 0.4380 = 0.062

Therefore the proportion of bulbs which can be estimated to burn less than 1900 hours is 6.2%

Example 12.3

In testing squash balls, one hundred standard balls were dropped from a height of 200 cm and the height to which the ball rose at the first bounce was measured. A ball was "fast" if it rose above 80 cm. The average height of bounce was 75 cm and the standard deviation was 2 cm. What would be the change of getting a "fast" standard ball?

Solution

$$z_2 = \frac{80 - 75}{2} = \frac{5}{2} = 2.5000$$

This means that the chance of getting a "fast" ball is 0.0062 as can be ascertained from the Normal tables.

Example 12.4

Sacks of flour automatically packed by a machine loader have a mean weight of 100 kg. It is found that 10% of the bags are over 102 kg. Determine the standard deviation.

Solution

This means that 10% = 0.100 is the area under the curve which we are interested in. We therefore need a standard variable Z_0 such that

$P(Z>Z_0) = 0.1000$. From the Normal tables we obtain $Z_0 = 1.280$.

Thus

$$Z_0 = \frac{x - \mu}{\sigma}$$

$$1.280 = \frac{102 - 100}{\sigma}$$

$$\sigma = \frac{102 - 100}{1.280} = \mathbf{1.56}$$

Example 12.5

If an unbiased coin is tossed 100 times, what is the probability that:

(i) There will be more than 60 heads?
(ii) There will be at least 45 and at most 5 heads?
(iii) There will be fewer than 43 heads?

Solution

In this case we have to use the theorem already given earlier because this is not truly a Normal distribution. However we have to modify the expression for the standardized variable Z because we are dealing with a discrete distribution. In other words we have to define the limits for the given X values. This is done in the following way.
For p(at most N successes) we have

$$\boldsymbol{Z} = \frac{(\boldsymbol{N} + \frac{\mathbf{1}}{\mathbf{2}} - \boldsymbol{np})}{\sqrt{\boldsymbol{npq}}} \qquad (12.3)$$

For p(at least N successes) we have

$$\boldsymbol{Z = \frac{(N - \frac{1}{2} - np)}{\sqrt{npq}}} \qquad (12.4)$$

Now we can proceed to solve out the problem.

μ = np = 100 × ½ = 50
σ^2 = npq = 100 × ½ × ½ = 25
σ = $\sqrt{25}$ = **5**

For X> 60 heads we have to use 60.5, that is

$$Z = \frac{60.5 - 50}{5} = 2.100$$

P(X > 60) = P(Z > 2.100) = **0.0179**

At least 45 and at most 55 heads we have

$$\mathrm{Z}_1 = \frac{44.5 - 50}{5} = -1.100$$

$$\mathrm{Z}_2 = \frac{55.5 - 50}{5} = 1.100$$

Therefore

P(45 < X < 55) = P(-1.100 < Z < 1.100) = **0.7286**

Fewer than 43 heads we have

$$\mathrm{Z} = \frac{42.5 - 50}{5} = -1.500$$

Therefore

P(X < 43) = P(Z < -1.500) = 0.0668

Example 12.6
A coin is tossed 400 times. Use the normal-curve approximation to find out the probability of obtaining:

(a) Between 185 and 210 heads inclusive.
(b) Exactly 205 heads.

Solution
μ= np = 400 × ½ = 200
σ^2 = *npq* = 40 × ½ × ½ = 100
Therefore σ = $\sqrt{\mathrm{npq}}$ = 10.

$$\mathrm{Z}_1 = \frac{184.5 - 200}{10} = -1.55$$

$$\mathrm{Z}_2 = \frac{210.5 - 200}{10} = 1.050$$

Therefore P(-1.55 < Z < 1.050) = **0.7925**.

(b)

$$\mathrm{Z}_1 = \frac{204.5 - 200}{10} = 0.450$$

$$\mathrm{Z}_2 = \frac{205.5 - 200}{10} = 0.550$$

Therefore P(0.450 < Z < 0.550) = 0.0352

EXERCISE 11

11.1 A normal distribution has mean μ = 12 and standard deviation σ = 2. Find the following areas under the curve:
- *(a) From x = 10 to x = 13.5*
- *(b) From x = 11.4 to x = 14.2*
- *(c) From x = 9.6 to x = 13.8*
- *(d) From x = 6 to x = 18*

11.2 A Normal distribution has a mean μ = 16 and a standard deviation σ = 2. Find the following probabilities:
- *(a) x = 14 to x = 17.5*
- *(b) x = 15 to x = 18.2*
- *(c) x = 13.6 to 17.8*
- *(d) From x = 10 to x = 22*

11.3 If X is normal, calculate:
- *(a) P(4.5 ≤ X ≤ 6.5) where μ= 5 and σ = 1*
- *(b) P(X ≤ 800) where μ= 400 and σ = 200*
- *(c) P(800 ≤ X) where μ= 400 and σ = 200*

11.4. Suppose that a proportion of men's heights is normally distributed with a mean of 68 inches, and standard deviation of 3 inches. Find the proportion of the men who:
- *(a) are over 6 feet*
- *(b) are under* $5\frac{1}{2}$ *feet*
- *(c) are between* $5\frac{1}{2}$ *and 6 feet.*

11.5. The weight of bars of soap produced in a certain factory are normally distributed. Given that 10% of the bars weigh less than 140 g and 20% weigh more than 165 g, find:
- *(i) The mean and variance of the distribution.*
- *(ii) The percentage of bars that would be expected to weigh less than 145 g.*

If the variance of the distribution is reduced by 30%, find the proportion of bars that would be expected to weigh less than 148 g.

11.6. The volume of soft drinks bottled by a certain company is approximately normally distributed with mean 300 mls and standard deviation 2 mls. Determine the probability that in a sample of 10 bottles at least two bottles contain less than 297.4 mls.

11.7 A certain firm sells maize flour in bags of mean weight 40 kg and standard deviation of 2 kg. Given that the weight is normally distributed, find:
- *(i) The probability that the weight of any bag taken at random will lie between 41.0. and 42.5 kg.*
- *(ii) The percentage of bags whose weight exceeds 43 kg.*
- *(iii) The number of bags rejected out of a 500 bag purchase by a retailer whose consumers cannot accept a bag whose weight is below 38.5 kg.*

11.8 A woman wrote to "ASK THE DOCTOR" saying that she had been pregnant for 310 days before giving birth. Complete pregnancies are normally distributed with a mean of 266 days and a standard deviation of 16 days. Find the probability that a pregnancy lasts longer than:
- *(i) 270 days*
- *(ii) 300 days*

11.9 The weight of 500 students are normally distributed with mean 75 kg and standard deviation 7.5 kg. Find the proportion of students who weigh:
- *(i) Between 60 and 80 kg.*
- *(ii) More than 90 kg.*

11.10 The masses of packages from a particular machine are normally distributed with a mean of 200g and standard deviation of 2g. Find the probability that a randomly selected package from the machine weighs:

(i) Less than 197g
(ii) More than 200.5g
(iii) Between 198.5g and 199.5g

11.11 The time taken by a milk man to deliver milk to town is normally distributed with mean of 12 minutes and standard deviation of 2 minutes. He delivers milk every day.
Determine the number of days during the year (365 days) when he takes:

(i) Longer than 17 minutes
(ii) Longer than 10 minutes
(iii) Between 9 and 13 minutes

PART C
ESTIMATION & TESTING

13.0
STATISTICAL INFERENCE

13.1 Introduction

Statistical inference can be defined as the process by which conclusions are drawn about some measure or attribute of a population i.e. the mean or standard deviation based upon analysis of sample data.

13.2: Types of inference

Statistical inference is divided into two main types;

(i) *Sampling theory (Estimation) and*
(ii) *Hypothesis testing.*

These are explained as below;

(i) Estimation (Sampling theory)

Sampling theory is a branch of statistical inference that deals with the estimation of population characteristics such as the population mean and standard deviation from sample characteristics such as the sample mean and standard deviation. Population characteristics are known as population parameters while sample characteristics are known as sample statistics.

(ii) Hypothesis testing

Hypothesis testing is a branch of statistical inference that deals with the process of setting up a theory or hypothesis about some characteristics of the population and then sampling to see if the hypothesis is supported or not.
Part of sampling theory has been discussed in the chapter 1; therefore the chapters that follow will examine estimation of population mean using sample data, followed by a systematic coverage of hypothesis testing.

13.3 SAMPLING THEORY

Sampling theory is a study of relationships existing between a population and samples drawn from the population. The theory is useful in estimation of unknown population quantities (such as the population mean, variance, etc.) often called population parameters, from a knowledge of corresponding sample quantities (such as sample mean, variance, etc.) often called sample statistics.

Sampling theory is also useful in determining whether observed differences between two samples are actually due to change variation or whether they are really significant. Such questions arise, for example, in testing a new vaccine for use in treatment of a disease or in deciding whether one production process is better than another. Their answers involve use of so-called tests of significance and hypotheses, which are important in the theory of decision.

Properties of good estimators
A good estimator must exhibit the following properties;

(i) Unbiased

An estimator is said to be unbiased if the mean of the sample means $\bar{x}$ of all possible random sample size, n, drawn from a population of size N, equals the population parameter, μ.
Thus, the mean of the distribution of sample means would equal the population mean.

(ii) Consistency

An estimator is said to be consistent if, as the sample size increases, the precision of the estimator of the population parameter also increases.

(iii) Efficiency

An estimator is said to be more efficient than another if, in repeated sampling, its variance is smaller.

(iv) Sufficiency

An estimator is said to be sufficient if it uses all the information in the sample in estimating the required population parameter.

Estimation of the population mean

Due to the inability of obtaining all the population values, it is difficult to calculate the population parameters. Therefore, sampling is required in order to estimate the population parameters. Therefore, to distinguish the sample parameters from those of the population, the following distinguishing symbols are used to denote the different parameters;

	Sample Statistic	**Population Parameter**
Arithmetic mean	$\bar{x}$	μ
Standard deviation	s	σ
Number of items	n	N

Therefore to make accurate estimation of population parameter, samples must be chosen so as to be representative of a population. A study of methods of sampling and the related problems which arise is called *the design of the experimen*t. One way in which a representative sample may be obtained is by a process *called random sampling* (refer to chapter 1), according to which each member of a population has an equal chance of being included in the sample.

Sampling with and without Replacement

If we choose a member from a population, we have a choice of replacing the member or not replacing the member back into the population before choosing another member. In the first case the member can come up again and again, whereas in the second case it can only come up once. Sampling where each member of a population may be chosen more than once is called *sampling with replacement*, while if each member cannot be chosen more than once is called *sampling without replacement*. Populations are either finite or infinite. A finite population in which sampling is with replacement can theoretically be considered infinite, since any number of samples can be drawn without exhausting the population. For many practical purposes, sampling from a finite population which is very large can be considered as sampling from an infinite population.

Sampling Distributions

Consider all possible samples of size **n** which can be drawn from a given population (either with or without replacement). For each sample we can compute a statistic, such as the mean, standard deviation, etc., which will vary from sample to sample. In this manner we obtain a distribution of the statistic which is called its *sampling distribution*.
For example, if the particular statistic calculated is the sample mean, the distribution obtained is called the *sampling distribution of means* or the *sampling distribution of the mean*. Similarly we could obtain the sampling distribution of standard deviations, variances, medians, proportions, and so on. Now for each sampling distribution, we can compute the mean, standard deviation, etc. Thus we can also speak of the mean of the sampling distribution of means, the standard deviation of the sampling distribution of means and so on.

Definitions

Definition 1. If $x_1, x_2, \ldots, x_n$ represents a random sample of size ***n***, then the ***sample variance*** has the value;

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad (13.1)$$

Example 13.1

Find the variance of the random sample whose observations are 12, 15, 17, 20.

Solution

$$\bar{x} = \frac{12 + 15 + 17 + 20}{4} = \frac{64}{4} = 16$$

$$\boldsymbol{s^2} = \sum_{i=1}^{4} \frac{(x_i - 16)^2}{4 - 1}$$

$$= \frac{(12 - 16)^2 + (15 - 16)^2 + (17 - 16)^2 + (20 - 16)^2}{4 - 1}$$

$$= \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{34}{3} = \mathbf{11\frac{1}{3}}$$

Note:

If $\bar{x}$ is a decimal number that has been rounded off, we accumulate a large error using the sample-variance formula in the above form, to avoid this, we use a more useful computational formula which falls under the following definition.

Definition 2. If s^2 is the variance of a random sample of size $\boldsymbol{n}$, we may write

$$\boldsymbol{s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}} \qquad (13.2)$$

Example13.2
Find the variance of the sample whose observations are: 3, 4, 5, 6, 6, 7.

Solution

$$\bar{x} = \frac{3 + 4 + 5 + 6 + 6 + 7}{6} = \frac{31}{6} = \mathbf{5.1666}$$

The above is a decimal round off, therefore the above formula will be adapted.

Consider the table below;

$\boldsymbol{x_i}$	$\boldsymbol{x_i^2}$
3	9
4	16
5	25
6	36
6	36
7	49
$\boldsymbol{\sum x_i = 31}$	$\boldsymbol{\sum x_i^2 = 171}$

Hence

$$s^2 = \frac{6(171) - (31)}{6(5)} = \frac{1026 - 961}{30} = \mathbf{2.17}$$

Definition 3: The sample standard deviation, denoted by the statistic, ***s***, is defined to be the positive square root of the sample variance. In the above example,

$s = \sqrt{2.17} = \mathbf{1.47}$

The standard deviation of the sampling distribution of a statistic is called the ***standard error*** of the statistic.

Sampling distributions of the mean

Consider sampling from a discrete uniform population consisting of the values 0, 1, 2, 3. Clearly each value has a probability ¼ chances to occur. We have

$$Themean, \mu, = E(x) = \sum_{x=0}^{3} \boldsymbol{xp(x)}$$

$$= \frac{0+1+2+3}{4} = \mathbf{\frac{3}{2}}$$

The variance is

$$\sigma^2 = [(X-\mu)^2] = \sum_{x=0}^{3} (x-\mu)^2\, p(x)$$

$$= \sum (X^2) - \mu^2$$

$$= \sum_{x=0}^{3} (x)^2\, p(x) - \mu^2$$

$$= 0 + \frac{1}{4} + \frac{4}{4} + \frac{9}{4} - \frac{3^2}{2} = \mathbf{\frac{5}{4}}$$

Theorems

Theorem 13.1

If all the possible random samples of size ***n*** are drawn *with replacement* from a finite population of size ***N*** with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$, then the sampling distribution of the mean $\bar{X}$ will be approximately normally distributed with mean $\boldsymbol{\mu_{\bar{x}} = \mu}$ and standard deviation $\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$. From which we have;

$$\boldsymbol{z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}} \qquad (13.3)$$

as the value of a standard normal variable ***Z***.

This theorem is valid for any finite population when **n≥ 30**. If **n< 30**, the results will be valid only when the population being sampled is not too different from a normal population.

Theorem 13.2

If all possible random samples of size **n** are drawn *without replacement* from a finite population of size ***N*** with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$, then the sampling distribution of the mean $\boldsymbol{\bar{X}}$ will be approximately normally distributed with mean and standard deviation given by;

$$\boldsymbol{\mu_{\bar{x}} = \mu}$$

$$\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \qquad (13.4)$$

Example 13.3

Given the population 1, 1, 1, 3, 4, 5, 6, 6, 6, 6, 7 find the mean and standard deviation for the sampling distribution of means for samples of size 4 selected at random without replacement.

Solution

x	f	fx	fx^2
1	3	3	3
3	1	3	9
4	1	4	16
5	1	5	25
6	3	18	108
7	1	7	49
	$\sum f = 10$	$\sum fx = 40$	$\sum fx^2 = 210$

$$Mean, \mu = \frac{\sum fx}{\sum x} = \frac{40}{10} = \mathbf{4.0}$$

$$\sigma^2 = \frac{\sum fx^2}{\sum f} - \mu^2 = \frac{210}{10} - 16 = \mathbf{5}$$

Using the theorems, the mean and standard deviation for the sampling distribution of means are;

$$\mu_{\bar{x}} = 4$$

$$\boldsymbol{\sigma_{\bar{x}} = \frac{\sqrt{5}}{\sqrt{4}}\sqrt{\frac{10-4}{10-1}}}$$

$$= \frac{\sqrt{5}}{2}\sqrt{\frac{6}{9}} = \mathbf{0.91}$$

Example 13.4

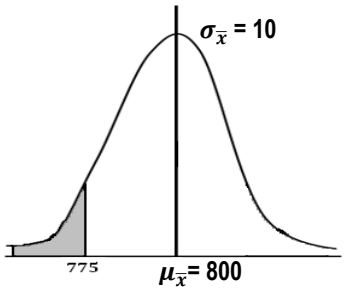
An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Solution

The sampling distribution of $\bar{X}$ will be approximately normal, with $\mu_{\bar{x}} = 800$ and

$$\sigma_{\bar{x}} = \frac{40}{\sqrt{16}} = 10$$

The desired probability is given by the area of the shaded region in the figure below.



Corresponding to $\bar{x}$ = 775, we obtain

$$z = \frac{775 - 800}{10} = -\frac{25}{10} = -\mathbf{2.5}$$

Therefore,

P (X < 775) = P (z < -2.5) = **0.0062**

Example 13.5

A population consists of the five numbers 2, 3, 6, 8, 11. Consider all the possible samples of size two which can be drawn with replacement from this population.

Compute

(a) the mean of the population
(b) the standard deviation of the population,
(c) the mean of the sampling distribution of the means,
(d) the standard error.

Solution

(a)

$$Mean\ , \mu = \frac{2 + 3 + 6 + 8 + 11}{5} = \frac{30}{5} = 6$$

(b)

$$\sigma^2 = \frac{\sum f x^2}{\sum f} - \mu^2$$

$$= \frac{4 + 9 + 36 + 64 + 121}{5} - 36$$

$$= \frac{234}{5} - 36 = 46.8 - 36.0 = \mathbf{10.8}$$

Therefore, σ = $\sqrt{10.8}$ = $\mathbf{3.29}$

(C)METHOD I

No.	Sample	x	No.	Sample	x	No.	Sample	x
1	2,2	2	9	3, 8	5.5	17	8,3	5.5
2	2,3	2.5	10	3, 11	7.0	18	8,6	7.0
3	2,6	4.0	11	6, 2	4.0	19	8,8	8.0
4	2,8	5.0	12	6,3	4.5	20	8,11	9.5
5	2,11	6.5	13	6,6	6.0	21	11,2	6.5
6	3,2	2.5	14	6,8	7.0	22	11,3	7.0
7	3,3	3.0	15	6,11	8.5	23	11,6	8.5
8	3,6	4.5	16	8,2	5.0	24	11,8	9.5
						25	11,11	11.0

	F	fx	fx^2
2.0	1	2	4.0
2.5	2	5	12.5
3.0	1	3	9.0
4.0	2	8	32.0
4.5	2	9	40.5
5.0	2	10	50.0
5.5	2	11	60.5
6.0	1	6	36.0
6.5	2	13.0	84.5
7.0	4	28	196.0
8.0	1	8	64.0
8.5	2	17	144.5
9.5	2	19	180.5
11.0	1	11	121.0
	$\sum f = 25$	**$\sum fx = 150$**	**$\sum fx^2 = 1035$**

Mean = $\frac{150}{25}$ = **6.0**

Hence

$$\sigma_{\bar{x}}^2 = \frac{\sum fx^2}{\sum f} - \mu^2$$

$$= \frac{1035}{25} - 36 = 41.4 - 36 = \underline{\mathbf{5.4}}$$

Therefore, $\sigma_{\bar{x}}$= standard error = $\sqrt{5.4}$ = **2.32**

METHOD II

A shorter method would be to use the following formula:

Standard error =$\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$
That is:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{10.8}{2}} = \sqrt{5.4} = \mathbf{2.32}$$

Example 13.6

Repeat Example **13.4** for sampling without replacement

Solution

(a) Mean of the population = 6 as in Example 13.5 (a) above.
(b) Standard deviation = 3.29 as in Example 13.5 (b) above.
(c) The samples are obtained by drawing one number then drawing another number different from the first. The sampling is illustrated in Table 13.3 below.

No.	Sample	x
1	2,3	2.5
2	2,6	4.0
3	2,8	5.0
4	2,11	6.5
5	3,6	4.5
6	3,8	5.5
7	3,11	7.0
8	6,8	7.0
9	6,11	8.5
10	8,11	9.5
		$\sum$ fx = 60.0

The mean of the sampling distribution of mean is

$$\frac{\sum x}{n} = \frac{60.0}{10} = 6.0$$

illustrating the fact that $\boldsymbol{\mu_x = \mu}$.

(d) The variance of the sampling distribution of means is:

$$\boldsymbol{\sigma_{\bar{x}}} = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\sqrt{\frac{\boldsymbol{N-n}}{\boldsymbol{N-n}}} = \frac{3.2}{\sqrt{2}}\sqrt{\frac{5-3}{5-1}} = \frac{1.814}{1.414}\sqrt{\frac{3}{4}} = \mathbf{2.01}$$

14.0 ESTIMATION THEORY

14.1 Introduction

We have seen in chapter 1 that if a sample is taken from a population, then the mean of the sample is a good estimate for the mean of the population itself. Let us now consider what one can say about the population variance.
Most samples have a variance which is less than that of the population from which the samples were taken. The reason for this is that the extremes of the population were probably not included in the sample. Another reason is that each sample has its variance calculated from its own mean and not from the mean of the population. Hence we need a way of estimating the variance of the population from that of the sample. This can be started in form of a theorem.

Theorem 14.1 If samples of size *n* are drawn at random from a parent population with variance σ^2, the sample variances, s^2, constitute a population with mean

$$Mean\ of \qquad s^2 = \frac{n-1}{n}\sigma^2 \tag{14.1}$$

This implies that

$$E(s^2) = \frac{n-1}{n}\sigma^2 \tag{14.2}$$

But since E (s^2) is still an approximation of σ^2, then we replace σ^2 by $\hat{\sigma}^2$stands for the approximate value of the population variance. Hence we can write (without loss of generality) e.g. equation as 14.1

$$s^2 = \frac{n-1}{n}\hat{\sigma}^2 \tag{14.3}$$

Or

$$\hat{\sigma}^2 = \frac{n-1}{n}s^2 \tag{14.4}$$

where $\mathbf{s^2}$ *is the mean sample variances.*

We note that as long as *n* is large the factor $\frac{n}{n-1}$ is very close to 1. We consider a sample to be small if it has a size of less than 30. Now a single number estimating a whole population is by itself of little value. It is more useful to know a set of values between which the parameter lies. This is why often we have to give what we call *confidence limits* or a *confidence interval* of population parameters. If we did not give an interval but only single values for $\hat{\mu}$ or $\hat{\sigma}^2$ then we would have given what are called *point estimates*. It is better to give an interval because we are more likely to have enclosed in the interval of the correct population parameter.

14.2 Confidence limits

(a) **The 95% confidence interval**.

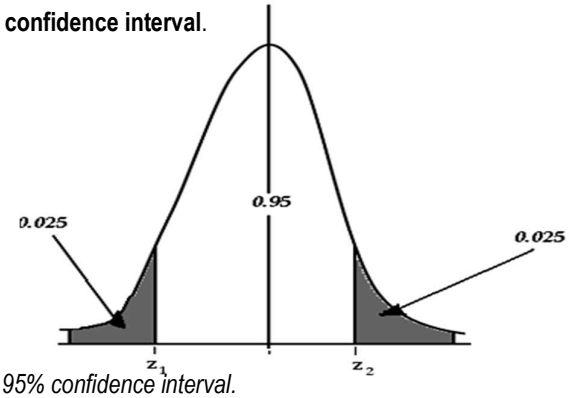


Fig. 14.1 *The 95% confidence interval.*

The 95% confidence interval for a population parameter implies that the population parameter lies within an area of 0.95 of the normal curve.

The z-values which determine this interval are

z_1 = **-1.96** and z_2 = **1.96**

(from the normal curve tables –Schedule I).

The 95% confidence limits are given by

$$\boldsymbol{\mu \pm \frac{1.96\sigma}{\sqrt{N}} = \mu \pm \frac{1.96s}{\sqrt{N}}} \qquad (14.5)$$

(b) The 99% confidence Interval.

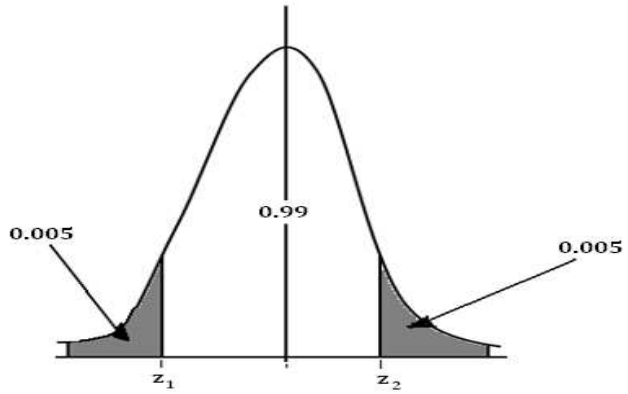


Fig. 14.2 *The 99% confidence interval.*

The 99% confidence interval for a population parameter implies that the population parameter lies within an area of 0.99 of the normal curve.

The z-values which determine this interval are

z_1 = **-2.575** and z_2 = **2.575**

(Again from the normal curve tables–Schedule I)

The 99% confidence limits are given by

$$\boldsymbol{\mu \pm \frac{2.58\sigma}{\sqrt{N}} = \mu \pm \frac{2.58s}{\sqrt{N}}} \qquad (14.6)$$

Example 14.1

In a sample of size 100 we have a mean of 40 and standard deviation of 11. Determine the 95% confidence interval for the mean of the population.

Solution

The 95% confidence limits are given by

$$\mu \pm \frac{1.96\sigma}{\sqrt{N}} \quad = \quad \mu \pm \frac{1.96s}{\sqrt{N}} = \quad 40 \pm 1.96 \times \frac{11}{\sqrt{100}} \quad = \quad 40 \pm \frac{1.96 \times 11}{10} = \quad \boldsymbol{40 \pm 2.16}$$

Implying that the population mean lies in the interval: **37.84 to 42.16**

We are 95% confident that the population mean lies between 37.84 and 42.16.

Example 14.2
A sample of size 70 has a mean of 65 with a standard deviation of 4.2. Find 98% confidence limits for the mean of the parent population.

Solution
The 98% confidence limits are given by

$$\mu \pm \frac{2.33\sigma}{\sqrt{N}} \quad = \quad \mu \pm \frac{2.33s}{\sqrt{N}} = \quad 65 \pm 2.33 \times \frac{4.2}{\sqrt{70}} = \quad 40 \pm \frac{233 \times 4.2}{8.37} = \mathbf{40 \pm 1.17}$$

Implying that the population mean lies in the interval: **63.83** to **66.17**

We are 98% confident that the population mean lies between 63.83 and 66.17.

Example 14.3
An anthropologist measured the heights (in inches) of a random sample of 100 men from a certain population, and found the sample mean and variance to be 71 and 9 respectively. Find the 99% confidence interval.

Solution
N = 100 men; μ = 71; s = √9 = 3

The 99% confidence limits are given by

$$\mu \pm \frac{2.575\sigma}{\sqrt{N}} \quad = \quad \mu \pm \frac{2.575s}{\sqrt{N}} = \quad 71 \pm 2.575 \times \frac{3}{\sqrt{100}} = \quad 71 \pm \frac{2.575 \times 3}{10} = \mathbf{71 \pm 0.77}$$

Implying that the population mean lies in the interval: **70.23** to **71.77**

We are 99% confident that the mean of the population lies between 70.23 and 71.77.

When the sample is large, that is greater than 30, we can use the sample variance (s^2) as an approximation to the population variance (σ^2).

14.3 SMALL SAMPLES

The t – distribution

By small sample we mean a case where n < 30. Here the situation is not the same as the case where n ≥ 30. In this case there is greater difficulty in estimating the variance σ^2 because the term $\frac{n}{n-1}$ is no longer close to 1.

Furthermore, if the sample size is smaller i.e. < 30, the arithmetic means of small samples are not normally distributed.

Therefore, for samples where *n*< 30 we now use a distribution called the *t-distribution*.

Where the statistic, t, is given by;

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{14.7}$$

Where, ***s***, is an estimate of ***σ***.

Features of the t-distribution

(i) It is an exact distribution which is unimodal and symmetric about 0.
(ii) It is flatter than the normal distribution i.e. the area near the tails are greater than the normal distribution.
(iii) As the sample size becomes larger, the *t*-distribution approaches the normal distribution.

Degrees of freedom

t-distributions are used hand in hand with the appropriate degrees of freedom denoted by;

$$\boldsymbol{\upsilon = n - 1} \tag{14.7}$$

*where; **n** is the number of items in the sample.*

The use of the t-distribution and its accompanying tables is best illustrated with an example.

Example 14.4
A random sample of size 20 taken from a normal distribution has a mean $\bar{x}$ = 32.8 and a standard deviation *s* = 4.51. Determine the 95% confidence interval of μ.

Solution
For small samples where n < 30, the 95% confidence limits are given by

$$\bar{\boldsymbol{x}} \pm \boldsymbol{t}_{\alpha/2}\frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}} \tag{14.9}$$

$\bar{\boldsymbol{x}}$ **= μ**

Where;

$$\boldsymbol{t}_{\alpha/2} \;=\; \boldsymbol{t}_{0.05/2} \;=\; \boldsymbol{t}_{0.025}$$

υ = n – 1 = 20 – 1 = 19 degrees of freedom.

Hence we want $t_{0.025}$ for 19 degrees of freedom from the tables. This is found to be **2.093**.

Hence the 85% confidence limits are given by;

$$= \;\; \bar{\boldsymbol{x}} \pm \frac{2.093 \times 4.51}{\sqrt{20}} \;=\; 32.8 \pm \frac{2.093 \times 4.51}{4.47} \;=\; \mathbf{32.8 \pm 2.11}$$

Implying that the population mean **μ**, lies between: **30.69 and 34.91**

We are 95% confident that the mean lies between 30.69 and 34.91.

Example 14.5
Estimate the 99% confidence interval of the parent population if a sample of ten drawn from it had a mean of 15.0 and a standard deviation of 1.8.

Solution
For small samples where n < 30, the 98% confidence limits are given by

$$\bar{\boldsymbol{x}} \pm t_{0.02/2}\frac{s}{\sqrt{n}}$$

In this problem **υ = n – 1 = 10 – 1 = 9** degrees of freedom.

Hence

$$\bar{\boldsymbol{x}} \pm t_{0.01}\frac{s}{\sqrt{n}} \;=\; 15.0 \pm \frac{2.821 \times 1.8}{\sqrt{10}} \;=\; 15.0 \pm \frac{2.821 \times 1.8}{3.16} \;=\; \mathbf{15.0 \pm 1.61}$$

Implying that the mean μ of the population lies between: **13.39 and 16.61**

We are 98% confident that the mean lies between 13.39 and 16.61.

Example 14.6

Sixteen weather stations at random locations in a state measure rainfall. They record an average of 10 inches and standard deviation of 1.5 inch. Construct a 99% confidence interval for the mean rainfall for the state.

Solution
This is a small sample; therefore we use the t-distribution.

99% → α = 1% = 0.01

which implies that α/2 = 0.005

υ = n – 1 = 16 – 1 = 15 degrees of freedom.
Hence the confidence limits are given by

$$\bar{\boldsymbol{x}} \pm t_{0.05}\frac{s}{\sqrt{n}} \;=\; 10.0 \pm \frac{2.947 \times 1.5}{\sqrt{16}}$$

$$= 10.0 \pm \frac{2.947 \times 1.5}{4}$$

$$= \mathbf{10.0 \pm 1.11}$$

Implying that the mean rainfall for the state lies between
8.89 and **11.11**

Example 14.7

The weights of 10 boxes of cereal are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 and 9.8 ounces. Find a 99% confidence interval for the mean of all such boxes of cereal, assuming an appropriate normal distribution.

Solution

The mean $\bar{\boldsymbol{x}}$ = 10.06.

$$s^2 = \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n-1} = \frac{(0.14)^2 + (-0.36)^2 + \cdots + (-0.26)^2}{(10-1)} = \frac{0.544}{9} = \mathbf{0.0604}$$

$$Therefore,\; s = \sqrt{0.0604} = 0.246$$

For the 99% confidence interval we have
$t_{\alpha/2} = t_{0.01/2} = t_{0.005}$

N = 10, hence υ = n – 1 = 9 degrees of freedom.

Hence the confidence limits are given by

$$\bar{\boldsymbol{x}} \pm t_{0.05}\frac{s}{\sqrt{n}} \;=\; 10.06 \pm \frac{3.250 \times 0.246}{\sqrt{10}}$$

$$= 10.06 \pm \frac{3.250 \times 0.246}{3.16}$$

$$= \mathbf{10.06 \pm 0.25}$$

Implying that the mean of all such boxes of cereal lies between

9.81 and **10.31**

14.4 Estimation of Population Proportions

Proportions represent an attribute of a population rather than the value of a variable. A proportion may represent the ratio of defective to non-defective goods produced in a factory or the proportion of consumers who plan to buy a given product or some similar information.

Estimating populations from sample statistics follows the same pattern as outlined for means, only that this time round binomial distributions are involved.

Theorem:
If **P** represents the proportion of successes then **q** will represent the proportion of failures in a sample of size **n**; the standard error of the sample proportion will then be given by;

$$\boldsymbol{S_{ps} = \sqrt{\frac{pq}{n}}} \qquad (14.10)$$

The confidence interval is given by;

$$\boldsymbol{p_s \pm Z_{cl} S_{ps}}$$

Where $\boldsymbol{Z_{cl}}$is the **Z** score at the given confidence level.

Example 14.8:

A random sample of 400 passengers is taken and 15% are in favour of the proposed new time tables. Determine at the 95% confidence level the proportion of all rail passengers infavour of the proposed time tables.

Solution

$$n = 400, p = 0.55, and\ q = 1 - 0.55 = 0.45$$

Hence;

$$S_{sp} = \sqrt{\frac{0.55 \times 0.45}{400}} = \mathbf{0.025}$$

Therefore, we are 95% confident that the population is between the two values;

$$P_s \pm 1.96 S_{ps} \quad since \quad Z_{95\%} = 1.96 \quad (see\ tables)$$

$$= 0.55 \pm 1.96(0.025) = 0.55 \pm 0.049 = \mathbf{0.501\ to\ 0.599}$$

EXERCISE 12

12.1. Suppose that the education level among adults in a certain country has a mean of 11.1 years and a variance of 9. What is the probability that in a random survey of 100 adults you will find an average level of schooling between 10 and 12?

12.2. An elevator is designed with a load limit of 2000 lb. It claims a capacity of 10 persons. If the weight of all the people using the elevator are normally distributed with a mean of 185 lband standard deviation of 22 lb, what is the probability that a group of 10 persons will exceed the load limit of the elevator?

12.3 An instructor wishes to estimate the mean performance of his students in a statistics class. He selects a random sample of 100 students, administers a test and obtains a mean of 15.5and standard deviation of 5. You are required to construct a 90% confidence interval for the mean performance of the students.

12.4. A random sample of 10 items is taken and found to have a mean weight of 60g and standard deviation of 12g. Determine the mean weight of the population:

(i) With 95% confidence (ii)with 99% confidence

12.5 From a random sample of 529 televisions off the production line it was found out that each set has 18 faults on average with a standard deviation of 3.45 faults. Determine the confidence limits for the production as a whole at:

(i) 99% level (ii)95% level

12.6 An inspection department is trying to determine the appropriate sample size to use. They wish to be within 1% of true population with 99% confidence. Past records show that the population defective is 3 in 100. What sample size should they use?

12.7: ABC Ltd manufactures light bulbs whose life span follows a normal distribution with mean of 1680 hrs and standard deviation of 20 hrs. In a certain period, ABC Ltd produced 2000 bulbs.

(a) Compute the number of bulbs that will last for less than 1700hrs.
(b) ABC Ltd guarantees to replace bulbs which burn for less than 1640 hrs. What percentage of the bulbs will they replace?
(c) A new machine was bought to improve the product. What should be the new guarantee if ABC Ltd replaces only 0.8% of the bulbs produced?
(d) Calculate the probability that the bulbs selected at random burns within the range of 1400 and 1600 hours.

12.8 A random sample of 10 packets was taken and found to have a mean weight of 50g and a standard deviation of 9g. Compute the mean weight of the population at both;

(a) 95% confidence and
(b) 99% confidence

Chapter 15
HYPOTHESIS TESTING

15.1 Introduction

A hypothesis is defined as a statement or theory about population parameters like the mean and variance, either because it is believed to be true or it is to be used as a basis for argument, but has not yet been proved.

15.2 Types of Hypothesis

There are four types of hypothesis i.e.

(a) Null hypothesis

The null hypothesis, $\mathbf{H_0}$, represents a theory that has been put forward, either because it is believed to be true or it is to be used as a basis for argument, but has not been proved.

Example 15.1

A null hypothesis may be that the new accounting policy is not better, *on average*, than the current accounting policy.

Therefore, we denote;

$\mathbf{H_0}$: there is no difference between the two accounting policies on average.

(b) Alternative hypothesis

The alternative hypothesis is a statement of what a statistical hypothesis is set up to establish.

Example 15.2

In a company's trial of the new accounting policy, the alternative hypothesis might be that "the new accounting policy adopted has a different application", on average, compared to the current accounting policy.

Therefore, we denote;
$\mathbf{H_1}$: the two accounting policies have a different application on average.
OR;
The alternative hypothesis might be that the new accounting policy is better, on average, than the current accounting policy.

Therefore we denote;

$\mathbf{H_1}$: the new accounting policy is better than the current accounting policy on average.

Special consideration is given to the null hypothesis. This is due to the fact that the null hypothesis relates to the statement being tested, whereas the alternative hypothesis relates to the statement to be accepted if or when the null is rejected.

The final conclusion once the test has been carried out is always given in terms of the null hypothesis.

i.e; We either *"reject H_0 in favour of H_1,"* OR
"Do not reject H_0."

We never conclude;

"Reject H_1" OR
"Accept H_1"

(c) Simple hypothesis

This is a hypothesis that specifies the population distribution completely.

(d) Composite hypothesis

This is a hypothesis which does not specify the population distribution completely.

15.3 TYPE I AND TYPE II ERRORS

(a) Type I Error

In a hypothesis test, a type I error occurs when the null hypothesis is rejected when it is in fact true. In other words, H_0 is wrongly rejected.

Type I error is also known as an error of the first kind.

Example 15.3

In a company's trial of a new accounting policy, the null hypothesis might be that the new accounting policy is not better on average than the current accounting policy; i.e.

H_0: there is no difference between the two accounting policies on average.

A type I error would occur if we conclude that the two accounting policies have a different application on average when in fact there is no difference between them.

(b) Type II Error

In a hypothesis test, a type II error occurs when the null hypothesis H_0, is not rejected when it is in fact false.

Type II error is also known as an error of the second kind.

From the above example, a type II error would occur if it was concluded that the two accounting policies have the same application i.e. there is no difference between the two accounting policies on average, yet in fact they have different applications.

The following table gives a summary of possible results of any hypothesis test;

		Decision	
		Reject H_0	**Don't reject H_0**
Truth	**H_0**	Type I Error	Right decision
	H_1	Right decision	Type II Error

A *type I error* is often considered to be *more serious*, and therefore more important to avoid, than a type II error. The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0.

Mathematical expression

Mathematically, the two types of errors are expressed in terms of the risk levels involved in committing such an error; The associated risk with the two types of error are denoted by α and β, thus;

P(Type I error) = **α**

P(Type II error) = **β**

The exact probability of a type II error is generally unknown and for any given set of data, type I and type II errors are inversely related, the smaller the risk of one, the higher the risk of the other.

Other Key terms

(i) Test Statistic

A test statistic is a quantity calculated from our sample of data. Its value is used to decide whether or not the null hypothesis should be rejected in our hypothesis test.

(ii) Critical Value(s)

The critical value(s) for a hypothesis test is a threshold to which the value of the test statistic in a sample is compared to determine whether or not the null hypothesis is rejected.

The critical value for any hypothesis test depends on the *significance level* at which the test is carried out, and whether the test is *one-sided* or *two-sided*.

(iii) Critical Region

The critical region CR, or rejection region RR, is a set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test. That is, the sample space for the test statistic is partitioned into two regions; one region (the critical region) will lead us to reject the null hypothesis H_0, the other will not. If the observed value of the test statistic is a member of the critical region, we conclude "Reject H_0"; if it is not a member of the critical region then we conclude "Do not reject H_0".

(iv) Significance Level

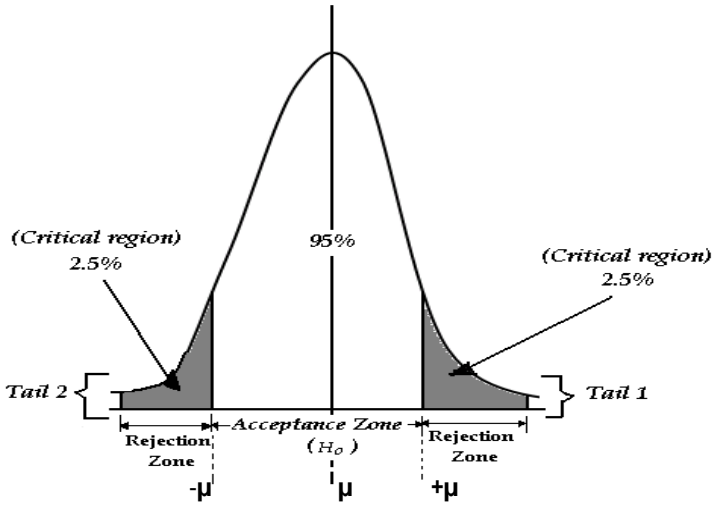
The significance level of a statistical hypothesis test is a fixed probability of wrongly rejecting the null hypothesis H_0, if it is in fact true.

It is the probability of a type I error and is set by the investigator in relation to the consequences of such an error. That is, we want to make the significance level as small as possible in order to protect the null hypothesis and to prevent, as far as possible, the investigator from inadvertently making false claims.

Accordingly, various levels of significance are chosen but usually, the significance level is chosen at the 0.05 level (or equivalently, 5%) or 0.001 level (equivalently 1%).

Therefore, the result of a particular test might be expressed as follows;“The difference between the sample mean and the hypothetical population mean is significant at the 5% level.”

Diagrammatically;(for a 5% significance level)



The above is a normal curve for a 5% significance level where by the standard normal Z computed must lie in the interval Z = -1.96 and Z= +1.96 if the null hypothesis is to be accepted. Otherwise, if its outside the range (greater than 1.96), the null hypothesis will be rejected and consequently the alternative hypothesis will be accepted.

(v) Two-Sided Test

A two-sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0 are located in ***both*** tails of the probability distribution.

In other words, a two sided test is a *two-tailed test of significance*.

(vi) One-sided Test

A one-sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0 are located entirely in one tail of the probability distribution.

In other words, a one-sided test is *a one-tailed test of significance*.

The choice between a one-sided and a two-sided test is determined by the purpose of the investigation or prior reasons for using a one-sided test.

Example 15.4

The mean age of employees in a company is known to be 40 years with a standard deviation of 6 years. A random sample of 36 employees is taken and the mean age is found to be 42.4 years. It is required to conduct a significance test at the 5% level.

Solution

We shall set up a hypothesis as follows;

"The population mean age, μ is 40 years" ……..**Null hypothesis** (H_0)

"The population mean is not equal to 40 years" …….**Alternative hypothesis** (H_1)

Hence we denote;
H_0: μ = 40 years
H_1: μ ≠ 40 years

We shall therefore compute the standard normal Z and compare the calculated

Z-score with the appropriate value for the required level of significance, i.e 1.96 for 5% significance.

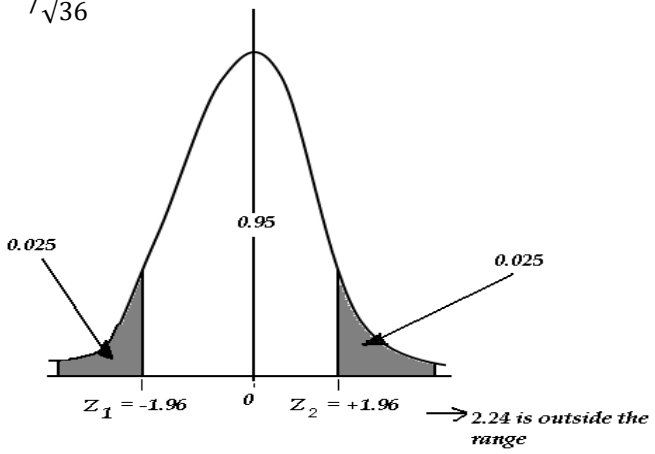
Therefore, following the same procedure as discussed before;

$\bar{x}$= 42.4 years
σ= 6 years
n = 36 employees

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{42.4 - 40}{6/\sqrt{36}} = \frac{2.24}{1}$$

= **2.24**

Graphically,



The calculated **Z** score i.e. 2.24 is outside the range ± 1.96, therefore H_0 is rejected and H_1 is accepted.

Example 15.5

We shall now re-compute the hypothesis but this time using a 1% level of significance.

H_0: μ = 40 years
H_1: μ ≠ 40 years

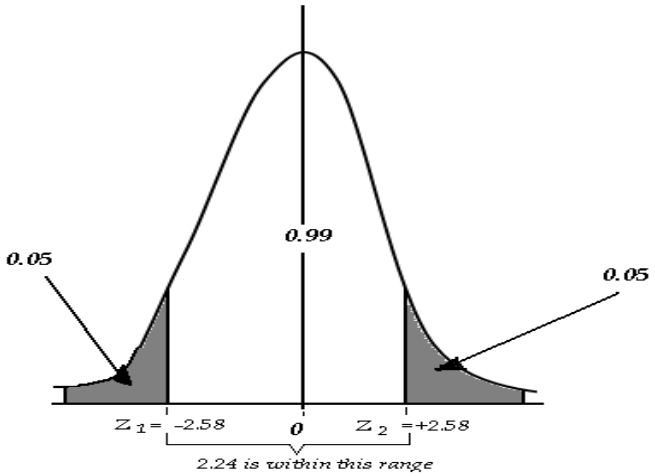
We shall therefore compute the standard normal Z and compare the calculated Z-score with the appropriate value at the 1% level of significance, i.e. 2.58

$\bar{x}$= 42.4 years
σ = 6 years
n = 36 employees

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{42.4 - 40}{6/\sqrt{36}} = \frac{2.24}{1}$$

= **2.24**

Graphically,



The calculated **Z** score i.e. 2.24 is this time in the range ± 2.58, therefore $\mathbf{H_0}$ is accepted at the 1% level.

Therefore the difference between the sample mean and the hypothetical mean is not significant at the 1% level.

NOTE:

(a) In example 1 we are likely to make a type I error while in example 2 we are likely to make a type II error.

The greater the level of significance, the greater the probability of making a type I error.

Therefore the 1% level (i.e. 99% level of significance) is more severe than the 5% level (i.e. the 95% level of significance).

(b) Example 1 and 2 exhibit a two tailed test of significance. Example 3 below will demonstrate hypothesis testing in a one tailed test of significance.

Example 15.6
The mean wage expense of employees in a company is known to be shs. 580 per day with a standard deviation of Shs. 100. A random sample of 100 employees is taken and the mean wage expense is found to be Shs. 600. You are required to conduct a significance test and determine whether the population mean wage is greater than shs. 580 using the 5% level of significance.

Solution

We shall set up a hypothesis as follows;

"The population mean wage, μ is shs. 580" ……..**Null hypothesis** (H_0)

"The population mean wage is greater than Shs. 580" …**Alternative hypothesis** (H_1)

Hence we denote;

$\mathbf{H_0}$: μ = shs. 580
$\mathbf{H_1}$: μ >shs. 580

This is a one tailed test since we are interested only in the right-hand tail of the distribution.

$\bar{x}$= shs. 600
σ = shs. 100*s*

n = 100 employees

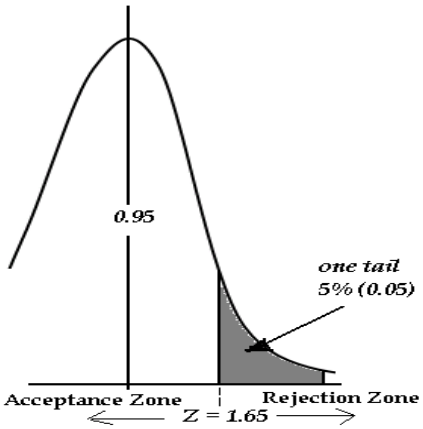
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{600 - 580}{100/\sqrt{100}} = \frac{20}{10}$$

= **2.0**

From the tables; P(z=???) = 0.05
P(z>???) = 0.5 – 0.05 = **0.4500**

The appropriate Z value that gives you the above probability = **1.65**

Graphically,



The calculated **Z** score i.e. 2.0 is located in the rejection zone since it is greater than 1.65, therefore $\mathbf{H_0}$ is rejected and H_1 is accepted. We shall then conclude that X and μ are significant at the 5% level of significance.

15.4 Hypothesis testing of the difference between means

When two random samples are taken, frequently it's required to know if there is a difference between the two means.

The distribution of sample mean difference is normally distributed with standard errors of the difference means given as;

$$\boldsymbol{S_{(\bar{X}_A - \bar{X}_B)}} = \sqrt{\frac{\boldsymbol{S_A^2}}{\boldsymbol{n_A}} + \frac{\boldsymbol{S_B^2}}{\boldsymbol{n_B}}} \qquad (15.1)$$

Where;- S_A = standard deviation of sample A, size n_A.
S_B = standard deviation of sample B, size n_B.

Therefore; the **Z**-score will be calculated from;

$$\boldsymbol{Z} = \frac{\boldsymbol{\bar{X}_A - \bar{X}_B}}{\boldsymbol{S_{(\bar{X}_A - \bar{X}_B)}}} \qquad (15.2)$$

Example 15.7

A company is known to depreciate all machines in the category A and category B of assets using the same depreciation rate. It is required to test if the mean depreciation rate of the two categories is the same. A random sample of 144 machines from category A had a mean depreciation rate of 36.4% with a standard deviation of 3.6%, while a random sample of 225 machines from category B had a mean depreciation rate of 36.9% with a standard deviation of 2.9%.
You are required to test if the mean depreciation rates are significantly different at the 5% level.

Solution

H_0: Mean of A = Mean of B
H_1: Mean of A ≠ Mean of B

This is a two tailed test, since we are concerned with the direction of the variation.

Therefore;

$$S_{(\bar{X}_A - \bar{X}_B)} = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}$$

$$S_{(\bar{X}_A - \bar{X}_B)} = \sqrt{\frac{3.6^2}{144} + \frac{2.9^2}{225}}$$

= **0.357**

$$Z = \left| \frac{36.40 - 36.90}{0.357} \right|$$

= **1.4**

The calculated **Z**-score i.e. *z* = 1.4 is within the range ± 1.96 (Acceptance zone), hence, H_0 is accepted.

Therefore; there is no significance difference between the mean depreciation rate of the two categories A and B of the asset group.

i.e. The depreciation rate of A = Depreciation rate of B

15.5 PROPORTIONS

15.5.1 Hypothesis testing of Proportions

Proportions are variables that follow a binomial distribution. The general standard error of proportions is given as;

$$s_p = \sqrt{\frac{pq}{n}} \tag{15.3}$$

With Z-score given as;

$$Z = \frac{p - \pi}{s_p} \tag{15.4}$$

Where; p = proportion found in the sample
π = the hypothetical proportion

15.5.2 Hypothesis testing of the difference between Proportions

If two samples are randomly selected from the same population, the best estimate of the two sample proportions is given as;

$$p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} \tag{15.5}$$

- The standard error of difference of two proportions is given by;

$$s_{(p_1-p_2)} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} \tag{15.6}$$

- Finally, the Z-score is computed as below**;**

$$\frac{(\boldsymbol{p_1} - \boldsymbol{p_2}) - (\boldsymbol{\pi_1} - \boldsymbol{\pi_2})}{\boldsymbol{s}_{(\boldsymbol{\pi_1}-\boldsymbol{\pi_2})}} \tag{15.7}$$

The relevant hypothesis is conducted as before at the required level of significance.

Ref:- See examples in the test exercises!

15.6 Hypothesis testing for small samples

(*t*-distribution)

As previously discussed, small distributions are those with a sample size less than 30.
The means of small samples are distributed around the population mean in a manner similar to but not exactly, a normal distribution.

The probability distribution of small sample means is calculated using the
t-distribution discussed in the previous chapter.

Example 15.8
A firm ordered sacks of chemicals with a normal weight of 50kg. a random sample of 8 sacks was taken and it was found that the sample mean was 49.2Kg with a standard deviation of 1.6Kg.
You are required to test whether the mean weight of the sample sacks is significantly less than the nominal weight, using a 5% level of significance.

Solution
This is a one-tailed test as we are interested in whether the mean weight is less than the nominal.

Therefore;
H_0: mean = 50kg
H_1: mean < 50kg

Degrees of freedom = n – 1 = 8 – 1 = **7**

Therefore;

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \left| \frac{49.2 - 50}{1.6/\sqrt{8}} \right| = \frac{49.2 - 50}{0.566} = \mathbf{1.413}$$

In order to find a one-tailed 5% probability point, we look up the 0.10 (two tailed) point in the t-distribution table. This value is 1.895.
Therefore, the calculated**t-**score i.e. 1.413 lies between the interval ±1.895 therefore we can accept the null hypothesis and reject the alternative hypothesis at the 5% level.

Note:
The 10% probability point in the **t**-distribution table is spread in both tails, i.e. 5% in each tail.

15.6.1 *t*-distribution tests of the difference between means

If the samples are small, the ***t***-distribution can be used to test for differences in population means.

The following should however be noted;

(i) The two sampled populations are normally or near normal distribution.
(ii) The difference between the two standard deviations is negligible and denoted as S_p.

The value of S_p is given as;

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_1 - 1)S_2^2}{n_1 + n_2 - 1}} \quad (15.8)$$

The above value is then used to compute the standard error for each sample mean as shown below;

Standard error for sample 1:

$$S_{\bar{x}_1} = \frac{S_p}{\sqrt{n_1}} \quad (15.9)$$

Standard error for sample 2:

$$S_{\bar{x}_2} = \frac{S_p}{\sqrt{n_2}} \quad (15.10)$$

The sampling error of the distribution of difference in sample means;

$$S_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{S_{\bar{x}_1}^2 + S_{\bar{x}_2}^2} \quad (15.11)$$

The value can be used to find the *t*-score in the normal manner;

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{(\bar{x}_1 - \bar{x}_2)}} \text{ with } n_1 + n_2 - 2 \text{ degrees of freedom} \quad (15.12)$$

Example 15.9
The monthly bonuses of two groups of salesmen are being investigated to see if there is a difference in the average bonus received. Random samples of 12 and 9 are taken from the two groups and it can be assumed that the bonuses in both groups are approximately normally distributed and that the standard deviations are about the same. A 5% level of significance is to be used.

The sample results are;

n_1= 12 n_2= 9
$\bar{x}_1$= shs. 1,060 $\bar{x}_2$= Shs. 970
s_1 = 63 s_2 = Shs. 76

Solution

H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

It will be seen that this is a *two-tailed* test.
The common standard deviation, S_p, is calculated first from;

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_1 - 1)S_2^2}{n_1 + n_2 - 1}}$$

$$S_p = \sqrt{\frac{(12 - 1)63^2 + (9 - 1)76^2}{12 + 9 - 2}}$$

= **68.77**

Standard error for sample 1:

$$S_{\bar{x}_1} = \frac{S_p}{\sqrt{n_1}} = \frac{68.77}{\sqrt{12}} = \mathbf{19.85}$$

Standard error for sample 2:

$$S_{\bar{x}_2} = \frac{S_p}{\sqrt{n_2}} = \frac{68.77}{\sqrt{9}} = \mathbf{22.92}$$

The sampling error of the distribution of difference in sample means;

$$S_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{S_{\bar{x}_1}^2 + S_{\bar{x}_2}^2} = \sqrt{19.85^2 - 22.92^2} = \mathbf{30.32}$$

The value can be used to find the ***t***-score in the normal manner;

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{(\bar{x}_1 - \bar{x}_2)}} = \frac{1{,}060 - 970}{30.32} = \mathbf{2.97}$$

From the tables, the 5% value with $(n_1+n_2 - 2) = (12 + 9 - 2) = 19$ degrees of freedom is **2.093**.

Since the calculated value is outside the range ±2.093(i.e. greater than), the null hypothesis can be rejected at the **5%** level.
We therefore conclude that "*there is a significant difference*" between the mean monthly incomes.

Chapter 16
NON-PARAMETRIC TESTS

THE CHI-SQUARED (χ^2) DISTRIBUTIONS

16.1 Introduction

The tests described in the previous chapters *i.e.* the ***t***-tests and others are used on data comprising specific measurements *i.e*. quantitative data. These are known as "*Parametric tests*."

The Chi-distribution tests are used on non-parametric data and such are tests that simply compare frequencies of occurrence. These tests are particularly used when one wishes to test a theory about qualitative observations e.g. the "*type of company*" or "*quality of a product*."

16.2 The Test statistic of a χ^2 distribution

Suppose we have a set of observed (Actual) frequencies $O_1, O_2, \ldots, O_n$ and we wish to test some hypothesis about these frequencies.
Assuming our hypothesis is exactly true then we would have expected the frequencies $E_1, E_2, \ldots, E_n$ the expected frequencies; and the test statistic for testing the differences between the observed and expected frequencies is;

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_n - E_n)^2}{E_n}$$

Summarised as;

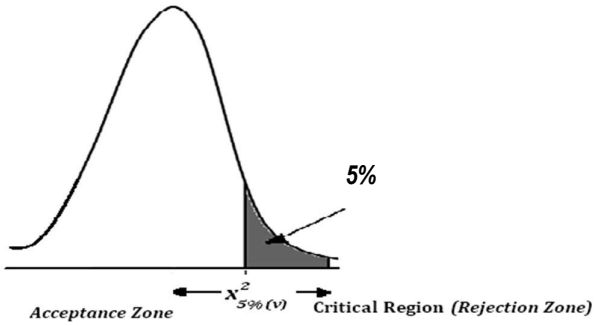
$$\chi^2 = \sum \frac{(\boldsymbol{O} - \boldsymbol{E})^2}{\boldsymbol{E}} \qquad (16.1)$$

16.3 Critical values and levels of significance

If the hypothesis is true, we would expect the χ^2 statistic to be small because the difference between the corresponding observed and expected frequencies would be small, but if the hypothesis was totally incorrect, then some of thedifferences would be large and hence the χ^2 statistic would be large.

The χ^2 statistic test is conducted as a one tailed (upper tail) test. The calculated value of the test statistic can lie either in the main bulk of the χ^2 distribution or the upper tail "*critical*" (or *rejection*) region; where the boundary of the critical region is called the *critical value*.
The critical value depends on the level of significance of the test often (5% or 1%).

Where: $\chi^2_{5\%(v)} = critical\ value$



If the test value lies in the critical region, then the null hypothesis $\boldsymbol{H_o}$ is rejected in favour of the alternative hypothesis $\boldsymbol{H_1}$.

Example 16.1
The table below shows a record of accidents at ***XYZ*** factory and the corresponding day of the week in which the accident occurred.

Day of week	Mon	Tue	Wed	Thu	Fri	Sat	Sun
No. of accidents	46	42	45	46	49	29	26

Required
Test the hypothesis that accidents are spread uniformly throughout the week at the 5% level of significance.

Solution
Hypotheses
$\boldsymbol{H_0}$**:** Accidents are spread uniformly throughout the week
$\boldsymbol{H_1}$**:** Accidents not spread uniformly
Therefore, expected number of accidents in each day of the week is given as;

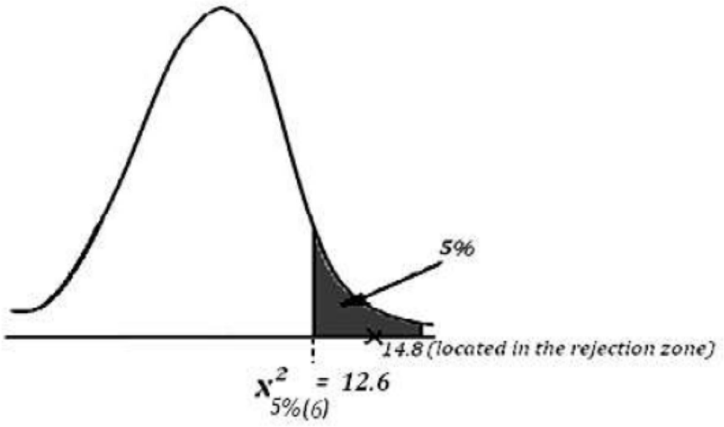
$$\boldsymbol{E} = \frac{46 + 42 + 45 + 46 + 49 + 2 + 26}{7} = \mathbf{40}$$

The test statistic is computed from the table below;

Day of the wk	$\boldsymbol{O}$	$\boldsymbol{E}$	$\frac{(\boldsymbol{O}-\boldsymbol{E})^2}{\boldsymbol{E}}$
Mon	46	40	0.90
Tue	42	40	0.10
Wen	45	40	0.625
Thu	46	40	0.90
Fri	49	40	2.025
Sat	29	40	3.025
Sun	23	40	7.225
			$\boldsymbol{\chi^2}$ **=14.8**

Degrees of freedom; $v = n - 1 = 7 - 1 = \boldsymbol{6\ degrees\ of\ freedom}$

The chi-squared tables, the critical value $\boldsymbol{\chi^2_{5\%(6)}}$= 12.6 i.e;



Since the calculated $\boldsymbol{\chi^2}$ statistic is located in the rejection zone, we conclude that accidents are not spread uniformly over the seven days.

Note:

(i) If the degree of freedom $\boldsymbol{v = 1}$, then use Yate's Correction;

$$\boldsymbol{\chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E}} \quad (16.2)$$

(ii) Any calculated value of expected frequency less than 5 must be added to the next preceding class of variables.

Example 16.2

The table below shows the age of **10** University students in the same year of study.

Student	P	Q	R	S	T	U	V	W	X	Y
Age (years)	24	20	27	26	22	18	20	21	22	20

Required
Test at the 5% significance level whether the students in the same year of study have the same age.

Solution

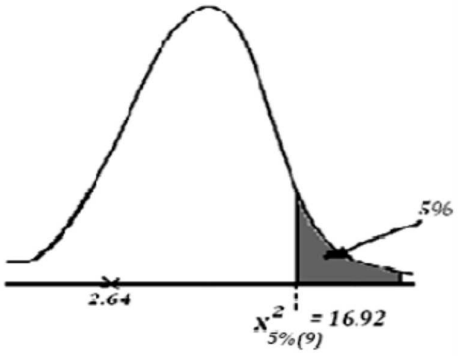
Hypothesis

$\boldsymbol{H_0}$: students in the same year of study have the same age.
$\boldsymbol{H_1}$: students in the same year of student have different age.

The expected age for each student will be;

$$\boldsymbol{E = \frac{24 + 20 + 27 + 26 + 22 + 18 + 20 + 21 + 22 + 20}{10} = 22\ years}$$

Student	$\boldsymbol{O}$	$\boldsymbol{E}$	$\boldsymbol{\frac{(O-E)^2}{E}}$
P	24	22	0.181
Q	20	22	0.181
R	27	22	1.140
S	26	22	0.730
T	22	22	0.000
U	18	22	0.730
V	20	22	0.181
W	21	22	0.045
X	22	22	0.000
Y	20	22	0.181
			$\boldsymbol{\chi^2 = 2.64}$



Critical value $\boldsymbol{\chi^2_{5\%(9)}}$=**16.92**(from tables).

Degrees of freedom; $v = n - 1 = 10 - 1 = \boldsymbol{9\ degrees\ of\ freedom}$
Since the calculated value of the test statistic is less than the critical value, accept H_0.

16.4 GOODNESS OF FIT

Goodness of fit tests are used to test whether a set of observed frequencies differs significantly from a specified probability distribution; for example binomial, Poisson, or normal.
Here, we calculate using the observed frequencies, the best estimate of the parameters belonging to specified probability distributions, and then use these estimates to calculate the expected frequencies for the distribution and compare them with the observed frequencies.

In goodness of fit tests, the number of degrees of freedom are given by;

$$v = \boldsymbol{number\ of\ classes - 1 - number\ of\ etimated\ parameters} \qquad (16.3)$$

Example 16.3 (*Goodness of fit for a binomial distribution*)

A farmer kept a record of the number of heifers born to each cow during the first years of breeding of the cow. The results are summarised in the table below;

No. of heifers	0	1	2	3	4	5
No. of cows	4	19	41	52	26	8

Required
Test at the 5% level whether this is a binomial distribution with parameters n = 5 and p = 0.5.

Solution

Let ***X*** be the number of heifer calves born to a cow in the first five years of breeding.

Hypotheses;

$\boldsymbol{H_0}: X \sim B(5,0.5) is a binomial distribution.$
$\boldsymbol{H_1}: X is not distributed in this way.$

We shall calculate the binomial probabilities using the normal formula; i.e.

$$P(X = x) = \frac{n!}{(n-x)!\,x!}.p^x q^{n-x} for x = 0,1,2,3,4,5.$$

The total number of cows is 150, so the expected frequencies are found by multiplying $P(X = x)$ by 150.

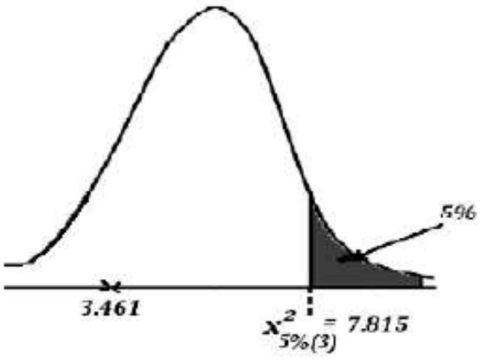
Therefore;

$\boldsymbol{x}$	$\boldsymbol{P(X = x)}$	**Expected frequency** $\boldsymbol{E = 150 \times P(X = x)}$
0	0.0313	4.7
1	0.1562	23.4
2	0.3125	46.9
3	0.3125	46.9
4	0.1563	23.4
5	0.0312	4.7

Since the expected frequency for the first and last classes is less than 5, combine them with the next classes.

Therefore, the test statistic will be given as per the computation below;

Student	O	E	$\frac{(O-E)^2}{E}$
0 or 1	23	28.1	0.925
2	41	46.9	0.742
3	52	46.9	0.554
4 or 5	34	28.1	1.238
	$\sum O = 150$	$\sum E = 150$	$\chi^2 = \mathbf{3.461}$



Degrees of freedom; $v = 4 - 1 = \mathbf{3\ degrees\ of\ freedom}$

Critical value, $\chi^2_{5\%(3)}$ **= 7.815**

Since the calculated χ^2 is less than the critical value do not reject H_0. We therefore conclude that X is a binomial distribution.

Note:

The expected frequency is computed using the formula;

$$E = n \times P(X = x) \qquad (16.4)$$

Where n represents the total number of observations.

Example 16.4 (*Goodness of fit for a Poisson distribution*)

A tool hire company has six mixtures for hire but complaints have been received about the non-availability of mixtures and the manager is wondering whether to increase the number available. Fortunately, records have been kept and the following data is available for the last 1000 working days.

Daily demand (No. of mixtures)	0	1	2	3	4	5	6	7	8	9	10
No. of Occasions	11	35	81	143	176	177	149	108	65	37	18

Required
Test the hypothesis that the observed data approximates to a Poisson distribution.

Solution

Probabilities of a Poisson distribution are obtained from;

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!} \quad for \quad x = 0,1,2,\ldots,n$$

$$But, \ \mu = \frac{(0 \times 11) + (1 \times 35) + (2 \times 81) + \cdots + (10 \times 18)}{1000} = \mathbf{4.9}$$

Hence for the above example, probabilities will be obtained from;

$$P(X = x) = \frac{4.9^x e^{-4.9}}{x!} \quad for \quad x = 0,1,2,\ldots,10$$

x	$P(X=x)$	$E = 1000 \times P(X=x)$	O	$\frac{(O-E)^2}{E}$
0	0.0074	7.4	11	1.75
1	0.0365	36.5	35	0.06
2	0.0894	89.4	81	0.79
3	0.1460	146.0	143	0.06
4	0.1789	178.9	176	0.05
5	0.1753	175.3	177	0.02
6	0.1432	143.2	149	0.23
7	0.1002	100.2	108	0.61
8	0.0614	61.4	65	0.21
9	0.0334	33.4	37	0.39
10	0.164	16.4	18	0.16
≥11	0.0119	11.9	0	11.90
				$\chi^2 = 16.23$

Degrees of freedom;

- There are 12 classes, **n = 12**.
- There are two restrictions i.e. ∑**E = 1000** and the mean of the Poisson distribution, **μ**.

Therefore;
$v = 12 - 1 - 1 = 10 degrees of freedom$

Critical value, $\chi^2_{5\%(10)}$= **18.30** (from tables)

Since the calculated $\chi^2 = 16.23$is less than the critical value (18.30) do not reject H_0. We therefore conclude that ***X*** is a Poisson distribution.

Example 16.5 (*Goodness of fit for a Normal distribution*)

The salary in thousands of shillings earned by Accountantsin their first year of employment is denoted by a random variable X. The value of *X* is obtained for a random sample of 86 employee accountants and the results are summarised in the table below;

X	< 35	35-45	45-55	55-65	> 65
Observed freq	10	18	28	18	12

Required
Carry out a χ^2-goodness of fit analysis to test, at the 5% level, the hypothesis that X is normally distributed.

Solutions

We shall compute the expected frequencies for each of the five classes by standardising each X value.

$P(X=x)$	$E = 86 \times P(X=x)$	O	$\frac{(O-E)^2}{E}$
$P(X < 35) = 0.1587$	13.7	10	0.999
$P(35 < X < 45) = 0.2109$	18.1	18	0.0005
$P(45 < X < 55) = 0.2608$	22.4	28	1.400
$P(55 < X < 65) = 0.2608$	18.1	18	0.005
$P(X > 65) = 0.01587$	13.7	12	0.210
			$\chi^2 = 2.61$

Degrees of freedom;

- There are five classes and one restriction i.e; ∑**E = 86**; therefore**;**

$$v = 5 - 1 = 4 degrees\ of\ freedom$$

Critical value, $\chi^2_{5\%(4)}$= **9.488** (from tables)

Since the calculated $\chi^2 = 2.61$is less than the critical value (9.488) do not reject H_0. We therefore conclude that ***X*** is a Normal distribution.

16.5: χ^2 CONTIGENCY TABLES

Chi-squared contingency tables are used to display results of sample data that is classified according to two different factors or attributes. Contingency tables are used to examine the independence of the two factors.

Again we have to decide whether or not the expected frequencies, calculated on the assumption that H_0 is true, differ significantly from the observed frequency.

The test statistic is as before; *i.e*;

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

The degrees of freedom are computed from;

$$v = (r - 1)(c - 1) degrees\ of\ freedom \quad (16.5)$$

Where;
$r = Number\ of\ rows$
$c = Number\ of\ columns$

The expected frequencies are computed from the formula;

$$Expected\ frequency = \frac{row\ total \times column\ total}{Grand\ total} \quad (16.6)$$

Example 16.6

The four production plants of Coca Cola (u) Ltd are based at Kampala, Jinja, Mbale and Njeru. A random sample of employees at each of these four plants has been asked to give their views on a productivity deal that the company is proposing.
The table below summarises these views.

	PRODUCTION PLANTS			
View	Kampala	Jinja	Mbale	Njeru
In favour	80	40	50	60
Against	35	30	40	25

Required
Test the hypothesis that there is no significance difference in views between the production plants.

Solution

The expected frequencies are computed as below;

	PRODUCTION PLANTS				
View	**Kampala**	**Jinja**	**Mbale**	**Njeru**	**Total**
In favour	$\frac{155 \times 230}{360} = \mathbf{73.47}$	$\frac{70 \times 230}{360} = \mathbf{44.72}$	$\frac{90 \times 230}{360} = \mathbf{57.50}$	$\frac{85 \times 230}{360} = \mathbf{54.31}$	**230**
Against	$\frac{155 \times 130}{360} = \mathbf{41.53}$	$\frac{70 \times 130}{360} = \mathbf{25.28}$	$\frac{90 \times 130}{360} = \mathbf{32.50}$	$\frac{85 \times 130}{360} = \mathbf{30.69}$	**130**
Total	**115**	**70**	**90**	**85**	**360**

Analysing the discrepancy between the observed and the Expected frequencies (the $\boldsymbol{\chi^2}$-statistic) we have;

Views	$\boldsymbol{O}$	$\boldsymbol{E}$	$\frac{(\boldsymbol{O}-\boldsymbol{E})^2}{\boldsymbol{E}}$
In favour	80	73.47	0.580
	40	44.72	0.498
	50	57.50	0.978
	60	54.31	0.596
Against	35	41.53	1.027
	30	25.28	0.881
	40	32.50	1.731
	25	30.69	1.055
	$\sum O = 200$	$\sum E = 200$	$\boldsymbol{\chi^2 = 7.35}$

Degrees of freedom;

$$\boldsymbol{v} = (\mathbf{2}-1)(\mathbf{4}-1) = 3\ degrees\ of\ freedom$$

Using the **0.05** level of the critical value will be;

$\boldsymbol{\chi^2_{5\%(3)}}$ = **7.84** (from tables)

Since the calculated $\boldsymbol{\chi^2}$statistic is less than the table value (7.84), we do not reject the null hypothesis (at the 5% significance level).

Example 16.7

In a survey of 200 (two hundred) boys of which 75 were intelligent, 40 had skilled fathers; while 85 of the intelligent boys had unskilled fathers.

Required
Using a $\boldsymbol{\chi^2}$test, determine whether the information supports the hypothesis that skilled fathers have intelligent boys at a 5% level of significance.

Solution

We shall display the information in a contingency table as below;

		Fathers		
		Skilled	**Unskilled**	**Total**
Boys (Sons)	Intelligent	40	35	**75**
	Unintelligent	40	85	**125**
	Total	**80**	**120**	**200**

The hypotheses are;

H_0: There is no relationship between the intelligence of boys and the fathers' skills.
H_1: There is a relationship.

To calculate the expected,E, frequencies, we have;

		Fathers		
		Skilled	**Unskilled**	**Total**
Boys (Sons)	Intelligent	$\frac{75 \times 80}{200} = \mathbf{30}$	$\frac{75 \times 120}{200} = \mathbf{45}$	**75**
	Unintelligent	$\frac{125 \times 80}{200} = \mathbf{50}$	$\frac{125 \times 120}{200} = \mathbf{30}$	**125**
	Total	**80**	**120**	**200**

Note the there are no expected frequencies less than 5.

Degrees of freedom;

$\boldsymbol{v = (2-1)(2-1) = 1}$, hence we shall use Yate's correction as shown in the table below;

Fathers	$\boldsymbol{O}$	$\boldsymbol{E}$	$\frac{(\lvert \boldsymbol{O-E} \rvert - \mathbf{0.5})^2}{\boldsymbol{E}}$
Skilled	40	30	3.00
	40	50	1.81
Unskilled	35	45	2.01
	85	75	1.20
	$\sum O = 200$	$\sum E = 200$	$\boldsymbol{\chi^2 = 8.02}$

Degrees of freedom; v = 1.

From tables, the critical value, $\chi^2_{5\%(1)}$= 3.841

We reject the null hypothesis since the calculated $\chi^2 = 8.02$ is greater than the critical value (3.841). We therefore conclude that there is no relationship between the boys' intelligence and the fathers' skills at the 5% significance level.

EXERCISE 13

13.1. A random sample of 400 households is classified by two characteristics, whether they own a colour T.V and type of householder. Householders are divided into three categories: landlord, tenant and servant.
The results of this investigation are shown in the table below;

	Observed frequencies		
	Landlord	Tenant	servant
Colour T.V	150	60	20
No Colour T.V	45	68	57

Require;
Carry out a test at the 5% level to determine whether ownership of a colour T.V is dependent on the nature of householder.

13.2 A survey was carried out in a firm about the smoking habits of men and women employees and the following results were obtained.

	Men	Women
Smokers	48	27
Non-Smokers	58	57

It is required to test at the 95% level, whether the survey reveals any difference in the smoking habits of men and women.

13.3 At Coca Cola a machine fills a bottle of soda with a mean volume capacity of 330mls. A random sample of 49 bottles is taken and the mean volume capacity is found to be 338mls with a standard deviation of 10mls. You are required to conduct a significance test at 5% level.

13.4 Briefly but clearly explain:

(i) The null hypothesis and the alternative hypothesis
(ii) The circumstances in which the paired t-test should be used to test the difference between two means.
(iii) The difference between parametric tests and non-parametric tests.
(iv) Type I and Type II errors.

13.5 An Accountant wishes to estimate the mean of accounts receivables. He takes a random sample of 10 accounts receivables and gets the following results in millions of shillings.

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

You are required to construct a 95% confidence interval of the true mean of accounts receivables. (Assume the accounts receivables are normally distributed).

13.6 The marketing manager of a company manufacturing detergents conducted a survey to assess customer's responses to a new packaging recently introduced. She classified the customers according to gender and from a random sample of 300 customers she got the following results.

	Preference		
	New Packaging	Old packaging	Undecided
Male	27	14	92
Female	73	82	12

You are required to test at 1% level of significance, if there is any significant relationship in packaging preference and gender.

13.7 The weights of 10 boxes of cereal are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 and 9.8 ounces. Find a 99% confidence interval for the mean of all such boxes of cereal, assuming an appropriate normal distribution.

13.8 Prior to a campaign, 35% of a sample of 400 house wives used a certain detergent. After the campaign, 40% of the second sample of 400 wives used the same product. Determine whether there was any significant increase in sales after the campaign.

13.9 The number of books borrowed from Aristock town library during a particular week were recorded as shown below;

Day of the week	Monday	Tuesday	Wednesday	Thursday	Friday	**Total**
No. of books borrowed	132	110	128	105	150	**625**

You are required to test the hypothesis that the number of books borrowed does not depend on the day of the week at the 1% level of significance.

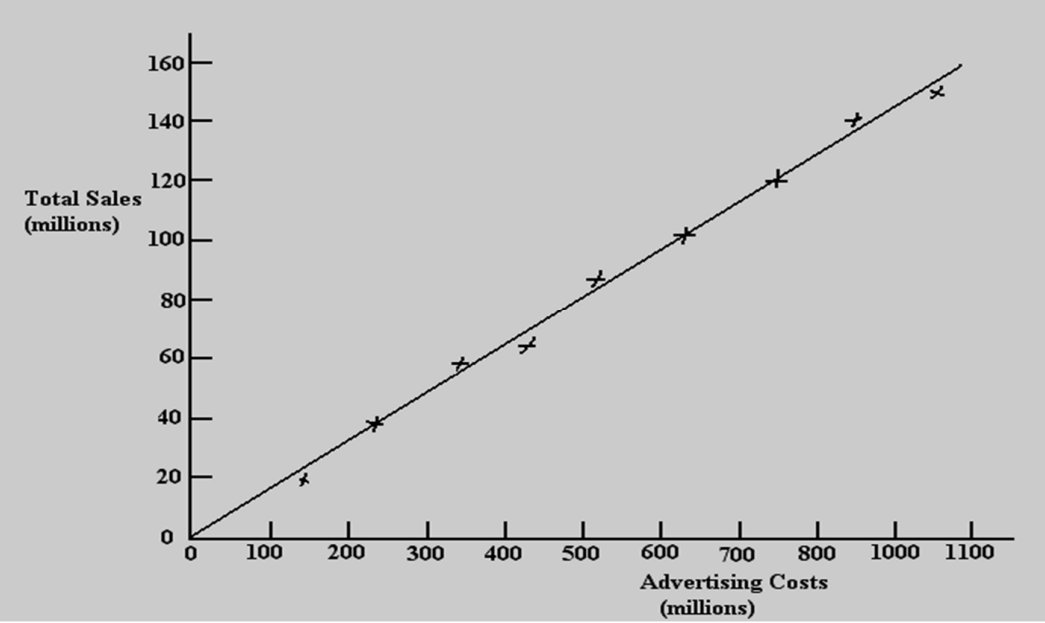
PART D
REGRESSION & CORRELATION

Chapter 17
REGRESSION ANALYSIS

17.1 Introduction

The methods of statistical analysis examined so far have been such that only one variable corresponding to each member of a particular universe has been considered, e.g. the marks obtained by a group of students in a certain subject or the age of a group of individuals. The statistical tests have been developed utilising the probability distribution of the particular single variable x. However it is good to note that one of the main tasks of management decision and business investigation is the analysis of the interrelations of the variables being investigated.

As frequently done in business administration, the interrelationships can be obtained by studying how one variable **x** changes with respect to another variable, **y**.
For example, as a manager of a company you may want to know how total costs are affected by an increase in output or how increase in sales has affected the increase in advertising costs.
The relationship between total sales and advertising costs will help the manager to make analysis about the movement of such a relation (i.e. whether the movement is positive or negative).



By plotting the two variables on a straight line graph we can obtain the nature of the relationship and which one variable can be deduced from another and vice versa.
Fig. 17.1above is an example of the graph of sales versus advertising costs.

It can be seen that the points lie almost on a straight line. By eye one can very easily draw a straight line that very nearly passes through all the points.

However one should note that there are cases where the relationship may not be clearly marked so that the relationship sought can only be approximated. Such cases are most likely to occur in instances where the variables concerned may be height and weight of N individuals from the N pairs of observations. One may be in need to determine the mathematical form of the relationship existing between the two variables. In this case the use of a scatter diagram becomes essential.

17.2: The Scatter diagram

Ten adults have their weight taken and their height measured. The results obtained are shown in Table 17.1 below.

Table 17.1: weight and heights of 10 adults

Adult	Weight (kg)	Height (ft.)
A	60	5.1
B	61	5.3
C	62	5.2
D	63	5.5
E	64	5.6
F	65	5.6
G	66	6
H	67	5.7
I	68	6.2
J	69	6.2

From such data one may want to ask – How does the height of an individual vary with his weight? One way to answer this question is to plot the above data in graphical form. By using coordinate paper and using suitable intervals we can plot on the y axis the height of the individuals and on the x axis their weight. The points are then marked off with say crosses as shown in Fig.13.2 below. The resulting figure is called *scatter diagram*.

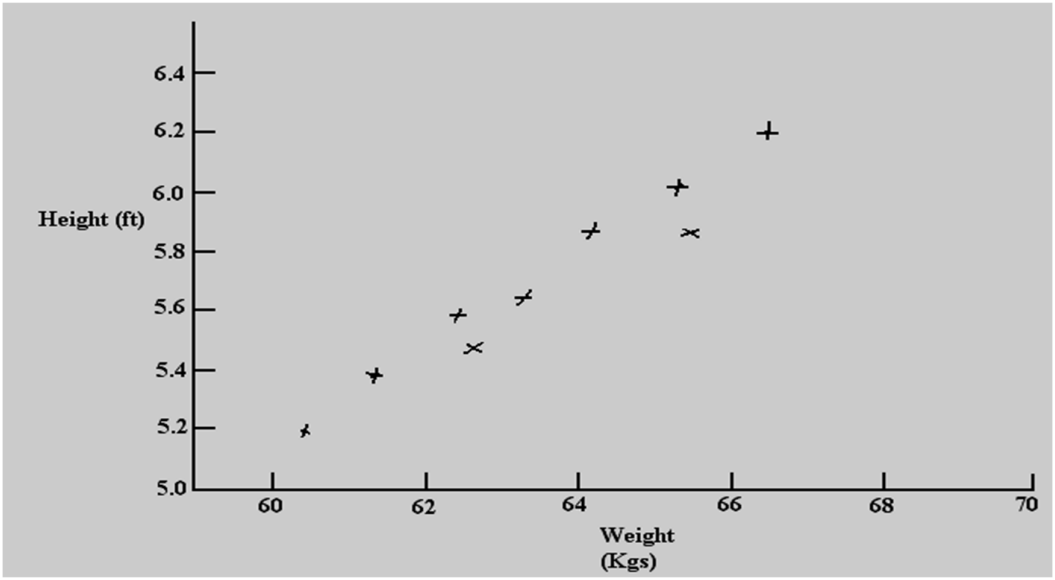


FIG.17.2 *Scatter diagram for table 17.1*

As we can see from the scatter diagram above the points show that there is an obvious linear relationship between the weight of an individual and his height. However, if one were to fit by freehand drawing to the set of points, then different lines could easily be obtained by different people. Hence in order to be consistent we should devise a mathematical technique by which line fitting to such data produces a line of "best" fit reproducible by everyone.

17.3 Curves of Regression

Under normal circumstances a relation between two variables has one of the variables being independent and the other being dependent. In a formula connecting two variables the subject is taken to be dependent variable, and when represented graphically, the independent variable is usually plotted along the "x-axis". However in most *bivariate*

distributions the same relation exists although either variable can be taken to be *dependent* or *independent*. This normally depends on what questions are being asked about the distribution in consideration.

Consider a bivariate distribution (x_1, y_1), (x_2, y_2), e.t.c. One may seek to find the formula for a curve which expresses x in terms of y or y in terms of x. In general these two formulae will not be the same. They can only be identical under the special circumstances when all pairs of the distributions are all identical. That is, the curve passes exactly through all the points. Therefore in practice we have two formulae.

In general terms, the equation which expresses y in terms of x is given by

$$y = a_1 + b_1x + c_1x^2 + ... \quad (17.1)$$

The one which expresses x in terms of y is given by

$$x = a_2 + b_2x + c_2x^2 + ... \quad (17.2)$$

Equation (17.1) is called the *equation of regression of y on x*. Equation (17.2) is called equation of regression of x on y. We call the resulting *curves of regression*. In the cases dealt with here, we are interested in linear relationships between two pairs of distributions. Hence strictly speaking we are interested in the *equation of regression of y on x* given by

$$y = m_1x + c_1 \quad (17.3)$$

And the equation of regression of x on y given by

$$x = m_2y + c_2 \quad (17.4)$$

Equations (17.3) and (17.4) are just special cases of equations (17.1) and (17.2) respectively.

Fig.17.3 *Below shows a line "best" fit drawn by eye*.

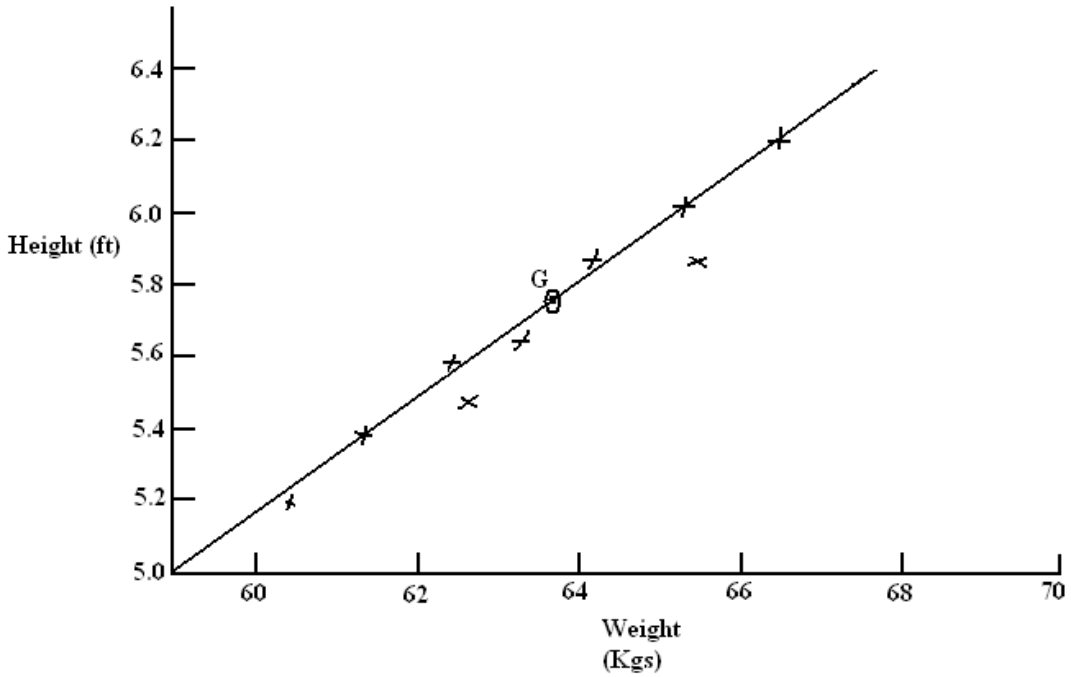


FIG. 17.3 *Line of best fit for the scatter diagram Fig.17.2*

The procedure for drawing a line of best fit is as follows: Find the mean centre, **G** of the estimated line of regression. This is the point G($\bar{x}$, $\bar{y}$), where$\bar{x}$ is the mean of the *x* values and $\bar{y}$ is the mean of the *y* values. Using a ruler, the line of regression is drawn to fit the plotted points as closely as possible while passing through **G**. The equation of the estimated line of regression can be obtained easily by first calculating its gradient (slope), *m*. Obtain this by using the co-ordinates of two points on the line which are far enough apart and then substitute the value of the gradient into the equation.

$$y - \bar{y} = m(x - \bar{x}) \qquad (17.5)$$

where $\bar{x}$ and $\bar{y}$ are as defined above.
Using this procedure the line of best fit of Fig.13.3 is found to be (depending on the person who draws the line)

$y = 0.125x - 2.42$

As we said above we must have a mathematical technique by which everyone should be able to reproduce a line of best fit. We shall now study this technique of how to obtain a line of best fit by calculation and then compare this equation with the one obtained by estimation.

17.4 Calculation of Regression Lines

Table 17.2 Calculation of regression lines

x	y	x^2	y^2	xy
60	5.1	3600	26.01	306.0
61	5.3	3721	28.09	323.3
62	5.2	3844	27.04	322.4
63	5.5	3969	30.25	346.5
64	5.6	4096	31.36	358.4
65	5.6	4225	31.36	364.0
66	6.0	4356	36.00	396.0
67	5.7	4489	32.49	381.9
68	6.2	4624	38.44	421.6
69	6.2	4761	38.44	427.8
$\sum$ x = 645	**$\sum$ y = 56.4**	**$\sum x^2$ = 41685**	**$\sum y^2$ = 319.48**	**$\sum xy$ = 3647.9**

In order to calculate a regression line we need a table with columns for x, y, x^2, xy as shown below. Consider the example of ten adults above. Consider the foregoing table.

$\bar{x} = \frac{645}{10} = 64.50$ *and* $\bar{y} = \frac{56.4}{10} = 5.64$

The gradient of the regression line of y on x is calculated from the equation

$$m_1 = \frac{\sum(xy) - n.\bar{x}.\bar{y}}{\sum x^2 - n.\bar{x}^2} \qquad (17.6)$$

The gradient of the regression line of x on y is calculated from the equation

$$m_2 = \frac{\sum(xy) - n.\bar{x}.\bar{y}}{\sum y^2 - n.\bar{y}^2} \qquad (17.7)$$

From the table above we have

$$m_1 = \frac{\sum(xy) - n.\bar{x}.\bar{y}}{\sum(x)^2 - n.\bar{x}^2} = \frac{3647.9 - 10 \times 64.5 \times 5.64}{41685.0 - 10 \times 64.5 \times 64.5}$$

$$= \frac{3647.9 - 3637.8}{41685.0 - 41602.5} = \frac{10.1}{82.5} = \mathbf{0.122}$$

Substituting for m_1 in equation (13.5) i.e. $\boldsymbol{y - \bar{y} = m(x - \bar{x})}$ we have

$y - 5.64 = 0.122(x - 64.5)$

$y - 5.64 = 0.122x - 7.87$

$y = 0.122x - 2.23$

Comparing with the estimated regression line one can see that the difference is very small. Hence the estimated regression line was quite accurate.

Example 17.2
The following table shows the number of units of a good produced and the total costs incurred.

Units	100	200	300	400	500	600	700
Total Cost Shs	40,000	45,00	50,000	65,000	70,000	70,000	80,000

Required

(i) *Calculate the regression line for Y on X.*
(ii) *Assuming **250** units were produced, compute the total costs incurred.*

SOLUTION

Determine the dependent variable, Y and which the independent variable X
For cost estimation, the values of the activities are the values of X and related costs are the values for Y.

X	**Y**	**XY**	**X^2**
100	40,000	4,000,000	10,000
200	45,000	9,000,000	40,000
300	50,000	15,000,000	90,000
400	65,000	26,000,000	160,000
500	70,000	35,000,000	250,000
600	70,000	42,000,000	360,000
700	80,000	56,000,000	490,000
2,800	**420,000**	**187,000,000**	**1,400,000**

Solution

From the general equation of the regression line ; $y - \bar{y} = m_1(x - \bar{x})$ *….regression of y on x.*

$$\bar{x} = \frac{2,800}{7} = \mathbf{400}\ and \bar{y} = \frac{420,000}{7} = \mathbf{60,000}$$

$$m_1 = \frac{\sum(xy) - n.\bar{x}.\bar{y}}{\sum(x)^2 - n.\bar{x}^2}$$

$$m_1 = \frac{187,000,000 - 7 \times 400 \times 60,000}{1,400,000 - 7 \times (400)^2} = \frac{19,000,000}{280,000} = \mathbf{67.86}$$

Substituting in the general equation of regression lines,

Y – 60,000 = 67.86 (x – 400)
Y – 60,000 = 67.86X – 27,144
Y = 67.86X – 27,144+60,000
Y = 67.86X + 32,856

(iii) If 250 units were produced then;
In this case, x = 250.

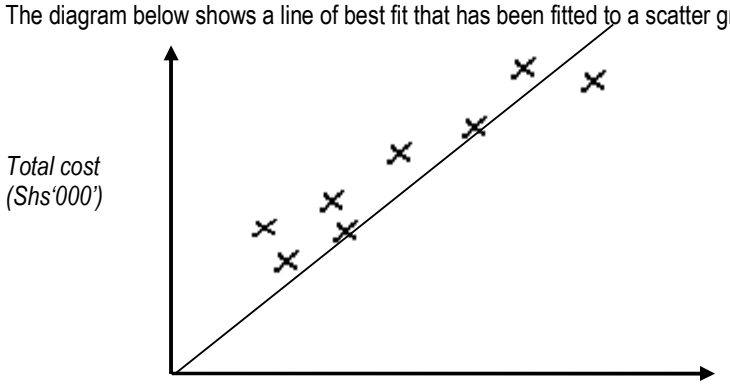
Therefore, substituting in the equation of regression, we shall have;

Y = (67.86 × 250) – 32,856
Y = 49,821
Thus the total cost were; **Shs. 49821.**

17.5: Application of line of best fit

A scatter graph can be used to make an estimate of fixed and variable costs by drawing a line of best fit through the band of points on a scatter graph, which best represents all the plotted points.

The diagram below shows a line of best fit that has been fitted to a scatter graph by 'eye' or judgment.



(a) Estimating fixed and variable costs using line of best fit

The amount of fixed costs can be found by looking at the point where the graph intercepts the Y-axis. In other words, look at total costs when activity (level of output) is zero. These must all be fixed costs.
To calculate variable costs, take any other point on your line of best fit and use these figures and your calculation of fixed costs to estimate variable cost per unit.

Example 17.3
In the previous diagram.
When X= 0, Y= 2000, therefore fixed costs are = Shs. 2000/-
Take point **P** on the line,
X = 5, Y = 6000Shs.
Total costs (TC) = 6000 Shs

TC = FC + VC (17.8)
VC = TC – FC
VC = 6000 -2000 = 4000Shs
Activity = 5 units
VC per unit = $\frac{4000}{5}$ = **800Shs**

17.6: INTERPOLATION AND EXTRAPOLATION

Regression lines can be used to calculate intermediate values of variables i.e. values within the known range. This is known as *linear interpolation* or simply interpolation. This is one of the main uses of regression lines.
It is also possible to extend regression lines beyond the range of values used in their calculation. It is then possible to calculate values of the variables that are outside the limits of the original data. This is known as *linear extrapolation.*
The problem with extrapolation is that it assumes that the relationship already calculated is still valid. This may or may not be so.
For example if output was increased outside the given range, there might come a point where economies of such a scale reduce costs and total costs might actually fall. So the graph as seen from the known range figure (a) might extend as in the figure (b).

Consider the previous example 2.

Units produced (X)	**Total costs** (Y)
100	40,000
200	45,000
300	50,000
400	65,000
500	70,000
600	70,000
700	80,000

Regression line of Y on X
Y= 32856 + 67.86x
Where x = 250
Y= 32856 + 67.86 × 250
Y = 49821

This is interpolation since x lies within the range of a given data i.e. between x = 200 and x = 300
Suppose x = 800
Y= 32856 + 67.86 × 800
Y = 87144

This is extrapolation since ***x*** does not lie within a range of the original data i.e. beyond x = 700

17.7: ANOTHER APPROACH FOR INTERPOLATION AND EXTRAPOLATION

Consider regression line **Y= a + bx**, drawn as follows; and take points A(X_1, Y_1) , B(X_2,Y_2) , C(X_3,Y_3)to lie on the line.

$$\boldsymbol{Gradient = \frac{Change\ in\ y}{Change\ in\ x}} \qquad (17.9)$$

Using points A and B

$$Gradient = \frac{y_2 - y_1}{x_2 - x_1}$$

But gradient is the same.

Thus

$$\boldsymbol{Gradient = \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_3 - y_1}{x_3 - x_1}} \qquad (17.10)$$

This expression can also be used.

NOTE: The method assumes that all points or values of original data lie on the regression line.

Summarized in the table below;

X	X_1	X_2	X_3
Y	Y_1	Y_2	Y_3

If X_2 or Y_2 are not known, but X_1, X_2, Y_1, Y_2 are known then obtain using interpolation.
Refer to previous example.

Activity (X)	200	250	300
Costs shs (Y)	45,000	—	50,000

Required;

a) Calculate Y when X = 250.
b) Calculate Y when X = 1,500

Solution;

(a) *This is linear interpolation since **X** lies within given range*
X_1= 200, X_2 = 250 ,X_3 = 300 Y_1 = 45,000 , Y_2 = ? , Y_3 = 50,000

Using;

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y_3 - y_1}{x_3 - x_1}$$

$$\frac{Y_2 - 45{,}000}{250 - 200} = \frac{50{,}000 - 45{,}000}{300 - 200}$$

$$\frac{Y_2 - 45{,}000}{50} = \frac{50{,}000 - 45{,}000}{100}$$

$Y_2 - 45{,}000 = 2{,}500$

$Y_2 = 45{,}000 + 2{,}500$ **= Shs. 47,500**

(b) *This is linear extrapolation since x = 1,500 does not lie in the given data.*

X	200	300	500
Y	45,000	50,000	—

X_1= 200, X_2 = 250 ,X_3 = 300 . Y_1 = 45,000 , Y_2 = 50,000, Y_3 =?

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y_3 - y_1}{x_3 - x_1}$$

$$\frac{50{,}000 - 45{,}000}{300 - 200} = \frac{Y_3 - 45{,}000}{1500 - 200}$$

$$\frac{5{,}000}{100} = \frac{Y_3 - 45{,}000}{1300}$$

$Y_3 - 45{,}000 = 65{,}000$

$Y_3 = 45{,}000 + 65{,}000$

Y_3 =110,000 Shs.

Chapter 18
CORRELATION ANALYSIS

18.1 Introduction

In the study of regression in the previous chapter we have examined how two variables can be related by two equations. These are the equations of regression of y on x and the equation of regression of x on y. In most situations, only one of the two regression lines is really necessary, and it is usually left to the investigator to decide which of the two lines is more important. Once this can be done, there is need of calculating the other one. Take the example of the marks in Mathematics and the marks in Physics given in the previous chapter.

Examining the two lines (Fig.17.4) reveals that the regression line y on x is the better of the two. This is to say that it is better to determine the physics marks from the mathematics marks instead of the other way round. This would seem to suggest that a high mark in physics correlates with a high mark in mathematics. Of course this suggestion is open to debate.

18.2 Definition

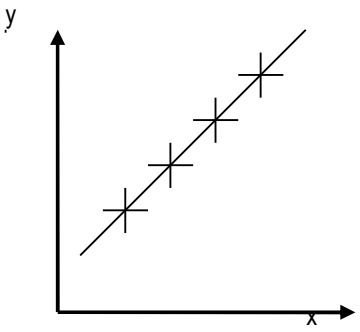
Correlation is defined as the degree of relationship between variables.

In correlation analysis we would like to determine *how well* the variables are related, in a quantitative sense. If the relationship between x and y is complete, then the equation describing this relationship is exactly linear and we say that the two variables are *perfectly correlated*. In the simplest case of only two variables we have *simple correlation*, while for more than two variables we have *multiple correlations*. Whether we have simple or multiple correlations, we can still have various kinds ofeach type.

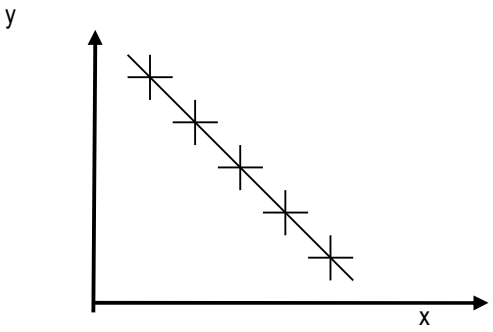
In ***Fig.18.1*** We have four scatter diagrams which show the kind of situation that can arise.

Examples of scatter graphs

(a) **Perfect positive linear**



(b) **Perfect negative linear**



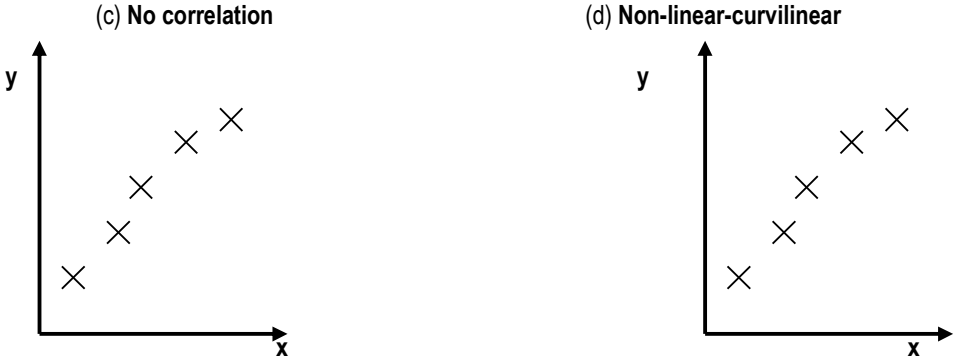


FIG.18.1 *Examples of different kinds of correlation*

(i) **Positive correlation** exists where the values of the variables increase together, for example when total costs arise as the level of activity arises.

(ii) **Negative correlation** exists where one variable increases as the other decreases in value.

Graph (a) is an example of perfect positive linear correlation since the points lie exactly on a straight line and as X increases so Y increases.

Graph (b) is an example of perfect negative linear correlation since the points again lie on a straight line, but as the X- values increase so the Y values decrease.

Graph (c) when the points are scattered all on the diagram, then there is little or no correlation between the two variables. We conclude that there is no cause and effect relationship between the two variables.

Graph (d) here the points lie on an obvious curve. There is a relationship between X and Y, but it is not a straight line relationship.

18.2 THE PRODUCT-MOMENT CORRELATION COEFFICIENT

If the correlation between x and y is perfect, then the two regression lines will merge into one and the slopes m_1 and m_2 of the two lines will be the reciprocal of each other. This means that the product of m_1 and m_2 is one. In other words, we have;

$$\mathbf{m_1.m_2 = 1}$$

$$\frac{\sum \boldsymbol{d_x d_y}}{\sum \boldsymbol{d_x^2}} . \frac{\sum \boldsymbol{d_x d_y}}{\sum \boldsymbol{d_y^2}} = \mathbf{1} \qquad (18.1)$$

Taking the square root we have;

$$\frac{\sum d_x d_y}{\sum d_x . \sum d_y} = \pm 1$$

When the quantity $\frac{\sum d_x d_y}{\sum d_x . \sum d_y}$ is calculated for any bivariate distribution, we obtain a number which lies between +1 and -1. This number is called the *product-moment correlation coefficient* and is normally represented by *r*. positive values of r indicate direct correlation while negative values indicate inverse correlation.
If ***r = 0***, then there is no correlation between ***x*** and **y**, although x and y could be associated in some other way, for example as in fig. 14.1 (d).

The product moment correlation coefficient is also known as "***Pearson's correlation coefficient***."

Example 18.1

Using the data already provided in example 2 of chapter 16.
However, we shall need to compute the value for y^2 and this has been done below;

X	Y	Y^2 (Shs'000')
100	40,000	1600,000
200	45,000	2,025,000
300	50,000	2,500,000
400	65,000	4,225,000
500	70,000	4,900,000
600	70,000	4,900,000
700	80,000	6,400,000
2,800	**420,000**	**26,550,000**

We also discovered that; $\mathbf{m_1 = 67.86}$., now we shall need to compute m_2 which is the gradient of the regression line of x on y.

Remember:- The regression line of x on y is given as;

$$m_2 = \frac{\sum(xy) - n.\bar{x}.\bar{y}}{\sum(y)^2 - n.y^2}$$

Whereby; (see previous example)
X =400
Y = 60,000
n = 7
∑xy = 187,000,000

Therefore by substitution;

$$\boldsymbol{m_2} = \frac{187{,}000{,}000 - 7 \times 400 \times 60{,}000}{26{,}550{,}000{,}000 - 7 \times (60{,}000)^2} = \frac{19{,}000{,}000}{1{,}350{,}000{,}000} = \mathbf{0.0141}$$

Therefore; the product-moment correlation coefficient, r, will be given as;

$= m_1 \times m_2$
$= 67.86 \times 0.0141$
$=$ **0.96 (2dps)**

$r = \sqrt{0.96} = \mathbf{0.978}$

This indicates a high degree of positive correlation.

18.3: Interpretation of results.

(i) The value of r obtained shows a very high degree of correlation because it is above 0.5 and very close to +1.

(ii) If r obtained was around +0.45, this would have indicated a fair positive correlation.

(iii) Don't forget that;

- If **r = +1**, there is perfect positive correlation.
- If **r = 0** , there is no correlation, and
- If **r = -1**, then there is perfect negative linear correlation.

18.4: Coefficient of determination

This is the square of the coefficient of correlation and so denoted by γ^2. it is a measure of how much of the variation in the dependent variable is explained by the variation of the independent variable. Where X is out put volume and Y is the total costs, the coefficient of the variations in total cost can be explained by variations in activity level. The variations not accounted for by variations in the independent variable will be to random fluctuations or to other specific factors.
In the previous example on output and costs, γ had a value of 0.977.

$$\gamma^2 = 0.977^2 = 0.955$$

$\gamma^2 \times 100 = 0.955 \times 100 = 95.5\%$

This indicates that the valuations in the number of units manufactured account for 95.5% of the variations in the total costs (good correlation).

If $\gamma = 0.42$, $\gamma^2 = 0.1764$

$100 \times \gamma^2 = 100 \times 0.1764 = 17.64\%$

So only 17.64% of the variations is explained by variations in the independent variable (poor correlation).

18.5: SPEARMAN'S RANK CORRELATION COEFFICIENT

The procedure for calculating the product-moment correlation coefficient is cumbersome and rather time-consuming. Instead of using the actual values, a method of ranking the data in order of size of importance will yield an approximate value of ***r*** in most cases. If we rank two variables *x* and *y*, we obtain the rank correlation coefficient which is the product-moment correlation coefficient of the ranks of the original data.

The rank-coefficient or the coefficient of rank correlation is given by;

$$\boldsymbol{R = 1 - \frac{6\sum D^2}{N(N^2 - 1)}} \qquad (18.2)$$

Where;
D = Difference between ranks of corresponding values of x and y.
N = number of pairs of values (x,y) in the data.

Example 18.2

The following table shows how 10 students, arranged in alphabetical order were ranked according to how they scored in both the laboratory and lecture parts of zoology course. Calculate spearman's coefficient of rank correlation for the data.

Lab.	8	3	9	2	7	10	4	6	1	5
Lect.	9	5	10	1	8	7	3	4	2	6

Solution

The difference in ranks **D** in the laboratory and lecture parts of the zoology course is given in the table below together with the calculated values of D^2.

Student	Rank(Lab.)	Rank(Lect.)	D	D^2
A	8	9	-1	1
B	3	5	-2	4
C	9	10	-1	1
D	2	1	1	1
E	7	8	-1	1
F	10	7	3	9
G	4	3	1	1
H	6	4	2	4
I	1	2	-1	1
J	5	6	-2	1
				$\sum D^2 = 24$

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

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

= **0.85**

Example 18.3

The marks of 10 students in arithmetic and Algebra are given in the table below;

Student	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Arithmetic	49	60	41	34	21	42	43	65	45	63
Algebra	39	74	33	32	13	31	34	71	57	40

Calculate Spearman's correlation between the performance in Arithmetic and algebra.

Solution

Student	Arithmetic	Rank	Algebra	Rank	D	D^2
A	49	4	39	5	-1	1
B	60	3	74	1	2	4
C	41	8	33	7	1	1
D	34	9	32	8	1	1
E	21	10	13	10	0	0
F	42	7	31	9	-2	4
G	43	6	34	6	0	0
H	65	1	71	2	-1	1
I	45	5	57	3	-2	4
J	63	2	40	4	-2	4
						$\sum D^2 = 20$

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

$$= 1 - \frac{6 \times 20}{10(10^2-1)}$$

= **0.88**

Example 18.4
In a drama competition, ten players were ranked by two judges as follows;

Player	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
$\mathbf{R_x}$	5	2	6	8	1	7	4	9	3	10
$\mathbf{R_y}$	1	7	6	10	4	5	3	8	2	9

You are required to calculate Spearman's rank correlation coefficient

Solution
Spearman's rank correlation coefficient

Player	$\mathbf{R_x}$	$\mathbf{R_y}$	**D**	$\mathbf{D^2}$
A	5	1	4	16
B	2	7	-5	25
C	6	6	0	0
D	8	10	-2	4
E	1	4	-3	9
F	7	5	2	4
G	4	3	1	1
H	9	8	1	1
I	3	2	1	1
J	10	9	1	1
				$\sum \mathbf{D^2 = 62}$

$$= 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6 \times 62}{10(10^2-1)} = \mathbf{0.62}$$

18.6: SIGNIFICANCE TESTS FOR CORRELATION COEFFICIENTS

Introduction
When the correlation coefficient ***r*** is calculated from sample data, i.e. correlation of sample paired values from an underlying population, the value of ***r*** obtained can be used to draw conclusions about the unknown population correlation coefficient, $\boldsymbol{\rho}$.

Remember that ***r*** is such that -1≤ r ≤ 1, where
r = -1 indicates perfect negative correlation
r = 0 indicates no correlation
r = 1 indicates perfect positive correlation

Illustration
If r, the correlation coefficient between X and Y is very close to zero then you would probably say that the two variables X and Y are not related at all. If r is very close to 1, for example r = 0.992, then you would probably say that there is a strong positive linear correlation between X and Y. For values of r such as 0.694 and – 0.5, a significant test would definitely be needed since it is difficult for one do draw conclusions about the degree of relation between the two variables X and Y.

Theorem: If two samples both from the same approximately normal distribution, then the test statistic;

$$\boldsymbol{t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}} \qquad (18.3)$$

has a ***t***-distribution with ***v***=*n-2* degrees of freedom

The hypothesis can be tested by comparing with a critical value of the t-distribution at a given level of significance.

The null hypothesis

In testing the significance of correlation coefficients, the null hypothesis is always that the correlation coefficient the two variables is zero i.e. there is no correlation between the variables.

The null hypothesis is stated as;

H_0: $\boldsymbol{\rho} = 0$*(There is no significant correlation)*

The alternative hypothesis

This depends on whether the test is one-tailed or two-tailed.

- **One-tailed tests**

If the hypothesis is that there is a *positive* correlation between the two variables (i.e. X and Y) then the alternative hypothesis will be stated as;

H_0: $\boldsymbol{\rho} > 0$*(The correlation is significantly greater than 0)*

If the hypothesis is that there is a *negative* correlation between the two variables (i.e. X and Y) then the alternative hypothesis will be stated as;

H_0: $\boldsymbol{\rho} < 0$*(The correlation is significantly less than 0)*

- **Two-tailed tests**

If one is looking for a correlation but not specifying whether it is positive or negative then the alternative hypothesis will be stated as;

H_1:$\boldsymbol{\rho} \neq 0$ *(There is a significant correlation between the two paired values)*

Example 18.5
The table below shows the monthly performance of the 10-outlets for Uniliver Uganda Ltd and the associated local advertising costs.

Outlet	Local advertising costs $\boldsymbol{x}$	Monthly sales **y**
Kampala	6	220
Masaka	8	230
Jinja	12	240
Kasese	12	340
Mbarara	2	420
Wakiso	8	460
Kisoro	16	520
Kabale	15	600
Bushenyi	14	720
Arua	20	800

Required

(i) Compute the product-moment correlation coefficient.
(ii) Test whether the correlation between sales and advertising costs is significant at the 5% level.

Solution

(a)**The product-moment correlation coefficient is given as**;

$$r = \sqrt{m_1 \times m_2}$$

Where;

$$m_1 = \frac{\sum(xy) - n.\bar{x}.\bar{y}}{\sum(x)^2 - n.\bar{x}^2}$$

$$m_2 = \frac{\sum(xy) - n.\bar{x}.\bar{y}}{\sum(y)^2 - n.\bar{y}^2}$$

Therefore we shall need the variables in the two equations which will be obtained from the table below;

x	y	xy	x^2	y^2
6	220	1320	36	48400
8	230	1840	64	52900
12	240	2880	144	57600
12	340	4080	144	115600
2	420	840	4	176400
8	460	3680	64	211600
16	520	8320	256	270400
15	600	9000	225	360000
14	720	10080	196	518400
20	800	16000	400	640000
∑x = 113	**∑y= 4,550**	**∑xy= 58,040**	**∑x²= 1,533**	**∑y²= 2,451,300**

$$\bar{x} = \frac{\sum x}{n} = \frac{113}{10} = 11.3 \ \ and \ \bar{y} = \frac{\sum y}{n} = \frac{4550}{10} = 455$$

Hence;

$$\boldsymbol{m_1} = \frac{58{,}040 - 10 \times 11.3 \times 455}{1533 - 10 \times (11.3)^2} = \frac{6{,}625}{256.1} = \mathbf{25.8688}$$

$$\boldsymbol{m_2} = \frac{58{,}040 - 10 \times 11.3 \times 455}{2{,}451{,}300 - 10 \times (455)^2} = \frac{6{,}625}{381{,}050} = \mathbf{0.017386}$$

Therefore the product moment correlation coefficient will be given as;

$r = \sqrt{m_1 \times m_2}$

$r = \sqrt{25.8688 \times 0.017386}$

$r = \sqrt{0.4497}$

$r = \mathbf{0.671} \ \ (3dps)$

(b) Hypothesis Test at 5.0% level

Null; H_0: $\boldsymbol{\rho} = 0$
Alternative; H_1:$\boldsymbol{\rho} \neq 0$

We further note that the above is a two tailed test!

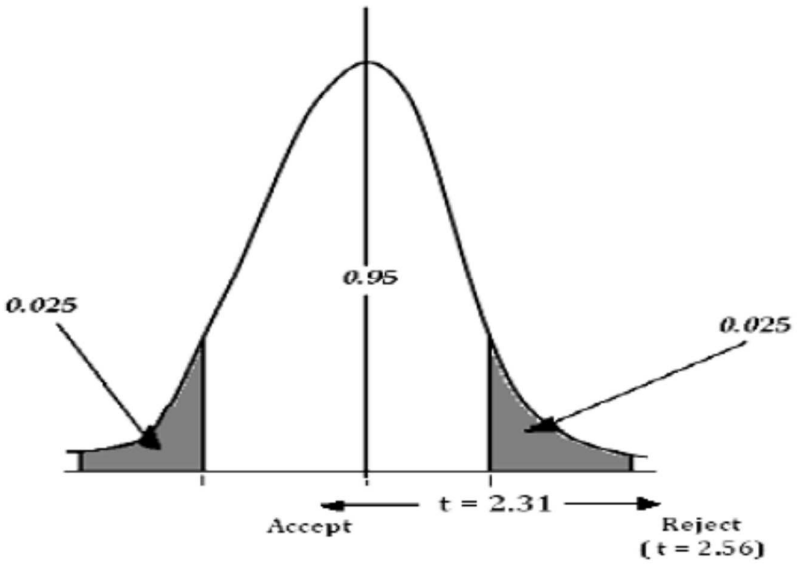
$t_{\alpha/2}$ = 0.05 and the test statistic is given as;

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Degrees of freedom; *v =10-2 = 8 degrees of freedom*

$$t = \frac{0.671 \times \sqrt{10-2}}{\sqrt{1-(0.671)^2}}$$

$$= \mathbf{2.56}$$



Since the calculated t-score is in the rejection zone, we shall reject the null hypothesis and accept the alternative hypothesis.

Therefore; the correlation between sales and advertising costs for the 10 outlets of Uniliver (u) Ltd is significant at the 5% level.

Example 18.6

(a) The job performance of ten employees at Uganda Martyrs University was ranked by two supervisors familiar with their work as follows;

	EMPLOYEE									
Supervisor	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
1	5	6	3	9	4	8	1	7	10	2
2	3	4	1	8	5	10	6	7	9	2

Required

(i) Compute Spearman's rank Correlation measure of the two sets of ratings.

(ii) Is it significantly greater than 0at the 2.5% level?

Solution

(a) Spearman's rank correlation measure

Employee	R_x	R_y	D	D^2
A	5	3	2	4
B	6	4	2	4
C	3	1	2	4
D	9	8	1	1
E	4	5	-1	1
F	8	10	-2	4
G	1	6	-5	25
H	7	7	0	0
I	10	9	1	1
J	2	2	0	0
				$\sum D^2 = 44$

$$= 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6 \times 44}{10(10^2-1)} = \frac{264}{990} = \mathbf{0.267}\ (3dps)$$

(b) Hypothesis Test at 5.0% level

Null; $H_0: \boldsymbol{\rho} = 0$
Alternative; $H_1: \boldsymbol{\rho} > 0$

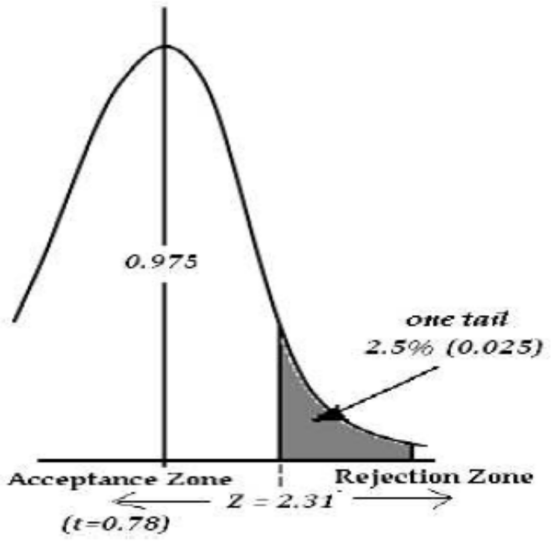
We further note that the above is a one tailed test!

$t_{\alpha/2}$ = 0.025 and the test statistic is given as;

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Degrees of freedom; *v =10-2 = 8 degrees of freedom*

$$t = \frac{0.267 \times \sqrt{10-2}}{\sqrt{1-(0.267)^2}} = \mathbf{0.78}$$



The calculated ***t*** score i.e.**0.78** is located in the Acceptance zone since it is less than 2.31, therefore $\mathbf{H_0}$ is accepted and $\mathbf{H_1}$ is rejected.

We shall then conclude that the calculated Spearman's Rank correlation is not significantly greater than 0 at the 2.5% level of significance

Example 18.7
A class of 15 students was examined in Mathematics and Physics. The following table gives the orders of merit of the students (arranged in alphabetical order) in both subjects. Calculate spearman's rank correlation and test whether the correlation between the performance in the two subjects is significant at the 1% level.

Maths	11	5	9	14	13	2	6	10	1	4	7	3	12	8	15
Physics	11	8	7	9	12	1	4	10	3	6	14	2	13	5	15

Solution

(a) Spearman's rank correlation coefficient

Student	Maths	Physics	D	D^2
A	11	11	0	0
B	5	8	-3	9
C	9	7	2	4
D	14	9	5	25
E	13	12	1	1
F	2	1	1	1
G	6	4	2	4
H	10	10	0	0
I	1	3	-2	4
J	4	6	-2	4
K	7	14	-7	49
L	3	2	1	1
M	12	13	-1	1
N	8	5	3	9
P	15	15	0	0
				$\sum D^2 = 12$

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

$$= 1 - \frac{6 \times 112}{15 \times 224}$$

= **0.80**

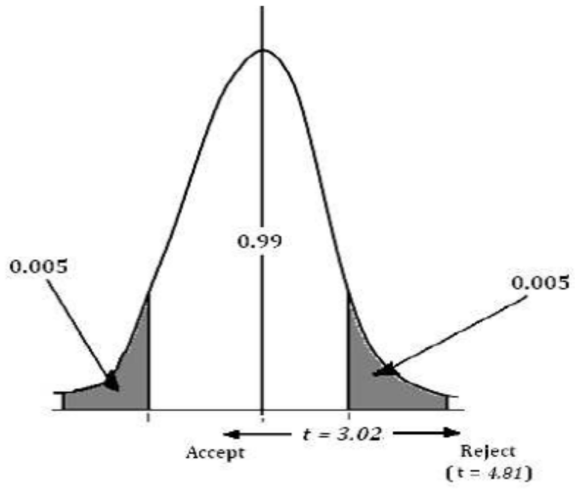
(b) Hypothesis Test at 1% level

Null; $H_0: \boldsymbol{\rho} = 0$
Alternative; $H_1: \boldsymbol{\rho} \neq 0$

We note that the above is a one tailed test.

$t_{\alpha/2}$ = 0.01 and the test statistic is given as;

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$



Degrees of freedom; *v =15-2 = 13 degrees of freedom*

$$t = \frac{0.80 \times \sqrt{15-2}}{\sqrt{1-(0.80)^2}}$$

$$= \mathbf{4.81}$$

The calculated ***t*** score i.e.**4.81** is located in the Rejection zone since it is greater than 3.02, therefore $\mathbf{H_0}$ is rejected and $\mathbf{H_1}$ is accepted.

We shall then conclude that the calculated Spearman's Rank correlation between the performance in the two subjects is significant at the 1% level

EXERCISE 14

14.1 Moses recorded the following costs for the past six months.

Month	**1**	**2**	**3**	**4**	**5**	**6**
Activity level (Units)	80	60	72	75	83	66
Total costs (Shs)	6,500	6,200	6,400	6,500	6,700	5,400

Required;

a) Establish the fixed costs per month and variable cost per unit using linear regression analysis.

b) Estimate costs for the following activity levels in a month.

(i) 750 units.

(ii) 90 units.

14.2 The table below shows marks given by two judges to ten teams in a music contest.

School	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Judge P	75	78	74	72	79	73	76	71	77	70
Judge Q	79	73	74	70	76	75	77	72	78	71

(i) Plot this information on a scatter diagram

(ii) Calculate the rank correlation coefficient

*(iii) Test whether **r** is significantly greater than 0 at the 5% level.*

14.3 Ten pairs of observations have been made on two random variables X and Y. The ten (x,y) values are (0,22); (-7,12); (-10,15); (-12,22); (-17,5); (-30,-5); (-32,-13); (10,30); (15,40); and (-12,8).

(a) Plot these results on a scatter diagram.

(b) Draw on a scatter diagram the lines of best fit for predicting Y from X and for predicting X from Y; identifying the lines clearly.

(c) State the coordinates of the points of intersection of the two lines.

(d) Estimate the expected value of Y corresponding to X = -7.

(e) Calculate the rank correlation coefficient for these data.

14.4 The costs of production in an operation have been obtained for nine time periods as follows.

Period	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Units"000"	2	1	0.5	6	8	4	3	4	6
Costs(Shs000)	12.9	9.9	22.6	25.0	32.4	18.3	15.8	19.2	25.2

Required;

a) Estimate the fixed costs per period and variable cost per unit using regression analysis.

b) Calculate the correlation coefficient.

14.5 (a)

(i) What does the regression coefficient represent?

(ii) Given a set of data for X and Y and the coefficient of Y on X and X on Y, is it possible that one coefficient has a positive value whereas the other has negative value?

(b) The table below shows marks awarded to **CPA** students in **QT** and **IT**

QT	57	58	59	59	60	61	62	64
IT	77	78	75	78	82	82	79	81

Required

(i) *Use the method of rank correlation to determine the relationship in performance between the two subjects.*

(ii) *Form two regression equations based on the above data.*

(iii) *Estimate the value of Y when X=65.*

14.6 The price of matooke is found to depend on the distance the market is away from the nearest town. The table below gives the average price of matooke for markets around Kampala city.

Distance, d (km)	40	8	17	20	24	30	10	28	16	36
Price, p(Shs)	120	160	140	130	135	125	150	130	145	125

(i) *Plot this data on a scatter diagram.*

(ii) *Draw the line of best fit on your diagram*

(iii) *Find the equation of your line in the form* $\boldsymbol{p = \alpha + \beta d}$ *where* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$ *are constants.*

(iv) *Estimate the price of matooke when* $\boldsymbol{d = 5}$.

14.7 The following shows the order in which six candidates were ranked in two tests X and Y.

X:	E	C	B	F	D	A
Y:	F	A	D	E	C	C

Calculate the coefficient of rank correlation and comment on your results.

14.8 The table below shows the marks obtained by candidates in two examinations

Exam A	15	20	28	36	47	48	51	66	75	77	80	81
Exam B	12	15	20	28	40	41	50	62	65	70	75	73

(i) *Plot a scatter graph for the data.*

(ii) *Draw a line of best-fit through the set of points on the graph. Hence find the Exam B mark for a candidate who scored 50 in Exam A.*

(iii) *Calculate the rank correlation coefficient for this data.*

PART E
FORECASTING & TIME SERIES ANALYSIS

Chapter 19
FORECASTING

19.1 Definition

Forecasting can be defined as a process of predicting or estimating future performances by looking at the current performance. Business data has certain characteristics which one needs to put into consideration in order to make accurate forecasts; these are discussed below;

19.2: Classification forecasting techniques

Forecasting techniques are classified into two;

(a) Quantitative techniques

Qualitative techniques are those that use statistical analysis to analyse past data of the item to be forecast.
These techniques rely on assumptions to enable use of data to forecast future performances.

Qualitative forecasting techniques include;

(i) Simple average method
(ii) Weighted average method
(iii) Moving average
(iv) Least squares
(v) Exponential smoothening

(b) Qualitative techniques

Qualitative techniques on the other hand are forecasting techniques which are used when data is scarce and no assumptions can be reliably made.
Therefore, these techniques employ human judgment and experience to turn qualitative information into quantitative estimates.
Qualitative methods of forecasting include;

(i) The Delphi method of forecasting
(ii) The market research method
(iii) Historical analog

19.3: FORECASTING TECHNIQUES

1. QUANTITATIVE TECHNIQUES (TIME SERIES MODELS)

There are many types of forecasting and these include;

(i) Simple averages
This refers to the method of averaging periodic data. It gives the same weight to each periodic data.
However, it is inappropriate in inflationary periods / seasons because in such situations, recent data should be given more weight to adjust recent values.

(ii) Weighted averages
This is an improvement on the simple average method. In this method, periodic data is given different weights.

(iii) Moving averages
This is another form of averaging and forms a link between several periodic data which makes it to adjust to recent values especially if the data periods involved in each average are many.
However, this method does not use all the available data and as such some information is unrepresented.

(iv) Least squares

This method obtains the average by obtaining the centre of data plotted on a graph. Its average becomes more accurate when the data is not scattered much otherwise it gives a misleading average.
However, it renders itself further to mathematical treatment because of its basic concept of minimizing deviations.

(v) Exponential smoothening

This is a better and desirable method. It has an in-built mechanism for assigning weights in a desirable way. It involves assigning small weights to older data and vice versa. The problem is that the initial weight assigned is done subjectively.

2. QUALITATIVE TECHNIQUES

(i) The Delphi method

This technique is mainly used for long term forecasting, designed to obtain expert consensus for a particular forecast, without the problem of submitting to pressure to conform to a majority view.
This method involves use of a panel of experts who independently answer a sequence of questionnaires in which the responses to one questionnaire are used to produce the next questionnaire. Thus any information available to some experts and not others is passed on to all, so that their subsequent judgments are refined as more information and experience becomes available.

(ii) The market research

This method uses opinion surveys, analyses of market data, questionnaires and other investigations to gauge the reaction of the market to a particular product, design, price etc.
Market research is often very accurate for the relatively short term, but longer term forecasts based purely on surveys are likely to be suspect because people's attitudes and intentions change.

(iii) Historical analogy

In this method, when past data on a particular item is not available, analysis is made on items similar to that under forecast and then such analyses are used to establish the life cycle and expected patterns of the item under consideration / forecast.
However, a lot of care is needed in using analogies which relate to similar items in different time periods!

19.4: TIME SERIES MODELS

19.4.1 Definitions

Time series is simply a process of collecting data about an event belonging to different time periods and then using such data to analyse the future behavior of such an event.
The time periods may be years, months, days, hours, e.t.c.

19.4.2 Importance of time series analysis

Time series is of practical importance to decision makers in the following ways;

(i) It helps in the analysis of past behaviour of the phenomenon under consideration.
(ii) It helps decision makers to predict or forecast the behaviour of the phenomenon under study in future, which is very essential for business planning.
(iii) It enables decision makers to make comparative studies of the values of different phenomenon under study over a given period of time.

19.4.3 Characteristics of business data that affect time series models

The following are the characteristics / features of business data;

(a) Trend characteristics

Business data normally follows a certain trend or pattern which is either falling or rising.
For example; prices rise because of inflation, demand increases, population growth and so on. Furthermore, infant mortality rates may fall because of improved health services.
Hence, such occurrences always make business information to oscillate up and down.

(b) Seasonal characteristics

Seasonal patterns in business data can be observed every after a regular interval which is usually less than a year. For example, there are more purchases of household consumer goods at weekends because that is the time members of households have time to shop for the week. More blankets are sold in winter because of coldness.

(c) Cyclical characteristics

Cyclical patterns in business data can be observed after a long-term period such as period more than one year.
Cyclical patterns usually reflect periods of a boom, depression or recovery. Because of the multiplicity of factors that bring about these cycles, it is very difficult to predict such characteristics / patterns. For example, a depression may occur because of unprecedented increase in petrol prices or stock price crush. This characteristic also does not have a fixed-period pattern.

(d) Random characteristics

Random characteristics in business data are difficult to isolate because they do not have a fixed pattern such as those possessed by the above characteristics.
These characteristics happen randomly and do not have a fixed period of occurrence, for example, a fall in output due to a cargo train accident or to any unusual causes cannot be predicted. These causes are not generally known.

19.4.4 Application of forecasting techniques in time series analysis

We shall now examine how the quantitative methods of forecasting are used in time series analysis.

(a) **Weighted average method**

This method is quite easy since as the name suggests, we shall need to get the average of the time series data and then assign forecasted figures for each period.
The weighted average method encompasses the simple average method as well, therefore we shall not necessarily make a distinction between the two methods in the example that follow.

However, one should note that the weighted average can be divided into semi-weighted average (where data is divided into two equal portions), quarter-weighted average (where data is divided into 4 equal portions) and so much more according to the way one needs to proportionally divide their data for forecasting purposes.

As we can observe, this method is easy but inappropriate for periods with fluctuating sales.

(b) **Least squares method**

As discussed before, this method applies an appropriate mathematical equation for the trend selected, and the constants that appear there are determined by the principle of least squares.
Provisions are made for linear($\boldsymbol{y = a + bx}$) , parabolic($\boldsymbol{y = ax^2 + bx + c}$), and exponential ($\boldsymbol{y = ab^x}$) relationships.

- **Linear Relationships**

In this method, linear trend relationships are obtained through the general linear trend equation;

$$\boldsymbol{y = a + bx} \qquad (19.1)$$

The unknowns **a** and **b** are obtained by solving two normal equations (i) and (ii) shown below simultaneously;

$$\sum y = an + b\sum x \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (i)$$
$$\sum xy = a\sum x + b\sum x^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (ii)$$

Example 19.1

The table below shows the annual sales in millions of shillings for *ABC Ltd* over a period of nine years. You are required to forecast the sales for year 10

Year	1	2	3	4	5	6	7	8	9
Sales	220	140	108	210	122	242	246	160	260

Solution

The variables in the least squares equations (i) and (ii) above will be obtained as shown in the table below;

x	y	xy	x^2
1	220	220	1
2	140	280	4
3	108	324	9
4	210	840	16
5	122	610	25
6	242	452	36
7	246	1722	49
8	160	1280	64
9	260	2340	81
∑x = 45	**∑y= 1,708**	**∑xy=8,068**	**∑x²=285**

(a) Computing the trend line

Using the general normal trend equation $\boldsymbol{y = a + bx}$with two normal equations;

$\sum y = an + b\sum x$.....*(i)*

$\sum xy = a\sum x + b\sum x^2$.....*(ii)*

In the above table, $\boldsymbol{x}$ represents the years while ***y*** represents the annual profits.

Therefore, by substitution

$1708 = a(9) + b(45)$ .....(1)

$\underline{8068 = a(45) + b(285)}$ .....(2)

$1708 = 9a + 45b$ $\times 45$

$\underline{8068 = 45a + 285b}$ $\times 9$ } $elimination\ of\ unknown\ variable\ \boldsymbol{a}.$

$76860 = 405a + 2025b$

$\underline{-72612 = 405a + 2565b}$ } $at\ this\ level\ subtract\ eqn\ (2) from\ (1) to\ eliminate\ \boldsymbol{a}.$

$4248 = 0 + -540b$

$$\frac{-540b}{540} = \frac{4248}{40}$$

b = -7.867

Unknown a will be obtained by substituting the value **b = -7.867**in any of the equations (1) or (2). Therefore by substituting in equation 1;

$$1708 = a(9) + 45(-7.867)$$

From which **a = 229.1**

Therefore the trend line will be given as;

$$\boldsymbol{y = 229.1 - 7.869x}$$

(b) <u>Forecasted sales for year 10</u>

Forecasted sales for year 10 are obtained by substituting the value **x=10** in the trend equation.

Hence; $y = 229.1 - 7.869(10) = 229.1 - 78.69 = \boldsymbol{150.41}$

Example 19.2
The table below shows the profit figures in millions of shillings of Seya and Sons Ltd over a period of 10 years.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Sales	10	20	12	15	24	13	13	23	8	14

(where*1999* represents *year 1*, *2000* represents *year 2* and so on)
Required

(i) Determine the trend line from the above data
(ii) Interpret each variable in the trend line
(iii) Compute the monthly growth rate in profits.
(iv) Forecast profits for 2009 and 2010.

Solution

(i) **Trend line**

The trend line will be determined as before; in this regard, we shall arrange the required variables as in the table below;

Years	x	**y**	xy	x^2
1999	1	10	10	1
2000	2	20	40	4
2001	3	12	36	9
2002	4	15	60	16
2003	5	24	120	25
2004	6	13	78	36
2005	7	13	91	49
2006	8	23	184	64
2007	9	18	162	81
2008	10	14	140	100
	$\sum$x = 55	**$\sum$y= 162**	**$\sum$xy=921**	**$\sum x^2$=385**

Using the general normal trend equation $\boldsymbol{y = a + bx}$with two normal equations;

$\sum y = an + b\sum x$*……(i)*

$\sum xy = a\sum x + b\sum x^2$*……(ii)*

In the above table, $\boldsymbol{x}$ represents the years while $\boldsymbol{y}$ represents the annual profits.
Therefore, by substitution

$$162 = a(10) + b(55) \quad \text{……(1)}$$
$$\underline{921 = a(55) + b(385)} \quad \text{……(2)}$$
$$\left.\begin{matrix} 162 = 10a + 55b \\ \underline{921 = 55a + 385b} \end{matrix}\right\} \begin{matrix} \times 55 \\ \times 10 \end{matrix} \} \; elimination\ of\ unknown\ variable\ \boldsymbol{a}.$$
$$\left.\begin{matrix} 8910 = 550a + 3025b \\ \underline{-9210 = 550a + 3850b} \end{matrix}\right\} \; at\ this\ level\ subtract\ eqn\ (2) from\ (1) to\ eliminate\ \boldsymbol{a}.$$
$$-300 = 0 + -83b$$

$$\frac{-83b}{83} = \frac{-300}{83}$$

b = 3.614

Unknown a will be obtained by substituting the value **b = 3.614** in any of the equations (1) or (2). Therefore by substituting in equation 1;

$$162 = a(10) + 55(3.614)$$ From which **a = 3.677**

Therefore the trend line will be given as;

$$\boldsymbol{y = 3.677 + 3.614x}$$

(ii) **Interpretations**

The value a = 3.677 million is the y-intercept of the trend line equation, the value of b = 3.614 million is the slope of the trend equation, which is also the annual growth rate in the profits.

(iii)

$$Monthly\ growth\ rate\ in\ profits = \frac{3.614}{12} = Shs.0.3\ million$$

(iv) **Forecasted sales for year 2009 and 2010**

- **Year 2009 (this represents year 11)**

$$y_{2009} = 3.677 + 3.614(11) = Shs.43.43\ million$$

- **Year 2010 (this represents year 12)**

$$y_{2010} = 3.677 + 3.614(12) = Shs.47.04\ million$$

- **Parabolic relationships**

The general equation of a regression line for exponential relationships is given by;

$$\boldsymbol{y = ax^2 + bx + c} \qquad (19.2)$$

This relationship is obtained by solving the three normal equations shown below simultaneously; i.e.

$$\sum y = a\sum x^2 + b\sum x + nc \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (i)$$

$$\sum xy = a\sum x^3 + b\sum x^2 + c\sum x \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (ii)$$

$$\sum x^2y = a\sum x^4 + b\sum x^3 + c\sum x^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (iii)$$

Where a, b and c are constants and as observed in linear relationships.

- **Exponential relationships**

The general equation of a regression line for parabolic relationships is given by;

$$\boldsymbol{y = ab^x} \qquad (19.3)$$

This relationship is obtained by solving the three normal equations shown below simultaneously; i.e.

$$a = antilog\frac{[\sum \log y]}{n} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (i)$$

$$b = antilog\frac{[\sum x \log y]}{\sum x^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (ii)$$

Where a, and b are constants as observed in linear relationships

(c) **Moving average method**

This is a forecasting technique which works on the principle that the average forecast for the next period in time is an average of several preceding periods.
It's a method which tries to smooth away any random fluctuations within the periodic data.

The moving averages computed are taken to be the trend values showing the direction of the trend in seasonal data and are later used when isolating seasonal variations from time series.

Example 19.3
The table below shows the demand in thousands of watches for Jama Jewelries Ltd over a period of 10 years.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Demand	100	250	162	155	204	130	103	123	800	814

Required
Compute a 3-year and 5 year moving average for the above demand figures

Solution

Years	***Demand***	***3-year moving total***	***3-year moving average***	***5-year moving total***	***5-year moving average***
1999	100				
2000	250				
2001	162				
2002	155	512	171		
2003	204	567	189		
2004	130	521	174	871	174
2005	103	489	163	901	180
2006	123	437	146	754	151
2007	800	356	119	715	143
2008	814	1,026	342	1,360	272

Workings;
(i) 3 year moving average for **2002**
= (1999 demand + 2000 demand + 2001 demand)÷ 3
=100 + 250 + 162 = **512** ÷ 3**= 171***(Placed against 2002)*

(ii) 3 year moving average for **2003**
= (2000 demand + 2001 demand + 2002 demand)÷ 3
=250 + 162 + 155 =**567**÷ 3 **= 189***(Placed against 2003)*

(iii) 3 year moving average for **2004**
= (2001 demand + 2002 demand + 2003 demand)÷ 3
= 162 + 155 + 204 =**521**÷ 3 **= 174** *(Placed against 2004)*

(iv) 3 year moving average for **2005**
= (2002 demand + 2003 demand + 2004 demand)÷ 3
= 155 + 204 + 130 =**489**÷ 3 **= 163** *(Placed against 2005)*

(v) 5 year moving average for **2004**
= (1999 demand + 2000 demand + 2001 demand+ 2002 demand+ 2003 demand)÷ 5
= 100+250+162+155+204 = **871**÷5=**174***(Placed against 2004)*

(vi) 5 year moving average for **2005**
= (2000 demand + 2001 demand+ 2002 demand+ 2003 demand+ 2004 demand)÷ 5
= 250+162+155+204+130 = **901**÷5=**180***(Placed against 2005)*

In summary, the moving totals are computed by summing the 3 year preceding results. To obtain the moving average, divide the moving total by 3. The moving average is then placed against the fourth year. The moving total of the following year is obtained by summing the next 3 results after skipping one of the oldest year results.
This applies to the 5-year moving averages and others as may be required.

Compute the remaining moving totals and averages to enhance your understanding!

Advantages of moving average method

(i) It is applicable to changing circumstances
(ii) It is simple to use as compared to the least square's method that requires a lot of mathematical models.

Limitations of the moving average method

(a) The method eliminates trend values, which relate to the extreme periods of time.
(b) The method cannot be used to predict or forecast future values which is one of the main objectives of time series / trend analysis.

(d) **Exponential smoothening method**

This is a method of forecasting that involves the use a smoothening constant, α, on past data. The weights decrease exponentially with time with the most current values receiving the greatest weighting while the oldest observations receives a decreasing weighting.

The exponential smoothening method greatly overcomes the limitations encountered in the above discussed moving average method.

To compute new periodic forecast the method uses the model shown below;

New forecast = Old forecast + α (Latest observation – old forecast) (19.4)

Example 19.4

The table below shows the profits in thousands of Capital Shoppers over a period of 10 years.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Profits	450	440	460	410	380	400	370	360	410	450

Required

Compute the periodic forecasted profits for capital Shopper using 0.1 and 0.5 smoothening constants.

		Exponential Forecasts	
Years	$\boldsymbol{Profits}$	$\boldsymbol{\alpha = 0.1}$	$\boldsymbol{\alpha = 0.5}$
1999	450		
2000	440	450	450
2001	460	449	445
2002	410	450.10	452.50
2003	380	445.69	431.25
2004	400	439.12	405.63
2005	370	435.21	402.82
2006	360	428.69	386.41
2007	410	421.82	373.21
2008	450	423.57	391.61

The forecast figures using the $\alpha = 0.1$ was computed as below;
Using;

2000 forecast = 1999 profit results
= **450**

2001 forecast = Old forecast + α(Latest observation – old forecast)

= 450 + 0.1(440 – 450)

= **449**

2002 forecast = Old forecast + α(Latest observation – old forecast)

= 449 + 0.1(460 – 449)

= **450.1**

Similar logic applies to exponential forecasts where α = 0.5.

19.5: TIME SERIES DECOMPOSITION

19.5.1 Introduction

The time series models i.e. moving average, exponential smoothening and least squares mentioned among others discussed in the previous section have been illustrated using data that is not seasonally adjusted to include factors such as inflation in prices, trend, seasonal, random and other variations.

As mentioned before, business data is affected by factors such as;

*(i) Trend variations (**T**)*
*(ii) Seasonal variations (**S**)*
*(iii) Cyclical variations (**C**)*
*(iv) Random variations (**R**)*

Therefore, to make reasonably accurate forecasts, it is necessary to make adjustments to business data to make provisions for the above 4 factors.

19.5.2 Definition

Decomposition of time series data is a process of ensuring that time series forecasts made are accurately computed by making adjustments for tread, cyclical and random variations.

19.5.3 Methods used in time series decomposition

There are 2 main models used in time series decomposition; i.e.

(i) Additive model and
(ii) Multiplicative model.

These are explained below;

(i) Additive model

This model assumes that the effects of the various factors components i.e. T, S, C and R of a time series are additive in nature. i.e;

$$\beta = T + S + C + R \quad (19.5)$$

where S, C, and R are expressed in absolute terms.

(ii) Multiplicative model

This model assumes that the effects of time series variations i.e. trend, seasonal, cyclical and random are not necessarily independent of each other but they are inter-related. i.e.

$$\beta = T \times S \times C \times R \quad (19.6)$$

Where: S, C, and R are expressed as percentages or proportions.

This is a preferred model since it adheres to most time series in the field of business and economics which are affected by many factors that are inter-dependent of each other as assumed by the model.

19.5.4 Deseasonalising time series data

This is the process of separating out the time series various from the above models to determine how each individual factor affects business data and how the two are interrelated.

For purposes of the syllabus, we shall examine how the trend variations and seasonal variations affect business data in isolation.

Measure of seasonal variations and seasonal indices

Definitions

(a) **Seasonal variations** are the short-term oscillations in the time series data which occur regularly in a definite season.
(b) **Seasonal indices** refer to the percentage of the seasonal averages of the seasonal data to the overall of the seasonal averages.

A seasonal index shows whether there is a fall or rise in data / performance. For example; if quarterly profits have an index of 107, it means profits increases by 7% in that particular quarter.

Measurement of seasonal variations and indices

Seasonal variations are measured using the additive model whereas seasonal indices are measured using the multiplicative model.

To measure variations and seasonal indices, the time series data is split into quarters and hence we obtain quarterly averages for each periodic time.

Example 19.5

The table below shows the profits in millions of shillings made by ABC Ltd for each quarter of the year from 2000 to 2003.

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2000	97	106	96	100
2001	101	104	80	118
2002	116	111	93	114
2003	120	101	101	106

Required;

(i) Compute the seasonal variations
(ii) Compute the seasonal indices

Solution

(i) Seasonal variations

Step 1:- We obtain the average of each quarter.

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2000	96	106	97	100
2001	80	104	101	118
2002	93	111	116	114
2003	101	101	120	106
TOTAL	**370**	**422**	**434**	**438**
Average	**92.5**	**105.5**	**108.5**	**109.5**

Step 2:- Obtain the average of all averages.

$$\text{Average of averages} = \frac{92.5 + 105.5 + 108.5 + 109.5}{4} = \mathbf{104}$$

Step 3:- The seasonal variations of each quarter is the deviation of each quarterly average from the average of averages.

Quarters	**Seasonal variations (ADDITIVE MODEL)**
Quarter 1	92.5 – 104 = **-11.5**
Quarter 2	105.5 – 104 = **1.5**
Quarter 3	108.5 – 104 = **4.5**
Quarter 4	109.5 – 104 = **5.5**

(ii) Seasonal indices

Here we simply get the percentage variations of quarterly averages above using the multiplicative model as shown below;

Quarters	**Seasonal indices (MULTIPLICATIVE MODEL)**
Quarter 1	$\frac{92.5}{104} \times 100 = 89\%$
Quarter 2	$\frac{105.5}{104} \times 100 = 101\%$
Quarter 3	$\frac{08.5}{104} \times 100 = 104\%$
Quarter 4	$\frac{109.5}{104} \times 100 = 105\%$

19.5.5 Measurement of Trend variations and indices

Trend variations are obtained by getting a regression line between two variables X and Y. (see regressions analysis). The general formula used to express trend variations per quarter is;

$$\boldsymbol{y = a + bx}$$

19.6 FORECASTING USING TREND AND SEASONAL VARIATIONS

In this case, seasonal variations will be computed by first determining the trend estimates using the above general regression equation.
All trend estimates for quarterly variations are obtained by substituting the respect quarter number into the trend equation.

Trend variations are obtained for each quarter using the formula below;

$$\frac{\boldsymbol{Actual\ Sales}}{\boldsymbol{Trend\ estimates}} \times \mathbf{100} \qquad (19.8)$$

The quarterly indices are then obtained by getting the average of all annual /yearly seasonal variations.

To illustrate this, two methods shall be adapted; i.e;

(i) the regression / least squares method and
(ii) the moving average method;

Example 19.6(*least squares method*);

The table below shows sales in thousands made by *Mukwano Industries* for one of its products Nomi-washing detergent for the period 2005 to 2008.

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2005	20	32	62	29
2006	21	42	75	31
2007	23	39	77	48
2008	27	39	92	53

Required

(i) Compute the trend line
(ii) Estimate the sales for each quarter.
(iii) Calculate the percentage seasonal variations of each quarter's actual sales from the estimates and the average seasonal variations.
(iv) Compute the average seasonal variations / indices.
(v) Compute the sales forecast.
(vi) Forecast sales for the year 2009

Solution

(i) The trend line

The trend line will be obtained using the least squares method. We shall therefore need to obtain the variables in the general and normal equations using the table below;

Year	$x\ (Quarters)$	$y\ (Sales)$	xy	x^2
2005	1	20	20	1
	2	32	64	4
	3	62	186	9
	4	29	116	16
2006	5	21	105	25
	6	42	252	36
	7	75	525	49
	8	31	248	64

2007	9	23	207	81
	10	39	390	100
	11	77	847	121
	12	48	576	144
2008	13	27	351	169
	14	39	546	196
	15	92	1,380	225
	16	53	848	256
Totals	$\sum x = 136$	$\sum y = 710$	$\sum xy = 6{,}661$	$\sum x^2 = 1{,}496$

Note: - There is continuity in assigning values to yearly quarters i.e. Quarter 1 for 2005 is assigned 1, Quarter 1 for 2006 is 5, Quarter 1 for 2007 is 9, and so on. The reader should note that 1, 5, 9 are all representative values for Quarter 1 in the respective years!

By substituting in the least squares equations, we have;

$$\sum y = an + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

$710 = 16a + 136b \quad \times 136$.....(1)

$6{,}661 = 136a + 1{,}496b \quad \times 16$ .....(2)

$96560 = 2176a + 18496b$

$-106576 = 2176a + 23936b$

$-10{,}016 = 0 + -5440b$

$$\frac{-5{,}440b}{5{,}440} = \frac{-10{,}016}{5440}$$

b = 1.84

Substituting in any of the equations (1) or (2), we obtain the value for unknown **a** as; **a = 28.74**

Therefore the trend line equation is given as;

$$\boldsymbol{y = 28.74 + 1.84x}$$

(ii) Estimated sales

Estimated sales are the trend variations obtained by substituting yearly quarter values into the above trend line equation.

For example;

2005 quarter 1 sales

$y = 28.74 + 1.84x = 28.74 + 1.84(1) = \mathbf{30.58}$

2006 quarter 1 sales

$y = 28.74 + 1.84x = 28.74 + 1.84(5) = \mathbf{37.94}$

2007 quarter 1 sales

$y = 28.74 + 1.84x = 28.74 + 1.84(9) = \mathbf{37.94}$

2008 quarter 1 sales

$y = 28.74 + 1.84x = 28.74 + 1.84(13) = \mathbf{52.66}$

You can note that we have simply substituted the quarterly values in the place of $\boldsymbol{x}$ to obtain the estimated / trend sales variations. All other trend estimates have been included in the table shown below.

(iii) Percentage Seasonal variations

Percentage seasonal variations are computed using equation (15) above; for example;

- *Seasonal variations for 2005 (Quarter 1)*

$$= \frac{Actual\ Sales}{Trend\ estimates} \times 100 = \frac{20}{30.58} \times 100 = \mathbf{65.0}\%$$

- *Seasonal variations for 2006 (Quarter 1)*

$$= \frac{Actual\ Sales}{Trend\ estimates} \times 100 = \frac{21}{37.94} \times 100 = \mathbf{55.0}\%$$

All other seasonal variations have been included in the table below;

Year	$\boldsymbol{x}$ $(\boldsymbol{Quarters})$	$\boldsymbol{y}$ $(\boldsymbol{Sales})$	$\boldsymbol{Trend\ estimates}$	Seasonal variations
2005	1	20	30.58	65
	2	32	32.42	99
	3	62	34.26	181
	4	29	36.10	80
2006	5	21	37.94	55
	6	42	39.78	106
	7	75	41.62	180
	8	31	43.46	71
2007	9	23	45.30	51
	10	39	47.14	83
	11	77	48.98	157
	12	48	50.82	94
2008	13	27	52.66	51
	14	39	54.50	72
	15	92	56.34	163
	16	53	58.18	91

(iv) Average seasonal variations / indices

Year	Quarter 1 %	Quarter 2 %	Quarter 3 %	Quarter 4 %
2005	65	99	181	80
2006	55	106	180	71
2007	51	83	157	94
2008	51	72	163	91
Totals	**222**	**360**	**681**	**336**
Averages(indices)	**56%**	**90%**	**170%**	**84%**

(v) Sales forecast (Multiplicative model)

The above average seasonal averages are used to forecast the yearly quarterly sales as shown in the table below, using the formula below;

Seasonally adjusted forecast = Trend estimate × seasonal variation (19.9)

Year	x $(Quarters)$	y $(Sales)$	$Trend\ estimates$	Seasonal Indices %	Seasonally adjusted forecasts
2005	1	20	30.58	56	17.12
	2	32	32.42	90	29.18
	3	62	34.26	170	58.24
	4	29	36.10	84	30.32
2006	5	21	37.94	56	21.24
	6	42	39.78	90	35.80
	7	75	41.62	170	70.74
	8	31	43.46	84	36.51
2007	9	23	45.30	56	25.37
	10	39	47.14	90	42.43
	11	77	48.98	170	83.27
	12	48	50.82	84	42.69
2008	13	27	52.66	56	29.49
	14	39	54.50	90	49.05
	15	92	56.34	170	95.78
	16	53	58.18	84	48.87

(vi) Sales forecast for the year 2009

Once the trend line and average seasonal indices has been determined, they can be used to forecast future sales. This is what is known as *extrapolation of seasonal variations* (in this case, extrapolation of sales)

The sales trend estimates and forecast for 2009 will be computed as below;

	x $(Quarters)$	$Trend\ estimates$	Seasonal Indices %	Seasonally adjusted forecasts
2009	17	60.02	56	33.61
	18	55.67	90	10.02
	19	108.29	170	20.58
	20	55.05	84	46.24

Example 19.7 (*Moving average method*)

Rework example 19.6 above using 3-point moving average method.

Solution

In this solution only relevant steps have been shown.
It is worth noting that this method is more robust and stable to any changes in seasonal data.

Year	x	y	3-point moving sales	$Trend\ estimates$	Seasonal variations %	Seasonal Indices %	Seasonally adjusted forecasts
2005	1	20		34.38	58	54	18.56
	2	32	38.0	35.70	90	89	31.77
	3	62	41.0	37.02	167	170	62.93
	4	29	37.3	38.34	76	86	33.02
2006	5	21	30.7	39.66	53	54	21.42
	6	42	46.0	40.98	102	89	36.47
	7	75	49.3	42.30	177	170	71.91
	8	31	43.0	43.62	71	86	37.51
2007	9	23	31.0	44.94	51	54	24.27
	10	39	46.3	46.26	84	89	41.17
	11	77	54.7	47.58	62	170	80.89
	12	48	50.7	48.90	98	86	42.05
2008	13	27	38.0	50.22	54	54	27.12
	14	39	52.7	51.54	76	89	45.87
	15	92	61.3	52.86	174	170	89.86
	16	53		54.18	98	86	46.59

Workings

- **Trend line equation** when computed in the normal way using the 3-point moving average sales its found to be;

$$y = 33.06 + 1.32x$$

This is what has been used to compute the estimated sales

- **Average seasonal variations**

These are averaged as in the previously example

Year	Quarter 1 %	Quarter 2 %	Quarter 3 %	Quarter 4 %
Averages(indices)	**54**	**89**	**170**	**86**

Forecast for 2009

	x $(Quarters)$	$Trend\ estimates$	Seasonal Indices %	Seasonally adjusted forecasts
2009	17	55.5	54	29.97
	18	56.82	89	50.57
	19	58.14	170	98.84
	20	59.46	86	51.13

A student is encouraged to fully work out the above summarised workings in order to gin more practice!

Example 19.8

The sales of a leasing company have a trend line that is represented by the equation $\boldsymbol{y = 55,000 + 1250t}$, where $\boldsymbol{t = 1}$ is the first quarter of 2006.

The four quarterly seasonal indices of sales are 107, 100, 82 and 111.

Required

(i) Approximate how sales have been increasing in each quarter up to the year ended 2008.

(ii) Forecast the quarterly sales for the year 2009.

Solution

(i)

Year	$\boldsymbol{t}$ $(\boldsymbol{Quarters})$	$\boldsymbol{Trend\ estimates}$	Seasonal Indices	Seasonally adjusted forecasts
2006	1	56,250	107	60,188
	2	57,500	100	57,500
	3	58,750	82	48,175
	4	60,000	111	66,000
2007	5	61,250	107	65,538
	6	62,500	100	62,500
	7	63,750	82	52,275
	8	65,000	111	72,150
2008	9	66,250	107	70,888
	10	67,500	100	67,500
	11	68,750	82	6,375
	12	70,000	111	77,700

(iii) Forecast for 2009

	$\boldsymbol{t}$ $(\boldsymbol{Quarters})$	$\boldsymbol{Trend\ estimates}$	Seasonal Indices %	Seasonally adjusted forecasts
2009	13	71,250	107	76,238
	14	72,500	100	72,500
	15	73,750	82	60,475
	16	75,000	111	83,250

Trend estimates are obtained by substituting quarterly values in the trend line equation.

EXERCISE 14

15.1 The following table gives a record of sales units from a supermarket in town on a yearly basis.

Year	1	2	3	4	5	6	7	8	9
Sales	450	440	460	410	380	400	370	360	410

(a) Find an equation relating sales to years in the form y = a + bx
(b) Using your equation estimate the sales units in the year 14.
(c) Calculate the 3 point moving average for the data.
(d) Draw a histogram of the raw data.

15.2 The number of auditing jobs performed by ***PKF*** auditing firm in each of the last nine months are listed below;

Months	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Jobs	353	387	342	374	396	409	399	412	408

Required;
PKF firm feels that if June auditing jobs are more than 410, they should hire an extra auditor.
Advise the firm whether they should do so if:

(i) The jobs assume a linear trend
*(ii) Assume a three-month period weighted moving average with weights (indices) of **0.60**, **0.30** and **0.10***

15.3 The table below gives the quarterly demand for the hotel accommodation in thousands of beds.

	Quarters			
Year	**1**	**2**	**3**	**4**
2005	19.4	20.6	19.5	22.8
2006	22.3	22.6	21.0	24.9
2007	23.3	24.1	22.2	25.6

Required;
(i) Compute the trend and the seasonal variations of the series.
(ii) *Forecast the future demand up to 2009*.

15.4 The table below gives monthly demand of some commodity in thousands of bags.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Demand 000'bags	3.5	4	4.2	4.5	5	6	5.5	4.8	6	7	4.8	8.2

Assuming that the forecast and actual demand for Dec of the previous year was 4,000 bags, use a smoothing constant of 0.3 to forecast demand for the commodity for the period given.

15.5 The quarterly tax revenues (in millions of shs) for a certain district for some three years are given below.

Year (yr)	**Yr 1**	**Yr 2**	**Yr 3**
Quarter 1	2.8	3.0	3.0
Quarter 2	4.2	4.2	4.7
Quarter 3	4.6	5.0	5.3

Calculate the equation of the line relating tax revenues (y) with the period (x).
Use your equation to estimate the tax revenue in the first quarter of year 4.

15.6 The following table gives a record of sales units from a supermarket in town on a monthly basis.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Sales Units	450	440	460	410	380	400	370	360	410	450	470	490

Obtain exponential forecasts for the period given using smoothing constants of 0.3 and 0.8. Use the Jan. value as both previous forecast and actual sales for the February forecast.

PART F
INDEX NUMBERS

Chapter 20
INDEX NUMBERS

20.1 Definition

An index number is a measure of relative change in a variable or group of variables in regard to time, geographical location, and other characteristics such as income.

20.2 Types of index numbers

There are 3 major types of index numbers and these include;

(i) **Price index numbers**;-

These measure the changes in commodity prices.

(ii) **Quantity index numbers**;-

These measure the changes in the volume of goods produced, consumed and distributed.

(iii) **Value index numbers**

These measure the changes in the total value of goods produced and sold i.e.

$$\boldsymbol{Price} \times \boldsymbol{Quantity} \qquad (20.1)$$

20.2 Classification of index numbers

Index numbers are classified into two major classes, i.e.;

1. *Un weighted indices*
2. *Weighted indices*

All the above 3 types of index numbers are reflected in the 2 classes (a) and (b) above.

1. Un-weighted index numbers

Definition

Un-weighted index numbers are those that do not consider the relative importance of the commodity for which they are computed.

Types of un-weighted index numbers

There are 3 types of un-weighted index numbers and these include;

(i) Simple index numbers or relatives
(ii) Simple aggregate index numbers
(iii) Simple average index relatives

These are discussed below;

(i) Simple index numbers or relatives

This is a ratio of the value of one variable of a single commodity (i.e. price, quantity or value) in a given year to its value in another period known as the base year / period.

Simple index numbers can be;

- Price relatives
- Quantity relatives
- Value relatives

Where;

- **Price relatives** are index numbers expressed as a ratio of the price of a single commodity in a given period to its price in another period called the base period. *i.e.*

$$\mathbf{Pricerelative} = \left(\frac{\mathbf{P}_n}{\mathbf{P}_o}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots asaratio$$

$$\mathbf{Pricerelative} = \left(\frac{\mathbf{P}_n}{\mathbf{P}_o}\right) \times \mathbf{100} \ldots\ldots\ldots\ldots\ldots asapercentage$$

- **Quantity / volume relatives** are index numbers expressed as a ratio of the quantity of a single commodity in a given period to its price in another period (base year).

i.e.

$$\mathbf{Quantityrelative} = \left(\frac{\mathbf{Q}_n}{\mathbf{Q}_o}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots asaratio$$

$$\mathbf{Quantityrelative} = \left(\frac{\mathbf{Q}_n}{\mathbf{Q}_o}\right) \times \mathbf{100} \ldots\ldots\ldots\ldots\ldots asapercentage$$

- **Value relatives** are index numbers expressed as the ratio of the value of a single commodity (i.e. price x quantity) in a given period to the price in another period (base year).

$$\mathbf{Valuerelative} = \left(\frac{\mathbf{P}_n . \mathbf{Q}_n}{\mathbf{P}_o . \mathbf{Q}_O}\right) \times \mathbf{100}$$

Example 20.1:

The table below shows the price and quantity of coffee produced in Bugiri district.

Year	Price of maize Shs "millions"	Quantity of Maize Units
2005	1,000	5,000
2006	1,200	3,800
2007	1,100	4,200

Required;

Compute; *(i)* *The price relative*
(ii) *Quantity relative*
(iii) *Value relative of the commodity in each year taking 2005 as the base year.*

Solution

(i) Price relative

2007 Price relative = $\left(\frac{P_{2007}}{P_{2005}}\right) \times 100 = \left(\frac{1,100}{1,000}\right) \times 100 = 110\%$

2006 Price relative = $\left(\frac{P_{2006}}{P_{2005}}\right) \times 100 = \left(\frac{1,200}{1,000}\right) \times 100 = 120\%$

Interpretation:- There was an increase in coffee price of 10% in 2007 and 20% in 2006, taking 2005 as the base year.

(ii) Quantity relative

2007 Quantity relative = $\left(\frac{Q_{2007}}{Q_{2005}}\right) \times 100 = \left(\frac{4,200}{5,000}\right) \times 100 = 84\%$

2006 Quantity relative = $\left(\frac{Q_{2006}}{Q_{2005}}\right) \times 100 = \left(\frac{3,800}{5,000}\right) \times 100 = 76\%$

Interpretation:- There was a drop per unit coffee production of 16% in 2007 and 24% in 2006, taking 2005 as the base year.

(iii) Value relative

2007 Value relative = $\left(\frac{P_{2007} \cdot Q_{2007}}{P_{2005} \cdot Q_{2005}}\right) \times 100 = \left(\frac{1,100 \times 4,200}{1,000 \times 5,000}\right) \times 100 = 92.4\%$

2006 Value relative = $\left(\frac{P_{2006} \cdot Q_{2006}}{P_{2005} \cdot Q_{2005}}\right) \times 100 = \left(\frac{1,200 \times 3,800}{1,000 \times 5,000}\right) \times 100 = 91.2\%$

Interpretation:- There was a fall in value of coffee of 92.4% in 2007 and 91.2% in 2006, taking 2005 as the base year.

(i) Simple aggregate index numbers

Definition

The simple aggregate index number refers to the ratio of the sum of the commodities' prices, quantities or values in a current year to the corresponding sum in the base year.

This implies that unlike the simple numbers which compute indices for a single commodity, the simple aggregate index numbers computes indices for a set of commodities. Simple aggregate index numbers are also divided into;

- *Simple aggregate price index number given by*;

$$= \left(\frac{\sum \mathbf{P}_n}{\sum \mathbf{P}_o}\right) \times \mathbf{100}$$

- *Simple aggregate volume / quantity index numbers given by;*

$$= \left(\frac{\sum \mathbf{Q}_n}{\sum \mathbf{Q}_o}\right) \times \mathbf{100}$$

- *Simple aggregate value index numbers given by*;

$$= \left(\frac{\sum \mathbf{P}_n \cdot \mathbf{Q}_n}{\sum \mathbf{P}_o \cdot \mathbf{Q}_O}\right) \times \mathbf{100}$$

Where;-

$\sum \mathbf{P}_n$ and $\sum \mathbf{Q}_n$ = sum of the commodities' prices and quantities in the current year respectively *and*

$\sum \mathbf{P}_o$ and $\sum \mathbf{Q}_o$ = sum of the commodities' prices and quantities in the base year respectively.

Example 20.2
The table below shows a set of commodities along with their prices (in shillings) and quantities.

Year	Base Year		Current year	
Commodity	**Price**	**Quantity**	**Price**	**Quantity**
Blue band	2,400	120	3,200	160
Milk	1,400	240	1,800	200
Bread	1,600	200	1,000	320
Tea leaves	4,000	80	6,000	80
Sugar	1,200	400	800	600

Required; Calculate

(a) Simple aggregate price index number
(b) Simple aggregate Quantity index number
(c) Simple aggregate Value index number

Solution

Year	Base Year			Current year		
Commodity	**Price** $\mathbf{P_0}$	**Quantity** $\mathbf{Q_0}$	$\mathbf{P_0Q_0}$	**Price** $\mathbf{P_n}$	**Quantity** $\mathbf{Q_n}$	$\mathbf{P_nQ_n}$
Blue band	2,400	120	288,00	3,200	160	512,000
Milk	1,400	240	336,000	1,800	200	360,000
Bread	1,600	200	320,000	1,000	320	320,000
Tea leaves	4,000	80	320,000	6,000	80	480,000
Sugar	1,200	400	480,000	800	600	480,000
	$\sum\mathbf{P_0}$ **=10, 600**	$\sum\mathbf{Q_o} = \mathbf{1,040}$	$\sum\mathbf{P_0Q_o}$ = **1,744,000**	$\sum\mathbf{P_n}$ = **12,800**	$\sum_{n} \mathbf{Q} = \mathbf{1,360}$	$\sum\mathbf{P_nQ_n}$ = **2,152,000**

(a) Simple aggregate price index = $\left(\frac{\sum \mathrm{P}_n}{\sum \mathrm{P}_\mathrm{o}}\right) \times 100 = \left(\frac{12,800}{10,600}\right) \times 100 = \mathbf{121}\%$

(b) Simple aggregate Quantity index = $\left(\frac{\sum \mathrm{Q}_n}{\sum \mathrm{Q}_\mathrm{o}}\right) \times 100 = \left(\frac{1,360}{1,040}\right) \times 100 = \mathbf{131}\%$

(c) Simple aggregate value index = $\left(\frac{\sum \mathrm{P}_n . \mathrm{Q}_n}{\sum \mathrm{P}_\mathrm{o} . \mathrm{Q}_\mathrm{O}}\right) \times 100 = \left(\frac{2,152,000}{1,744,000}\right) \times 100 = \mathbf{123}\%$

(ii) Simple average index relatives

A simple average index relative is an index number got by finding the average of the relatives of the commodities. Simple average of relatives are divided into three parts, i.e.

- *A simple average of price relatives given by*;

$$\frac{\sum\left(\frac{\mathbf{P_n}}{\mathbf{P_0}}\right)}{\boldsymbol{n}}$$

Where; $\sum\left(\frac{\boldsymbol{P_n}}{\boldsymbol{P_0}}\right)$ = sum of the price relatives of all commodities and

$\boldsymbol{n}$ = number of commodities

- *A simple average of volume relatives given by*;

$$\frac{\sum\left(\frac{Q_n}{Q_0}\right)}{n}$$

Where; $\sum\left(\frac{Q_n}{Q_0}\right)$= sum of the volume relatives of all commodities and n= number of commodities

- *A simple average of value relatives given by*;

$$= \left(\frac{\sum P_n . Q_n}{\sum P_0 . Q_0}\right) \times \frac{1}{n}$$

Example 20.3:

The table below shows the commodity prices and quantities.

Year	Base year		Current year	
Commodity	**Price**	**Quantity**	**Price**	**Quantity**
P	500	80	800	70
Q	420	50	450	100
R	300	120	210	180
S	900	75	750	140

Required;- Compute;

(i) Price average relative
(ii) Volume average relative
(iii) Value average relative

Solution

Year	Base Year			Current year				
Comm	**Price** P_0	**Qty** Q_0	P_0Q_0	**Price** P_n	**Qty** Q_n	P_nQ_n	$\frac{P_n}{P_0}$	$\frac{Q_n}{Q_0}$
P	500	80	40,000	800	70	56,000	1.6	0.9
Q	420	50	21,000	450	100	45,000	1.1	2.0
R	300	120	36,000	210	180	37,800	0.7	1.5
S	900	75	67,500	750	140	105,000	0.8	1.9
	$\sum P_0$ = 2,120	$\sum Q_o$ = 325	$\sum P_0Q_0$ = 164,500	$\sum P_n$ = 2,210	$\sum Q_n$ = 490	$\sum P_nQ_n$ = 243,800	$\sum\left(\frac{P_n}{P_0}\right)$ = 4.2	$\sum\left(\frac{Q_n}{Q_0}\right)$ = 6.3

(i) Price average relative = $\frac{\sum\left(\frac{P_n}{P_0}\right)}{n} = \frac{4.2}{4} = \mathbf{1.05}$

(ii) Volume average relative = $\frac{\sum\left(\frac{Q_n}{Q_0}\right)}{n} = \frac{6.3}{4} = \mathbf{1.58}$

(iii) Value average relative = $\left(\frac{\sum P_n . Q_n}{\sum P_o . Q_o}\right) \times \frac{1}{n} = \frac{243{,}800}{164{,}500} \times \frac{1}{4} = \mathbf{0.37}$

2. WEIGHTED INDEX NUMBERS

Definition
These are index numbers that take into consideration the relative importance of the commodities for which they are being computed. Weighted index numbers are of two types; i.e.

1. *Weighted price index*
2. *Weighted quantity index numbers*

These are described below with mathematical illustrations;

1. Weighted price index

These are index numbers that show the relative changes in the commodity prices. They include the following index numbers;

(a) Weighted average price index
(b) Laspeyre's price index
(c) Paasche's price index
(d) Edeworth-Marshall's price index
(e) Fisher's ideal index
(f) Dorbish-Bowley's index

These are described as below;

(a) Weighted average price index

This is computed using weights of any typical year and not necessarily the base year or the current year.
It is expressed as;

$$Weightedaveragepriceindex = \left(\frac{\sum \mathbf{P}_n \mathbf{W}_i}{\sum \mathbf{P}_0 \mathbf{W}_i}\right) \times \mathbf{100}$$

Where; $\mathbf{w_i}$ = weights of the i^{th} commodity.

Example 20.4
The table shows the prices as per the current and base year with the corresponding weights of four commodities, X, Y, W, Z.

Commodity	Base year Price	Current Year Price	Commodity weights
X	500	800	70
Y	420	450	100
W	300	210	180
Z	900	750	140

Required:
Compute the weighted average price index

Solution

	Base Year			Current year		
Comm	**Price** $\mathbf{P_0}$	**Weights** $\mathbf{W_i}$	$\mathbf{P_0 W_i}$	**Price** $\mathbf{P_n}$	**Weights** $\mathbf{W_i}$	$\mathbf{P_n W_i}$
X	500	80	40,000	800	70	56,000
Y	420	50	21,000	450	100	45,000
W	300	120	36,000	210	180	37,800
Z	900	75	67,500	750	140	105,000
			$\sum P_0 W_i$ = **164,500**			$\sum P_n W_i$ = **243,800**

From the expression;

$$Weightedaveragepriceindex = \left(\frac{\sum P_n W_i}{\sum P_o W_i}\right) \times 100 \ = \left(\frac{243{,}800}{164{,}500}\right) \times 100 \ = \mathbf{148.2}\%$$

(b) Laspeyre's price index

This is computed using the base year quantity weights (Q_o) and is given by;

$$Laspeyres' spriceindex \ = \left(\frac{\sum \mathbf{P_n \cdot Q_0}}{\sum \mathbf{P_o \cdot Q_O}}\right) \times \mathbf{100}$$

(c) Paasche's price index

This is a weighted index number computed using current year quantity weights i.e. Q_n.

It is given by;

$$Paasche' spriceindex \ = \left(\frac{\sum \mathbf{P_n \cdot Q_n}}{\sum \mathbf{P_o \cdot Q_n}}\right) \times \mathbf{100}$$

Example 20.5:
The table below shows the prices and quantities of selected food products;

Year	**Base Year**		**Current Year**	
Commodity	**Price**	**Quantity**	**Price**	**Quantity**
Rice	800	15	1,400	15
Potatoes	350	25	200	27
Flour	2,000	12	1,500	15
Wheat	3,000	18	2,000	20

You are required to calculate and interpret

a. Laspeyre's price index
b. Paasche's price index

Solution

From the table below;

(a) $Laspeyres' spriceindex = \left(\frac{\sum P_n \cdot Q_0}{\sum P_o \cdot Q_O}\right) \times 100 = \frac{80{,}000}{98{,}750} \times 100 = \mathbf{81.0}\%$

(b) $Paasche' spriceindex = \left(\frac{\sum P_n \cdot Q_n}{\sum P_o \cdot Q_n}\right) \times 100 = \frac{88{,}900}{111{,}450} \times 100 = \mathbf{79.77}\%$

Year	**Base Year**			**Current year**				
Comm.	**Price** $\mathbf{P_0}$	**Qty** $\mathbf{Q_0}$	$\mathbf{P_0Q_0}$	**Price** $\mathbf{P_n}$	**Qty** $\mathbf{Q_n}$	$\mathbf{P_nQ_n}$	$\mathbf{P_nQ_0}$	$\mathbf{P_0Q_n}$
Rice	800	15	12,000	1,400	15	21,000	21,000	12,000
Potato	350	25	8,750	200	27	5,400	5,000	9,450
Flour	2,000	12	24,000	1,500	15	22,500	18,000	30,000
Wheat	3,000	18	54,000	2,000	20	40,000	36,000	60,000
			$\sum \mathbf{P_0Q_0}$ =98,750			$\sum \mathbf{P_nQ_n}$ =88,900	$\sum \mathbf{P_nQ_0}$ =80,000	$\sum \mathbf{P_0Q_n}$ = 111,450

(d) Edeworth-Marshall's price index

This is a weighted index number computed using the sum of quantities in the base year and the current year as the weights.

i.e. **$W = (Q_o + Q_n)$**

It is given by;

$$Edgeworth - Marshall's index = \left(\frac{\sum \mathbf{P_n}.(\mathbf{Q_o} + \mathbf{Q_n})}{\sum \mathbf{P_o}.(\mathbf{Q_o} + \mathbf{Q_n})}\right) \times \mathbf{100}$$

Example 20.6:
Using the table in example 5, compute Edeworth-Marshall's price index.

Solution

	Base year		Current year				
Comm	**Price P_0**	**Qty Q_0**	**Price P_n**	**Qty Q_n**	$(Q_0 + Q_n)$	$P_n(Q_0 + Q_n)$	$P_0(Q_0 + Q_n)$
Rice	800	15	1,400	15	30	42,000	24,000
Potato	350	25	200	27	52	10,400	18,200
Flour	2,000	12	1,500	15	27	40,500	54,000
Wheat	3,000	18	2,000	20	30	60,000	90,000
						$\sum P_n(Q_0 + Q_n)$ **=152,900**	$\sum P_n(Q_0 + Q_n)$ **= 182,200**

Hence;

$$Edgeworth - Marshall's index = \left(\frac{\sum P_n.(Q_o + Q_n)}{\sum P_o.(Q_o + Q_n)}\right) \times 100 = \frac{152{,}900}{182{,}200} \times 10 = \mathbf{83.9\%}$$

(e) Fisher's ideal index

This refers to the geometric mean of the Laspeyre's price index and Paasche's price index.

Whereby;

$$Laspeyres' spriceindex = \left(\frac{\sum \mathbf{P_n}.\mathbf{Q_0}}{\sum \mathbf{P_o}.\mathbf{Q_O}}\right) \times \mathbf{100}$$

$$Paasche' spriceindex = \left(\frac{\sum \mathbf{P_n}.\mathbf{Q_n}}{\sum \mathbf{P_o}.\mathbf{Q_n}}\right) \times \mathbf{100}$$

Therefore;

$$Fisher's\ idealindex = \sqrt{(Laspeyre'sindex) \times (Paasche'sindex)}$$

$$Fisher'sidealindex = \sqrt{\left(\frac{\sum \mathbf{P_n}.\mathbf{Q_0}}{\sum \mathbf{P_o}.\mathbf{Q_0}}\right) \times \left(\frac{\sum \mathbf{P_n}.\mathbf{Q_n}}{\sum \mathbf{P_o}.\mathbf{Q_n}}\right)} \times \ 100$$

Hence from the workings given in *example 5*, Fisher's ideal index would be given as;

$$Fisher's\ idealindex = \sqrt{\mathbf{81.0 \times 79.77}}$$

$$= \mathbf{80.38\%}$$

(f) Dorbish-Bowley's index

This is the arithmetic mean of Laspeyre's and Paasche's price indices.
It is given by the formula;

$$Dorbish-Bowley's\ index = \left[\left(\frac{\sum \mathbf{P_n}.\mathbf{Q_0}}{\sum \mathbf{P_o}.\mathbf{Q_O}}\right) + \left(\frac{\sum \mathbf{P_n}.\mathbf{Q_n}}{\sum \mathbf{P_o}.\mathbf{Q_n}}\right)\right] \times \frac{1}{2}$$

Therefore, from *example 5*, Dorbish-Bowley's index is given as;

$$Dorbish-Bowley's\ index = [(81.0) + (79.77)] \times \tfrac{1}{2} = 160.77 \times \tfrac{1}{2} = \mathbf{80.385}$$

2. Weighted quantity index numbers

These are weighted index numbers which show the relative changes in the quantities of the commodities sold.

The weighted quantity indices are obtained from the above weighted price index formulae by interchanging the prices (P) and the quantities (Q).

That is;

$$(i) Weightedaveragepriceindex = \left(\frac{\sum \mathbf{Q_n}\mathbf{W_i}}{\sum \mathbf{Q_o}\mathbf{W_i}}\right) \times \mathbf{100}$$

$$(ii) Laspeyres'spriceindex = \left(\frac{\sum \mathbf{Q_n}.\mathbf{P_0}}{\sum \mathbf{Q_o}.\mathbf{P_O}}\right) \times \mathbf{100}$$

$$(iii) Edgeworth-Marshall'sindex = \left(\frac{\sum \mathbf{Q_n}.(\mathbf{P_o} + \mathbf{P_n})}{\sum \mathbf{Q_o}.(\mathbf{P_o} + \mathbf{P_n})}\right) \times \mathbf{100}$$

$$(iv) Fisher's\ idealindex = \sqrt{\left(\frac{\sum \mathbf{Q_n}.\mathbf{P_0}}{\sum \mathbf{Q_o}.\mathbf{P_O}}\right) \times \left(\frac{\sum \mathbf{Q_n}.\mathbf{P_n}}{\sum \mathbf{Q_o}.\mathbf{P_n}}\right)} \times 100$$

$$(v) Dorbish-Bowley's\ index = \left[\left(\frac{\sum \mathbf{Q_n}.\mathbf{P_0}}{\sum \mathbf{Q_o}.\mathbf{P_O}}\right) + \left(\frac{\sum \mathbf{Q_n}.\mathbf{P_n}}{\sum \mathbf{Q_o}.\mathbf{P_n}}\right)\right] \times \frac{1}{2}$$

Calculate and interpret using information in example 5;

I. Weighted average quantity index
II. Laspeyre's quantity index
III. Paasche's quantity index
IV. Edgeworth-Marshall's quantity index
V. Fisher's ideal index
VI. Dorbish-Bowley's index

20.3 Shifting the base year of index numbers

The base year of index numbers may be shifted in the following situations;

(i) When comparisons need to be made between index numbers having different base period or years.
(ii) If the base year is too distant in the past and hence no longer useful for comparison purposes.

In such situations, the new index number known as the recast index number must be computed from;

Recast index number = Old index number of that year x 100
Index number of the new base year

20.4 Splicing of index numbers

This refers to a technique of completing an index series that was discontinued due to some reasons, or completing a new index series from the base period backwards.

20.5 Testing the consistency of index numbers

Consistency of index numbers can be verified by carrying out certain tests on the formulae used in computing such index numbers.

There are 3 major tests that can be carried out to check the consistency of index number formulae and these include;

a) Time reversal test
b) Factor reversal test
c) Circular test

These are explained below;

(a) The time reversal test

This test requires that the index number formula must possess the time consistency property when working both forward and backward with respect to time i.e. for any index number computed, after reversing the base year and the current year, the result obtained must be the reciprocal of the original index number.

For example, the price index numbers;

$$\frac{P_n}{P_o} \times \frac{P_o}{P_n} = 1$$

Where; P_n= price in the current year
P_o = Price in the base year

(b) Factor reversal test

This test requires that for an index to satisfy such a test, the product of the corresponding price index (i.e. P_o / n) and the corresponding quantity index (i.e. Q_o / n) give the true value index. For example; for simple index numbers;

Price index = P_n / P_o and Quantity index = Q_n / Q_o

Test = $(P_n / P_o) \times (Q_n / Q_o) = (P_nQ_n / P_oQ_o)$ which is the value index.

Therefore the test is satisfied.

20.6 Features of a well constructed index number

The following are some of the features of a well constructed index number;

(a) It is constructed for the rightful purpose using the most appropriate formula / correct method.
(b) It is constructed using data that is representative of the entire population. Its not always necessary to include the entire population in the construction of index numbers since this makes computation quite tiresome.
(c) It is constructed using a base period that represents normal and stable economic conditions and close to the past period.
(d) It is constructed from data that has been systematically collected at regular intervals of time from prominent business firms / standard retail outlets, which are strategically located and often visited by majority of customers.
(e) It is constructed using well selected weighing systems i.e. well constructed prices, quantities and value weights.

20.7 Importance of index numbers

(i) They are used in formulating suitable business and economic policies.
(ii) They are important in forecasting future economic activities e.g. they are used in time series analysis to study long-term trend, seasonal variation and cyclical developments.
(iii) They are used in measuring the cost of living over a given period of time i.e. cost of living indices.
(iv) They are useful in converting nominal wage to real wage (i.e. deflating of index numbers).
(v) They reveal trends and tendencies in business, social and economic activities and as thus are important in making conclusions as regards cyclical or irregular movements.
(vi) They are to make comparisons between different periods e.g. price levels between 2 or more periods.

20.8 Limitations of index numbers

(a) Index numbers are computed using many methods, which may give different results. This at times creates confusion.
(b) Most index numbers are computed from sample data. This increases the risk of bias and reduces on the accuracy of index numbers.
(c) Index numbers can be misused in order to draw desired conclusions; meaning that the computed index numbers are as good as the person computing them.
(d) Some index numbers are based on selected items, they simply depict the broad trend and not the complete reality.

20.9 PUBLISHED INDEX NUMBERS

There are two major types of published index numbers and these include;

(c) The real Income (wage) index number

Real wage index numbers are computed from;

Real income/wage index = Real income/wage of current year x 100 %
Real income / wage of base year

However, one has to note that in inflationary situations, real income / wage has to be deflated in order to make an allowance for the effect of changing price levels.
In deflating real income, we shall obtain the deflated income and this shall be taken as the new real income for index number computational purposes.

The formula for deflated income is given by;

Deflated income / wage = Nominal or money wage / income x 100
(Real income) Price index

(d) The cost of living index numbers

Cost of living index numbers also known as consumer price indices (CPI) are special purpose index numbers that are constructed to measure the relative change in the cost of living.

The cost of living index is given by;

$$The\ cost\ of\ living\ index = \left(\frac{\sum \mathbf{P}_n \mathbf{W}}{\sum \mathbf{P}_o \mathbf{W}}\right) \times \mathbf{100}$$

These index numbers are generally constructed for each week. Where the average of the weekly index numbers is taken as the index number for the month and the average of the monthly index numbers gives the cost of living for the whole year.

20.9.1 Importance of cost of living index numbers

(a) They are purposely used in determining remunerations, so as to maintain the same standard of living in other years as that of the base year.
(b) Their reciprocal is used in measuring the purchasing power of money
(c) They are also used to find real wages by the process of deflation.
(d) Give guidelines to the government on its decisions about income and general economic policy, wage policy, and price policy among others.

EXERCISE 16

16.1

(a) What is an index number? Explain areas where index numbers are applied.

(b) The table below shows values of consumer price index (CPI) for 1995 through 1999.

Year	1995	1996	1997	1998	1999
Consumer price index	130.7	136.2	140.7	144.5	148.2

Required;

Determine the purchasing power of the shilling for each of these years in terms of the value of the shilling in the 1995 base period.

16.2 (a) Explain the following terms as used in index numbers

(i) Price index

(ii) Quantity index

(iii) Value index

(b) The following prices and quantities reflect weekly consumption patterns of a certain family for the years 2007 and 2008

Year	**Year 2007**		**Year 2008**	
Item	**Price (P_o) Shs**	**Quantity(Q_O)**	**Price (P_t) Shs**	**Quantity (Q_t)**
Oranges (Kg)	15	2	25	1
Milk (Litres)	30	2	35	2
Bread (loaves)	30	3	40	3
Eggs (Dozens)	30	1	65	1

Required; Compute

(i) Price relative for each item

(ii) Laspeyres price index

(iii) Paaches price index

(c) What are some of the limitations of index numbers?

16.3 In Kampala the average weekly wages for a certain group of wage earners was Shs. 3,022.50. In 1998 the average weekly pay for the same group of wage earners showed an earning of Shs. 4,836. In 1998 the consumer price index using 1990 as a base was 165.

Required;

Determine whether the wage earners were better off or worse off in 1998 than 1990.

PART G
NETWORK ANALYSIS

Chapter 21
NETWORK ANALYSIS

21.1 Definition

Network analysis refers to a family of related techniques developed to aid management to plan and control projects.
Network analysis is used to evaluate projects that are large and complex; where restrictions in terms of time and resources exist.

21.2: Importance of network analysis

(i) At the planning stage, network analysis helps managers to plan on how to get started on the project by indentifying the activities of the project.
(ii) It helps in showing the inter-relationship of various jobs or tasks which make-up the overall project and identifying the time durations for each.
(iii) It clearly indentifies the critical parts of the project.
(iv) Provides planning and control of information on time, cost and resource aspects of a project.
(v) Network analysis helps in determining the total completion time to the project if everything goes as to the plan.

21.3: Key terms

(i) **Project Planning**

Project is the task of breaking down a project into its constituent activities and deciding on their time relationships.
Project planning helps management to establish objectives and to be able to pinpoint courses of action.

(ii) **Activity**;-

This is a task or job of work which takes time and resources e.g. verifying the debtors in the sales day book.
An activity is represented by an arrow where the head indicates the end point and the tail represents the start point of a task.

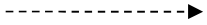
(iii) **Event**;-

This is a point in time and indicates the start or finish of an activity, or activities e.g. debtors verified.
An event is represented by a circle or node as shown below;

All activities must start and end with an event.

(iv) **Dummy activities;-**

This is an activity that does not consume time or resources. It is used merely to show clearly, logical dependencies between activities so as not to violate the rules for drawing networks. It is represented on a network by dotted arrow thus;



Dummy activities are not usually listed with the real activities but become necessary as the network is drawn.

21.4: Rules for drawing networks

(i) A complete network should have only one point of entry – a start event and only one point of exit i.e. a finish event.
(ii) Every activity must have one preceding or 'tail' event and one succeeding or 'head' event. Several activities may use the same tail event and heat event.
(iii) No activity can start until its tail event is reached.
(iv) An event is not complete until all activities leading into it are complete.

(v) Loops .i.e. a series of activities which lead back to the same event, are not allowed since networks are a progression of activities always moving onwards in time.
(vi) All activities must be tied into the network.

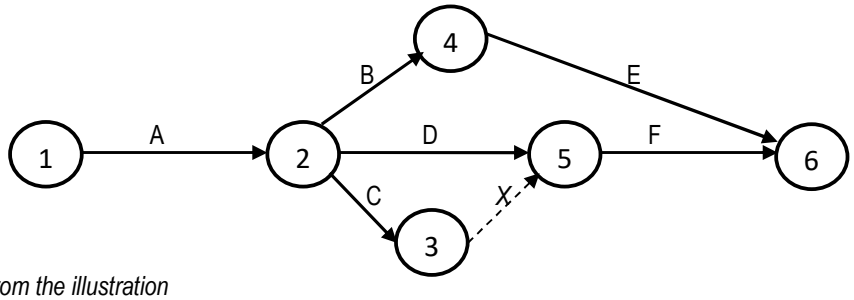
Other guiding conventions
(vii) Networks proceed from left to right
(viii) Networks are not drawn to scale ***i.e.*** the length of the arrow does not represent time elapsed.

Example 21.1
Consider a network shown below the values 1,2,……,6 represent events while letters A,B …., F represent activities.

Activity	Event	Preceding activity
A	2	-
B	4	A
C	3	A
D	5	A
E	6	B
F	6	C,D

Required;
(a) Construct a network showing the relationship between activities and events.
(b) Interpret the network relationships depicted by the network diagram.
(c) Outline the possible paths within the network

Solution
(a)



(b) From the illustration
- Events B, C and D cannot begin until A has completed.
- Activity F cannot begin until D and C are completed.
- There is only one start and end events i.e. event (1) and event (6).

Note:
Activities *C* and *D* have the same start and end events except for the extra events 3 and its associated dummy activity *X*.

(c) Possible paths within the network

These include
Path 1: - ABE
Path 2:- ADF
Path 3:- AC*X*F (Where *X* is a dummy activity)

Example 21.2
The table below shows activities required to complete a certain project and their respective preceding activities.

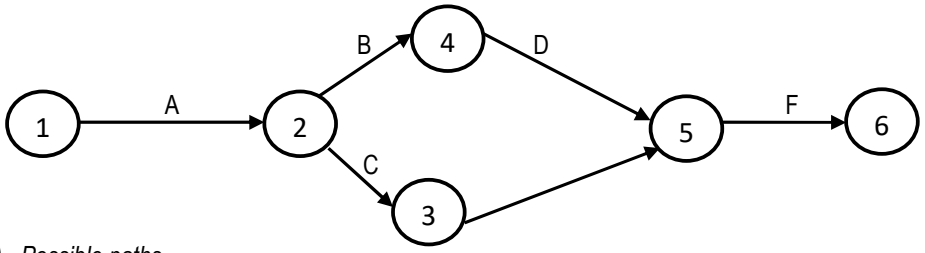
Activity	Preceding activity
A	-
B,C	A
D	B
E	C
F	D,E

Required;

(a) Construct a network for the above project showing the possible paths that can be taken during the project.
(b) Outline the possible paths within the network.
(c) Interpret the network relationships depicted in the paths

Solution

(a)



(b) Possible paths

Path 1:- ABCDF
Path 2:- ACEF

(c) Interpretation

(i) Activities B and C cannot begin until A has finished.
(ii) Activity F cannot begin until D and E are completed.
(iii) There is one start event (1) and stop event (6).

21.5: TIME ANALYSIS

This is the process of assessing the duration of a project. Analysis of project duration is done by inserting activity duration times in a fully drawn network.

This helps in analysis any time lags or critical time that may be encountered during the project life span.

Basic time analysis of a network involves calculating the critical path.

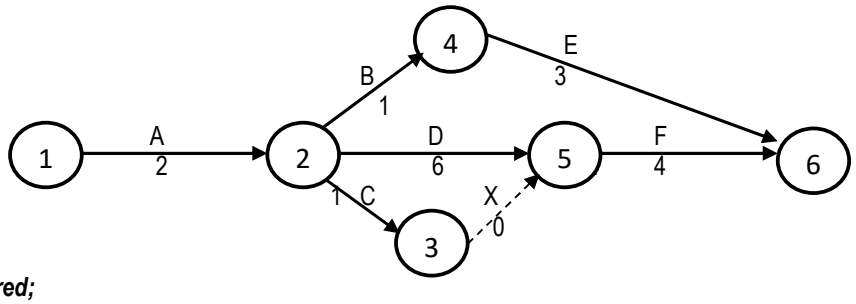
Definitions

(i) Critical path

The critical path of a network gives the shortest time in which the project can be completed. It is the chain of activities with the longest duration times.
There may be more than one critical path in a network and it is possible for a critical path to go through a dummy.

Example 21.3

Consider the network diagram in example 1 above;
Assume the network has been drawn to completion and activity duration times estimated in days as shown below;



Required;

(a) Compute the duration time for each path in the network.
(b) What is the critical path and project duration?

Solution

(a) *Duration time for each path*;

Path 1; - ABE = 2+1+3 = 6 days
Path 2:- ADF = 2+6+4 = 12 days *** *Critical path*
Path 3:- ACXF = 2+1+0+4=7days

(b) Critical paths Vs Project duration

Path ADF has the longest duration time compared to other paths; and therefore qualifies to be the critical paths.

Note:

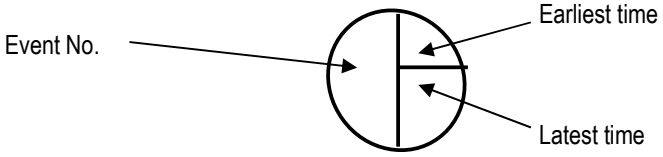
A dummy consumes no time in term of duration period i.e. days = 0, for a dummy.

(ii) Earliest Start time (EST) and Latest start time (LST)

The critical path can also be established by calculating the "Earliest start time (EST)" and the "Latest start time (LST)" for each event and comparing them.
The critical path is then obtained from the chain of activities where the ESTs and LSTs are the same.

The **EST** and the **LST** are more systematic computerized methods of identifying a critical path.

The notation for making events is shown below;



Definitions;

(a) **Earliest start time (EST)**;

This is the earliest possible time at which a succeeding activity can start.
The EST assesses the total project time.

Procedure for computing EST (*forward pass*).

- The EST of a head event is obtained by adding onto the EST of the tail event the linking activity duration starting from event 0, time 0 and working forward through the network.
- Where two or more routes arrive at an event, the longest route is taken.
- The EST in the finish event is the project duration and is the shortest time in which the whole project can be completed.

(b) **Latest start time (LST)**;

This is the latest possible time at which a proceeding activity can finish without increasing the project duration.

The LST helps in isolating the critical path.

Procedure for computing EST (*Backward pass*)

- Starting at the finish event, insert the LST and work backwards through the network deducting each activity duration from the previously calculated LST.
- Where the tails of activities join the lowest number is taken.

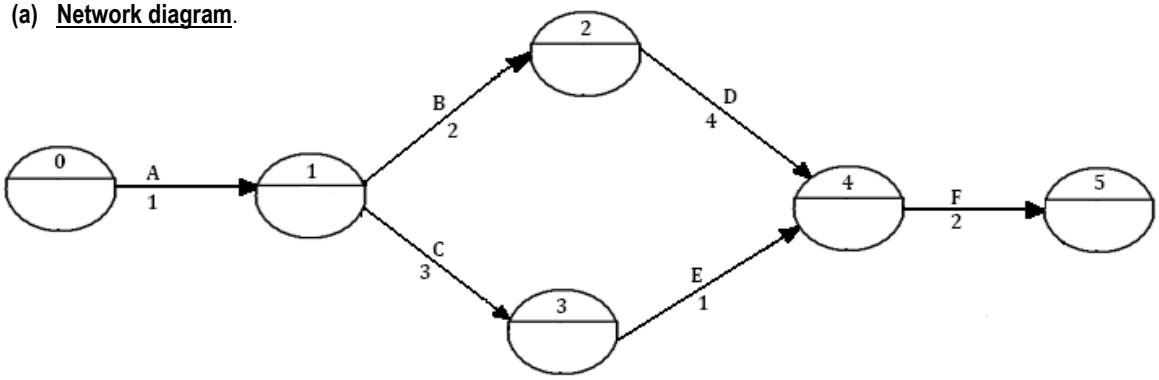
Example 21.4
The table below shows activities required to complete a certain project and their respective preceding activities.

Activity	Days	Preceding activity
A	1	-
B	2	A
C	3	A
D	4	B
E	1	C
F	2	D,E

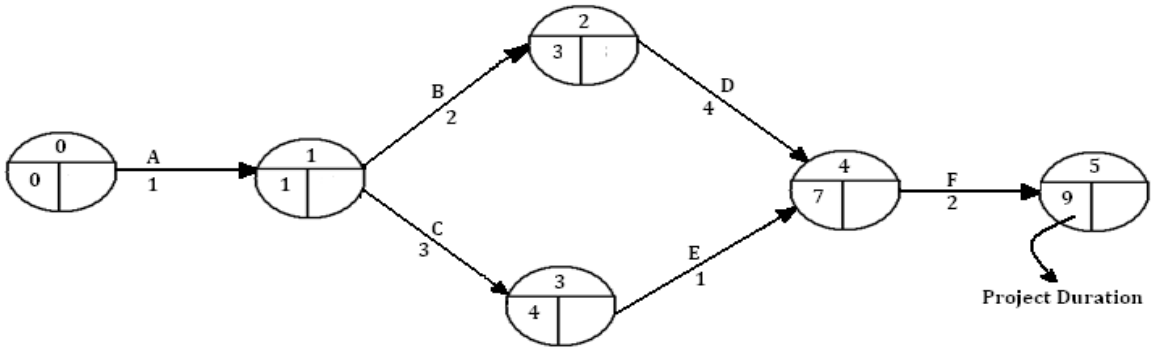
Required;
(a) Construct a network for the above project showing the possible paths that can be taken during the project.
(b) Indicate the EST within the network.
(c) Indicate the LST within the network.
(d) Determine the critical path.

Solution

(a) Network diagram.



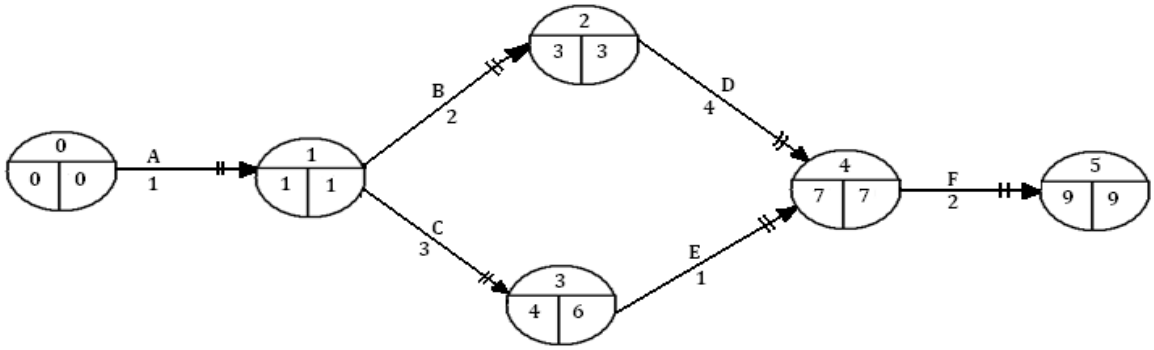
(b) Earliest start time within the network (Forward Pass)



Activity D and E arrive at event 4 and the longest route time i.e. ABD(7 days) must be taken instead of the shortest route time i.e. ACE (5days).

Reason:- The reason for the above treatment is that activity F depends on completion of activity D and E. However, E is completed by day 5 and D is not completed until day 7.
Therefore F cannot start before day 7.

(c) Latest start time (Backward Pass) within the network.

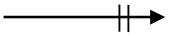


Tails of activities B and C join event 1 where the LST for C is day 3 while that for B is day 1, hence the lowest LST is taken.

Reason;- If event 1 occurred at day 3 then activities B and D could not be completed by day 7 as required and consequently the project would be delayed.

Observations

From the above example, path ABDF has EST and LST which are identical (see solution). This is the "*Critical path"* which is the chain with the longest duration.
An activity is critical when EST = LST of the tail event, and EST = LST of the head event and the EST of the head event *minus* the EST of the tail event equals the activity duration.

The critical path is indicated on the network by two small transverse lines across the arrows on the critical path
i.e. 

Point to note

The activities along the critical path are vital activities and must be completed by their EST and LST to prevent delays in the completion of the project.

(iii) Float

Non-critical activities (C and E, in example 4) have "spare time" and this is known as "float."
Activities C and E could take up additional 2 days in total without delaying the project duration.
If it is required to reduce the overall project duration, then the time of one or more of the activities on the critical path must be reduced perhaps by using more labour, better equipment or some other method of reducing job times.

Definition
Float is defined as the spare time available on non-critical activities.

Types of float

There are 3 types of float; i.e.

- **Total float**;- This is the amount of time a path of activities could be delayed without affecting the overall project duration.
  Total float is given by the formula below;

Total float = maximum available time – activity duration time
= latest time of end event – earliest time of start event - activity duration time

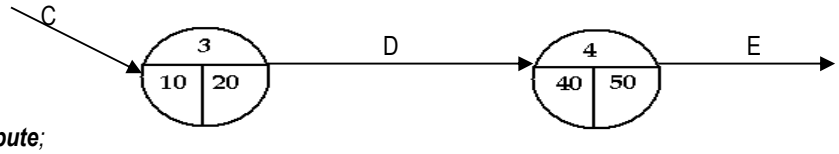
- **Free float**;- This is the amount of time an activity can be delayed without affecting the commencement of a subsequent activity at its EST, but may affect the float of a previous activity.
  Free float is given by the formula below;

free float = Duration between the earliest times of both events – activity duration time
= Earliest head time – earliest tail time - activity duration time

- **Independent float**;- This is the amount of time an activity can be delayed when all preceding activities are completed as late as possible and all succeeding activities completed as early as possible. The independent float therefore does not affect the float of either preceding or subsequent activities.
  Independent float is given by the formula below;

Independent float = minimum available time – activity duration time
= Earliest time of end event – latest time of start event - activity duration time

Example 21.5
Consider a network diagram shown below;



***Compute**;*

(i) Total float
(ii) Free float
(iii) Independent float

Solution
(i) Total Float = LST of end event – EST of start event – Activity duration time
= 50 – 10 - 10 = **30 days**
(ii) Free float = Earliest head time – Earliest tail event – Activity duration time
= 40 – 10 - 10 = **20 days**

(iii) Independent float = EST of end event – LST of start event – Activity duration
= 40 -20-10 = **10 days**

21.5: COST ANALYSIS

Cost analysis in network analysis is of importance as it helps one to be able to calculate the cost of various project durations.

The normal duration of a project incurs a certain cost and by more labour, working overtime, more equipment etc. the duration could be reduced but an expense of higher costs.
Some ways of reducing the project duration are cheaper than others and therefore cost analysis in network is necessary in identifying the cheapest way of reducing the overall duration.

Early completion in most project analysis would attract a bonus while late completion would attract a fine under cost analysis.
Therefore it is important to analyse costs to establish whether it is worthwhile paying extra to reduce the project time so as to save penalty.

The following are key terms used in cost analysis;

(i) **Normal cost**;- This is the cost associated with a normal time estimate for an activity. It's a time estimate for a situation where resources (i.e. men, machines, etc) are used in the most efficient manner.

(ii) **Crash cost**; - This is the cost associated with the minimum possible time for an activity. They are usually higher than the normal cost because of extra wages, overtime premiums, and extra facility among others.

(iii) **Crash time**; - This is the minimum possible time that an activity is planned to take. It comes up because of application of extra resources e.g. more labour or machinery.

(iv) **Cost slope**; - This is the average cost of shortening an activity by one time unit (day, week, month etc). the cost slope is generally assumed to be linear and calculated from;

$$Cost\ slope = \frac{\text{Crash cost} - \text{Normal cost}}{\text{Normal time} - \text{Crash time}}$$

(v) **Least cost scheduling or 'crashing'**;- This is a process which seeks to indentify the least cost method of reducing the overall project time period by time period.

Examples 21.6
ABC Ltd has provided you with the normal and crash durations together with their associated costs for the various activities involved in the repair work. The direct cost for supervision of the work is also indicated as below;

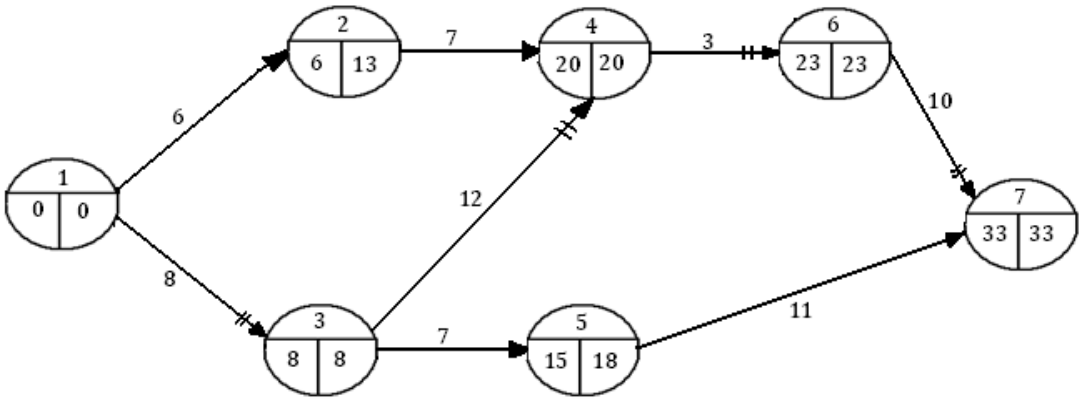
	TIME (Days)		COST (Shs)	
Activity	**Normal**	**Crash**	**Normal**	**Crash**
1-2	6	2	4,000	12,000
1-3	8	3	3,000	6,000
2-4	7	4	2,800	4,000
3-4	12	8	9,000	11,000
4-6	3	1	10,000	13,000
5-6	5	2	4,900	7,000
3-5	7	3	1,800	5,000
5-7	11	5	6,600	12,000
6-7	10	6	4,000	8,400

The indirect cost of the project is Shs. 2,000 per day.

***Required**;*
(i) Draw a network diagram for this project.
(ii) Find the normal cost and duration of the project.
(iii) The crash duration of the project.
(iv) The total crash cost of the project if the total cost is to be minimized.
(v) The total cost of the entire project.

Solution

(i) Network



(ii) Normal Cost

The normal cost is found by summing up the normal costs of all activities i.e;

= 4,000 + 3,000 + 2,800 + 9,000 + 10,000 + 4,900 + 1,800 + 6,600 + 4000
= **Shs. 46,000**/=

Duration of the project = **33 days** (see network diagram)

(iii) Crash duration

This comes from the critical paths marked in the diagram;

Activity	Crash Time
1-3	3
3-4	8
4-6	1
6-7	6
Total	**18 days**

(iv) Total crash cost

- We need to compute the crash cost per unit time for each critical activity using the formula for the cost slope; i.e;

$$Cost\ slope = \frac{\text{Crash cost} - \text{Normal cost}}{\text{Normal time} - \text{Crash time}}$$

Activity	Crash cost per unit
1-3	$= \frac{6{,}000 - 3{,}000}{8 - 3} = \mathbf{600}$
3-4	$= \frac{11{,}000 - 9{,}000}{12 - 8} = \mathbf{500}$

4-6 $$= \frac{13,000 - 10,000}{3 - 1} = \mathbf{1,500}$$

6-7 $$= \frac{8,400 - 4,000}{10 - 6} = \mathbf{1,100}$$

- Activities to be crashed(critical activities)

Critical activities will be crashed at the above cost per unit day as to obtain the total crash cost as shown below;

1-3	by	3 days @ 600	= 1,800
3-4	by	8 days @ 500	= 4,000
4-6	by	1 day @ 1,500	= 1,500
6-7	by	6 days @ 1,100	= 6,600
		18 days	13,900

- Total cost of the entire project

Normal cost	46,100
Add crash cost	13,900
Direct cost	60,000
Add Indirect cost (2,000 × 33 days)	66,000
Total cost	**126,000**

21.6: RESOURCE ANALYSIS

For a proper resource scheduling, all resource requirements for each activity must be specified.
Various resources involved such as men, machinery etc must be classified and the availability and constraints specified. After calculating the critical path in time analysis, a resource aggregation profile is prepared i.e. the amount of the resources required in each time period of the project based on the ESTs of each project based on the ESTs of each activity.
In any case, if the resource aggregated indicates that a constraint is being exceeded, and float is available the resource usage is smoothed i.e. the start of the activities is delayed. Example 21.7 below demonstrates the steps followed in network resource analysis.

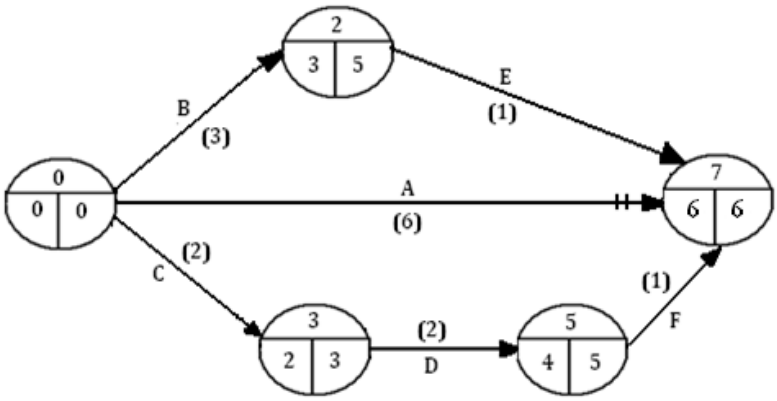
Example 21.7
A project has the following activity durations and resource requirements;

Activity	Preceding activity	Duration (days)	Resource requirements (units)
A	-	6	3
B	-	3	2
C	-	2	2
D	C	2	1
E	B	1	2
F	D	1	1

Assuming no restrictions;
(a) Show the network for the above project
(b) Indicate the critical path and resource requirements on a day by day basis.
(c) Assuming there is a Resource constraint of 6 units, what would be the plan?

Solution

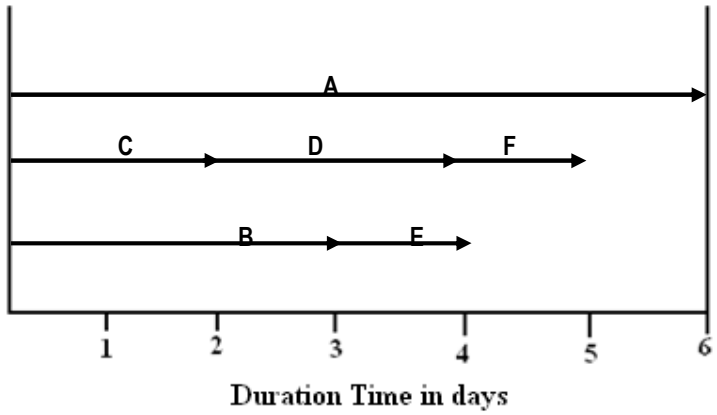
(a) Network



(b) Time resource network

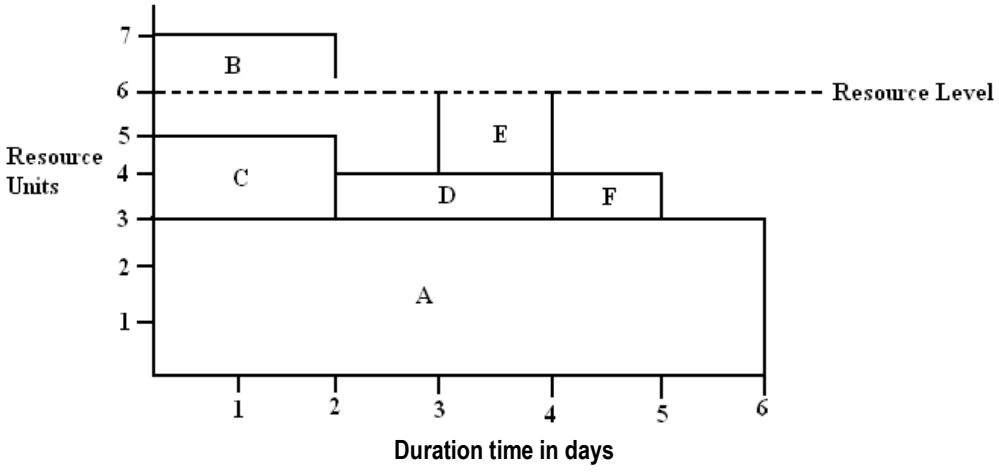
A time resource network is simply a diagram showing network information organised on a Gantt or Bar Chart based on the Earliest Start Times (ESTs) of project activities.

In our particular example, activities A, B, C and D are plotted against their respective duration times as shown below;



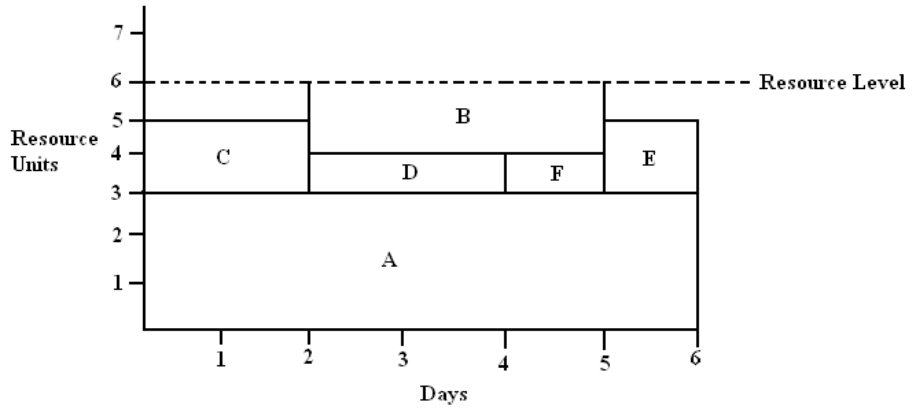
(c) Resource aggregate profile

The Resource Aggregate profile is drawn basing on the time resource network shown above. It shows the total resource requirements in each period of time.



Analysis of the above profile shows that at times resources are required than are available if activities commence at their ESTs. The ESTs / LSTs on the network show that float s available for activities B, C, D and E. Having regard to these floats, its necessary we 'smooth out' the resource requirements so that the resources required do not exceed the resource constraints, i.e. delay the commencement of activities (within their float) and if this procedure is not sufficient then delay the project as a whole. Carrying out this procedure results in the following resource profile.

(d) Smoothed resource requirement profile



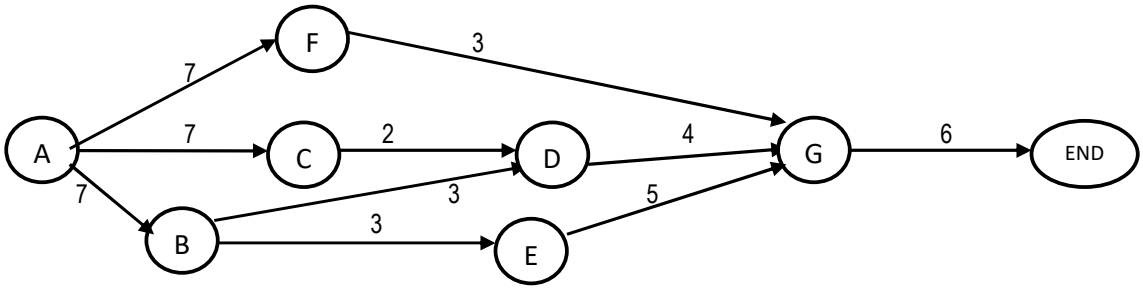
PLAN:

Because of the resource constraint of 6 units, it will be necessary to delay the start of activities B and E until day 2.

In cases where a project may have to be delayed as a whole, and management still requires the project to be completed in the shortest time period, then the only plan would be to provide more resource units than the given resource constraints.

EXERCISE 17

17.1 The network for a small building project is shown below; together with the time in days, required to complete each task.



You are required;

(i) To list the possible paths through the network and the length of the path (in days) in each case.

(ii) To state the critical path and the project duration.

(iii) If D takes 8 days, rather than 4, to explain how much, if at all, the project will be delayed.

17.2 (a) A small construction project involves the following activities;

		Normal		Crash	
	Activity	**Time(days)**	**Cost (Shs)**	**Time(days)**	**Cost (Shs)**
1-2	Clear the ground	6	60,000	5	70,000
1-3	Lay foundation	5	30,000	3	50,000
2-4	Build walls	3	10,000	2	15,000
3-4	Roofing and piping	7	40,000	4	55,000
3-5	Painting	4	20,000	3	30,000
4-5	Landscaping	2	10,000	1	17,500

***Required**;*

(i) Determine the shortest time and the associated cost to finish this project.

(ii) If a penalty of Shs. 4,500 must be charged for every day beyond 12 days, what is the most economical time for completing the project?

17.3 Builders Construction company limited has a project to renovate a bungalow in Muyenga. The major activities and their duration are as follows;

	Activity	**Preceding Activity**	**Duration (days)**
A:	Replace windows	-	5
B:	Re-wiring	-	4
C:	Re-plastering	A	2
D:	Fit-lights	B	1
E:	Decorate bedrooms	B	5
F:	Install plumbing	B	5
G:	Decorate living room	C,D	4
H:	Decorate kitchen	F	3
I:	Decorate bathroom	F	2

***Required**;*

Draw a network for the above project and identify the critical path.

17.4 Gilbert is the project manager of Kaka Construction Company. The company is bidding on a contract to install telephone lines in a small town. It has identified the following activities along with their predecessor restrictions, expected times and worker requirements.

	Activity	**Preceding Activity**	**Duration (days)**	**Crew size workers**
A:	Replace windows	-	4	4
B:	Re-wiring	-	7	2
C:	Re-plastering	A	3	2
D:	Fit-lights	A	3	4
E:	Decorate bedrooms	B	2	6
F:	Install plumbing	B	2	3
G:	Decorate living room	D,E	2	3
H:	Decorate kitchen	F,G	3	4

Gilbert has agreed with the client that the project should be completed in the shortest duration.

***Required**;*

(i) Draw a network for the project.

(ii) Determine the critical path and the shortest project duration.

(iii) Gilbert will assign a fixed number of workers to the project for its entire duration and so he would like to ensure that the minimum number of workers is assigned and that the project will be completed in 14 weeks. Draw schedule showing how project will be completed in 14 weeks

(iv) Comment on the schedule you have drawn in (ii) above.

PART H

LINEAR ALGEBRA & CALCULUS

Chapter 22
LINEAR ALGEBRA AND CALCULUS

22.1: Introduction

This chapter is divided into 2 main parts i.e.;

(i) Linear Algebra, and
(ii) Calculus

Where, linear algebra covers linear equations, quadratics and simultaneous equations; while Calculus discusses differentiation and integration.

Linear algebra has not been included in this chapter since all the parts have been integrated in earlier chapters of the text. However trial questions and solutions have been provided at the back of the text.
Calculus has been covered fully; Integration has been included to help the reader conceptualise the techniques applied in continuous random variables.

22.2: DIFFERENTIATION

Differentiation establishes the slope of a function at a particular point.

Its denoted by; $\boldsymbol{\frac{dy}{dx}}$

Where;
y = dependent variable and
x = independent variable

Consider a function; $\boldsymbol{y = mx + c}$, then;

$$\boldsymbol{\frac{dy}{dx}} = \boldsymbol{m}\ (gradient\ of\ the\ function)$$

Therefore, $\boldsymbol{m}$ is the gradient of the linear relationship and it is known as the *derivative* of the function.
Derivatives can be obtained by other relationships i.e. parabolic and exponential.

For example, the derivative for a parabolic relationship $\boldsymbol{y = ax^2 + bx + c}$ will be;

$$\frac{dy}{dx} = 2x + b$$

where $2x + b$ is the derivative of the parabolic relationship.

22.3: Rules of finding derivatives

(i) Basic Rule

Where the function is $\boldsymbol{y = x^n}$ then

$$\boldsymbol{\frac{dy}{dx} = nx^{n-1}}$$

E.g. $y = x^2$ ***And*** $y = x^2$

$\frac{dy}{dx} = 2x$ $\frac{dy}{dx} = 10x^9$

(ii) Functions with coefficients

If is $\boldsymbol{y = kx^n}$ (*k* being a coefficient) then;

$$\frac{\boldsymbol{dy}}{\boldsymbol{dx}} = \boldsymbol{nkx^{n-1}}$$

E.g. $y = 3x^2$ ***and*** $y = 8x^4$

$\frac{dy}{dx} = 6x$ $\frac{dy}{dx} = 32x^3$

(iii) Functions with constants

If $\boldsymbol{y = x^n + c}$ and ***c*** represents a constant, then:

$$\frac{\boldsymbol{dy}}{\boldsymbol{dx}} = \boldsymbol{nx^{n-1}}$$

E.g. $y = x^3 + 8$ ***and*** $y = 6x^4 + 27$

$\frac{dy}{dx} = 3x^2$ $\frac{dy}{dx} = 24x^3$

Note:
Constants disappear on differentiation of functions with constants.

(iv) Where the function is a sum

If $\boldsymbol{y = x^n + x^m}$; then;

$$\frac{\boldsymbol{dy}}{\boldsymbol{dx}} = \boldsymbol{nx^{n-1} + mx^{m-1}}$$

E.g. $y = x^2 + 6x^4$ ***And*** $y = \frac{1}{6}x^2 - 12x^5$

$\frac{dy}{dx} = 2x \quad + 24x^3$ $\frac{dy}{dx} = \frac{1}{3}x \quad - 60x^4$

(v) Where the function is a product

If ***m*** and ***n*** represents functions of ***x*** and ***y = mn***; then;

$$\frac{\boldsymbol{dy}}{\boldsymbol{dx}} = \frac{\boldsymbol{d(mn)}}{\boldsymbol{dx}} = \boldsymbol{m}\frac{\boldsymbol{dn}}{\boldsymbol{dx}} + \boldsymbol{n}\frac{\boldsymbol{dm}}{\boldsymbol{dx}}$$

E.g. $y = (8x + 4)(2x^3 + 6)$

Let ***m*** = $8x + 4$ and

n = $2x^3 + 6$

then;

$\frac{dm}{dx} = 8$ ***And*** $\frac{dn}{dx} = 6x^2$

$\frac{dy}{dx} = (8x + 4)(6x^2) + (2x^3 + 6).8$

$= \boldsymbol{64x^3 + 24x^2 + 48}$

(vi) Where the function is a quotient

If ***m*** represents the function of ***x*** which is the numerator and ***n*** represents the function which is the denominator, then:

$$\boldsymbol{y = \frac{m}{n}}$$

$$\boldsymbol{\frac{dy}{dx} = \frac{n\frac{dm}{dx} + m\frac{dn}{dx}}{n^2}; n \neq 0}$$

$$E.g. \qquad y = \frac{4x^3 + 2}{x^6} \quad where \ m = 4x^3 + 2 \ and \ n = x^6$$

$$\frac{dm}{dx} = 12x^2 \qquad \textbf{\textit{And}} \qquad \frac{dn}{dx} = 6x^5$$

$$\frac{dy}{dx} = \frac{x^6.(12x^2) - (4x^3 + 2).(6x^5)}{(x^6)^2}; \ n \neq 0$$

$$= \frac{\boldsymbol{12 + 12x^3}}{\boldsymbol{x^7}}$$

(vii) Function of functions

If $\boldsymbol{y = (2x + 6)^3}$ and the expression in brackets is a differentiable function, say $\boldsymbol{m = 2x + 6}$, the whole expression can be written as $\boldsymbol{y = m^3}$

i.e.

$$y = m^3 \quad where \ m = 2x + 6$$

Then;

$$\frac{dy}{dx} = \frac{dy}{dm} \times \frac{dm}{dx}$$

$$\frac{dy}{dm} = 3m^2 \ and \ \frac{dm}{dx} = 2$$

Then;

$$\frac{dy}{dx} = \frac{dy}{dm} \times \frac{dm}{dx} = 3(2x + 6)^2.2$$

$$= \boldsymbol{6(2x + 6)^2}$$

22.4: PRACTICAL APPLICATION OF DIFFERENTIATON

There are many practical areas in business where differentiation comes into play at a greater extent. These are situations where one variable changes in response to changes in man-hours employed, or in theory of the firm costs may change in response to output changes.

Differentiation helps to show the rate of change of the dependent variable (e.g. cost) with respect to an infinitely small increment in the value of the independent variable (e.g. activity like production).

In economics small increments of revenue and cost are known as marginal revenue and marginal cost respectively. Such small increments are obtained by the principle of differentiation following the same rules described above.

SMALL CHANGES IN ECONOMIC VARIABLES

(i) Marginal cost (MC)

Marginal cost is the measure of the rate of change of cost as output changes.
Therefore, an extra cost incurred when an extra unit of output is produced is what is termed as *marginal cost*.
Marginal cost is the gradient of the total cost curve and is obtained by differentiating the cost curve.

Example 22.1

Suppose that the total cost of producing *Q* units of a commodity is represented by a function of *Q* which is denoted by *C = 500 + 3Q*. The marginal cost of this function will be given as;

$$\boldsymbol{MC = \frac{dC}{dQ} = 3}$$

Therefore, the extra cost that will be incurred in production of an extra unit of output in the above particular case is Shs. 3/-. You will discover that this is the variable cost of production.
Furthermore, Shs. 500 in the cost function represents the fixed cost.

However, example 1 presents a situation where *MC* is constant. Example 2 below illustrates a situation where *C* contains higher powers of *Q*, where *MC* itself is a function of *Q*.

Example 22.2

Given that the total cost of producing *Q* units of a commodity is represented by a cost function $C = 200 + 5Q+3Q^2$. The marginal cost of this function will be given as;

$$MC = \frac{dC}{dQ} = 6Q + 5$$

Therefore, the extra cost that will be incurred in production of an extra unit of output in the above particular case is itself a function of Q i.e. 6Q + 5.

(ii) Average cost (AC)

Average cost is the cost per unit of output.

The average cost function is denoted by;

$$\boldsymbol{AC = \frac{C}{Q}}$$

Where *C* = total cost and
Q = Units produced / output

Example 22.3

The cost function of XYZ Ltd is given as $C = 50 + 4Q + Q^2$. Find

(i) The fixed cost
(ii) The variable cost
(iii) The marginal cost
(iv) The average cost

Solution

(i) The fixed cost is the cost of producing zero units, so it can be found by substituting Q = 0 in the total cost function, which gives in this case;
Fixed costs = $50 + 4(0)+ (0)^2$ = **Shs. 50**.

(ii) The variable cost is the remaining part of the total cost function i.e. $4Q+Q^2$. This is the function which varies with the quantity produced.

(iii) Marginal cost is given by

$$MC = \frac{dC}{dQ} = 4 + 2Q$$

(iv) Average cost is given by;

$$AC = \frac{C}{Q} = \frac{50 + 4Q + Q^2}{Q} =$$

$$= \frac{\mathbf{50}}{\boldsymbol{Q}} + \mathbf{4} + \boldsymbol{Q}$$

(iii) Marginal Revenue (MR)

Marginal revenue is the measure of the rate of change of Revenue as output changes.
In other words, extra revenue earned by the firm as a result of producing an extra unit of output is what is termed as marginal revenue. Marginal revenue is the gradient of the revenue curve and is obtained by differentiating the revenue function as shown below;

$$\boldsymbol{MR = \frac{dR}{dQ}}$$

Where *Q* is the quantity of output

(iv) Average Revenue (AR)

Average Revenue is the Revenue per unit of output. The average revenue function is denoted by;

$$\boldsymbol{AR = \frac{R}{Q}}$$

Where *R* is the revenue function and *Q* being the quantity of output
Average revenue is also represented by the symbol ***P*** since it is the price per unit. The price ***p*** as a function of ***Q*** (quantity) expresses the link between price and quantity demanded and is referred to as the *demand function*.

(v) Point of maximum profit

This is the point where the firm's marginal cost (MC) is equal to its marginal revenue (MR).

At this point, the gradient for both the cost and revenue functions intersect each other;
i.e;

$$\frac{\boldsymbol{dC}}{\boldsymbol{dQ}} = \frac{\boldsymbol{dR}}{\boldsymbol{dQ}}$$

$$\boldsymbol{MC = MR}$$

(vi) Breakeven point

Break even occurs when the firm's total cost is equal to the firm's total Revenue.
At this point, the firm is deemed to be making no profit or loss i.e. profit / loss = 0. The quantity required to be produced by the firm to break even is obtained by equating the Cost function to the revenue function.

Example 22.4

The revenue function of ABC Ltd is given by $R = 400Q - 4Q^2$ and the cost curve is given by $C = Q^2 + 10Q + 30$; where Q is the number of units sold.
Required;

(i) Compute the firm's marginal revenue and marginal cost functions.
(ii) Determine the quantity at that should be sold in order to maximise profits
(iii) At what price should each unit be sold to maximise profits?
(iv) Compute the amount of profit that will be maximised.
(v) What is the firm's break even quantity?

Solution

(i) Marginal revenue and marginal cost;

$$MR = \frac{dR}{dQ} = \mathbf{400 - 8Q} \qquad and \qquad MC = \frac{dC}{dQ} = \mathbf{2Q + 10}$$

(ii) Quantity to be sold a maximum profit level.

At Maximum profit;

$$MR = MC$$

$$400 - 8Q = 2Q + 10$$

From which; **Q = 39 units**

(iii) Price per unit output at maximum profit level

$$Price = \frac{Revenue}{Number\ of\ units} = \frac{400(39) - 4(39)^2}{39}$$

$$= \frac{15{,}600 - 6.084}{39}$$

$$= \boldsymbol{Shs.244\ per\ unit}$$

(iv) Amount of profit to be maximised

$$Profit = Revenue - Cost$$

$$= 9{,}516 - (39)^2 + 10(39) + 30$$

$$= \boldsymbol{Shs.7{,}575}$$

Example 22.5

A firm has the following average revenue curve function (demand Curve):
P=100 – 0.01Q, where *Q* is the weekly production and *P* is the price in shillings per unit.
The firm's cost function is given by *C = 50Q + 30,000*.

Required;
Assuming the firm maximises profits, determine the equilibrium level of production Q, and profits per week.

Solution

- **The revenue function**

$$Average\ revenue = \frac{\boldsymbol{Revenue}}{Quantity} = p$$

Therefore the Revenue curve will be given as;

Revenue function = Price × Quantity

= (100 – 0.01Q) × Q

$R = 100Q - 0.01Q^2$

- **Equilibrium level of production**

$$MR = \frac{dR}{dQ} = \mathbf{100 - 0.02Q} \qquad and \qquad MC = \frac{dC}{dQ} = \mathbf{50}$$

At maximum profit;

$$MR = MC$$

$$100 - 0.02Q = 50$$

From which; **Q = 2,500 units**

- **Price per unit**

From the average revenue function;
P = 100 – 0.01Q
= 100 – 0.01 ×2,500
= **Shs. 75**/-

- **Profits per week**

Profit = Revenue - cost

$$= [100(2{,}500) - 0.01(2{,}500)^2] - [50(2{,}500) + 30{,}000]$$
$$= 187{,}500 - 155{,}000$$
$$= \boldsymbol{Shs.32,500}$$

22.5: MINIMUM COST AND MAXIMUM REVENUE

A firm may need to know the quantity at which it incurs a minimum cost or a point at which it earns the maximum revenue.

These can only be obtained by looking at the turning points of the respective functions. The turning points are the points of minimum cost or maximum revenue.
These points are points of Zero slope or gradient. If the derivative of the respective function is equated to zero, then the resulting value will represent a turning point.

For functions with more than one turning point, a second derivative may be required whereby if the second derivative is negative, the turning point is a maximum and if the second derivative is positive, the turning point is a minimum.

Second derivatives are denoted by;

$$\frac{\boldsymbol{d^2y}}{\boldsymbol{dx^2}}$$

Steps in finding the maximum and minimum value of a function

Step 1: - Find the first derivative of the function i.e. $\frac{dy}{dx}$

Step 2: - Set $\frac{dy}{dx}$ to Zero and calculate the turning point.

Step 3: - Find the second derivative, i.e. $\frac{d^2y}{dx^2}$, by differentiating the first derivative.

Step 3: - If $\frac{d^2y}{dx^2}$ is negative, the turning point is maximum and if $\frac{d^2y}{dx^2}$ is positive, the turning point is minimum.

Example 22.6
Find the point of maximum value of the Revenue function $R = 400Q - 4Q^2$

Solution

Step 1: -

$$\frac{dr}{dQ} = 400 - 8Q$$

Step 2:-
At the turning point $\frac{dr}{dQ} = 0$

Therefore; 400 – 8Q = 0, whereby Q = **50**
Step 3:-

$$\frac{dr}{dQ} = 400 - 8Q \; therefore$$

$$\frac{d^2r}{dQ^2} = -8 \; i.e. negaive \; for \; all \; values \; of \; Q.$$

Step 4:-

As $\frac{d^2r}{dQ^2}$ is negative, the turning point when Q = 50 is a maximum and the revenue at that point is;

$$R = 400Q - 4Q^2$$

$$= 400(50) - 4(50)^2$$

$$= \boldsymbol{Shs.\,10,000}$$

22.6: PARTIAL DIFFERETIATION

Partial differentiation applies in situations where a function contains more than one independent variable. E.g.

$$y = x^2 + 2z^2$$

Where x and z are independent variables.

In the above case, we differentiate each variable separately, keeping the others as a constant; C.
e.g.

$$for \; y = x^2 + 2z^2 \quad we \; let \; C = 2z^2$$

Then we differentiate the function $y = x^2 + C$ to get the function $\frac{dy}{dx} = 2x$.
Proceed to differentiate the second variable by having $C = x^2$ so that the function is represented as $y = C + 2z^2$ hence $\frac{dy}{dZ} = 4Z$.

The partial derivatives will therefore be 2x and 4Z.

EXERCISE 18

18.1

Company A sells all its output to company B for Shs. 200 per unit. The cost of the sales per week in company A are given by the function $C=2q^2 + 40q + 80$ where q is the value of weekly sales. Company B uses the output of company A to manufacture a product whose demand is dependent on the sale price. The revenue per week of company B is given by the function;

$R=1000q - 16q^2$ and the cost per week of company B excluding cost of the products bought from company A are given by the function;

$C=2q^2 + 80q + 400.$

Company A can restrict the weekly supply of its product to company B, but cannot raise the unit price above Shs. 200. The two companies are considering whether to merge together into a single company.

Required;

(i) *At what weekly sales would company A maximise profits?*

(ii) *What would be the profit or loss of company B if company A were able to supply the profit maximizing quantity of its products each week.*

(iii) *At what level of weekly sales would company B maximise profits?*

(iv) *If the two companies merge into one, what would be the profit maximizing output per week and what would be the weekly profit?*

18.2

An electronics firm carries out a small scale test launch of a low priced pocket calculator. It estimates from this test that if it went full-scale production it would sell between 1,000 and 2,500 calculators per month and that the monthly revenue in thousands of shillings over this range of sales could be represented by the equation $R = -x^2 + 5x$ where x is the monthly output in thousands of calculators (it is assumed that it sells the entire output).

From experience of calculator production, the firm estimates that its marginal costs in thousands of shillings could be represented by the equation $MC = x^2 - x + 2$ and that its fixed costs will be Shs. 500 per month.

Required

(i) *Determine the average cost and revenue equations for this firm.*

(ii) *Determine the profit maximizing output, the price that should be charged to maximise profit, and how each calculator will then cost to make.*

18.3

Kato and Wasswa certified public Accountants have recently started to give business advice to their clients. Acting as consultants, they have estimated the demand curve of a client firm to be;

$$AR = 200 - 8Q$$

where AR = Average Revenue in millions of shillings and Q is the output in units.

Investigations of the firm's cost profile show the marginal cost (MC) given by

$$MC = Q^2 - 28Q + 211 \ (in\ millions\ of\ shillings)$$

Further investigations have shown that the firm's cost when not producing is Shs. 10 million.

Required;

(i) *The equation of total cost.*

(ii) *The equation of total revenue.*

(iii) *An expression for profit.*

(iv) *The level of output that maximise profit.*

(v) *The equation of marginal revenue*.

18.4 At ABC factory a machine costs USD 12,000. The operational cost from the time of purchase of the machine at anytime t years is given by $20t^2 + 15t$. The machine's resale value at anytime is given by $6{,}880 - 60t^2$.

Required

(i) Express the average cost of the machine at the time of replacement of the machine in terms of t.

(ii) Find the optimum time for the replacement of the machine.

18.5 A chemical company is planning to launch a new fertilizer and carries out a market research in order to determine the likely demand. The survey indicates that the company can expect to sell between 1,000 and 2,000 tons per month and that the relationship between price charged and quantity demanded will be as follows;

Price (Shs '000' per ton	16	15	14	13	12
Monthly demand in thousands of tons	1.00	1.25	1.50	1.75	2.00

The company estimates that marginal cost (in Shs '000') of producing the fertilizer can be represented by the equation;

$$MC = 2x^2 - x + 5$$

Where x is the monthly output in thousands of tons. The fixed costs will be Shs. 100,000 per month.

Required

(i) Determine the quantity and price which should be produced and charged respectively to maximise revenue and profit.

(ii) Determine the maximum revenue and profit.

(iii) Why is the value of maximum revenue different from maximum profit in (ii) above?

PART I
DECISION THEORY

Chapter 23
DECISION MAKING

23.1: Introduction
Decision making is a process of selecting from a set of alternative courses of action, the best option that is considered to meet the objectives of the decision problem more satisfactorily than others as judged by the decision maker.

The best possible decision (or option) is arrived at through abstraction, observation, investigation, experimentation, measurement and analysis of data in the relevant real world environment.

Therefore a good decision is always based on logic, uses data, applies techniques of quantitative analysis and considers all possible alternatives.
A good decision may sometimes result in unexpected unfavourable outcomes but it still remains a good decision whereas a bad decision may sometimes lead to favourable occurrences but still it remains a bad decision.

23.2: Decision making concepts

(i) Decision maker
This is an individual or group of persons charged with the responsibility of making the decision by selecting one best alternative from a set of possible courses action.

(ii) Alternatives (or acts)
These are events or outcomes that are available to the decision maker and are within the control of the decision maker e.g. construction of a large or small facility launch or do not launch the product.

(iii) States of nature (SN)
States of nature are the circumstances (events) that affect the outcome of the decision, but are beyond the control of the decision maker. e.g. Strong market Vs poor market.

(iv) Outcome / payoff
For each combination of an act and event there will exist an outcome. It may be expressed in terms of profits, present value or in terms of some non-monetary measure. We refer to the outcome as payoff.

23.3: The Decision making process
Before a decision can be taken, the decision maker must carry out the following steps;

(i) *Define the problem at hand*.
This involves making clear and concise statements about the problem.

(ii) *Gather data on the problem.*
This involves collection of data about the current awareness of the problem. Several techniques can be employed to execute this process.

(iii) *List alternative courses of action*.
Having defined the problem and collected data about it, it is now possible to develop a list of all possible alternatives to making the decision.

(iv) *Identify possible outcomes*.
Where possible, all outcomes should be identified for informed decision.

(v) Estimate the profits or payoffs of the alternatives.
Estimation of profits or payoffs will require one's knowledge of the market behaviour.

(vi) *Develop a mathematical model* and use the model to generate alternative solutions to the problem.

(vii) *Choose from the alternative solutions*.

23.4: The decision making Environments

A decision making environment is simply a combination of all forces that may affect the best alternative choice. Therefore, it is of significant importance for any decision maker to be aware of such forces governing the decision environment.

There are basically three types of decision making environments and these include;

(i) Decision making under certainty
(ii) Decision making under risk.
(iii) Decision making under uncertainty

These are explained as below;

(i) Decision making under certainty.

This is a situation in which the decision maker knows for certain the consequences of every alternative taken. The decision maker will always choose the best alternative to maximise his objective in case of profits or to minimize the objective in case of loss or costs.

In other words, there would exist only one outcome for the decision maker. For example if an investor knows for sure that there is better business in a service sector than in farming, then he will invest in the service sector which provides better returns.

(ii) Decision making under Risk.

This is a situation in which the decision make knows the probability with which each outcome is likely to happen. If the probability of an occurrence is known, then the decision maker will attempt to maximise his objective by maximizing the expected monetary value i.e. the EMV or by minimising the expected loss.

(iii) Decision making under Uncertainty.

This is a situation where nothing is known about any choice that may be taken.
Under conditions of uncertainty, the decision maker is constrained because of the lack of knowledge about the probabilities of occurrence of the outcome of possible courses of action.

For example, a potential investor may not know the political environment of the country in the next five year period to necessitate him to take a proper investment decision.

23.5: Theorems under decision making Environments

23.5.1 Decision making under risk

There are two criteria that the decision maker may use under the conditions of risk and these include;

(i) Expected monetary value, EMV and
(ii) Opportunity loss

These are explained as below;

(i) **Expected monetary value, EMV**

Once the conditional values or payoffs and probabilities of each state of nature (SN) and for each alternative are given (in a decision table), the EMV of each alternative is the product of the payoff of the states of nature by their respective probabilities.

In other words;

EMV (alternative k) = *(payoff of SN_1) × $P_r(SN_1)$ + payoff of SN_2 × $P_r(SN_2)$ + ……+ payoff last SN × P_r(last SN).*

Expected monetary value is further discussed under two headings;

- **Expected value of perfect information (EVoPI)**

If the decision maker knows what state of nature was going to prevail, he or she would simply choose the strategy giving the best return for the state of nature.

The *perfect information* about the state of nature would help to remove all the uncertainties, and hence the decision maker's problems would be solved.
Unfortunately, this is not possible without incurring a cost for market research or some other appropriate method of investigation about the state of nature.

By finding out how much better off the decision maker would be if he or she had perfect information helps in setting an upper limit (bound) on what it is worth paying to seek real information.

Definition:
The expected value of perfect information refers to the maximum or minimum payment a decision maker may require as payment for provision of any technical information that may be employed in order to make any decision environment change from uncertainty into that of certainty.
In otherwords;

$$\mathbf{EV_oPI = EV_wPI - EMV_{max}}$$

Where;
EV_wPI is the Expected monetary value with perfect information.
EMV_{max} is the maximum expected monetary value.

- **Expected value of perfect information (EVoPI)**

This is the upper price limit (bound) paid for any perfect information used in making the decision.
It is computed using the formula below;

EV_wPI = (Best outcome of SN_1) × (Probability of SN_1) +
(Best outcome of SN_2) × (Probability of SN_2) + ……. +
(Best outcome of SN_n) × (Probability of SN_n)

Example 23.1
XYZ Ltd wishes to set up a processing plant in order to launch its newly developed production Uganda. The managers have come up with the following information about the possible courses of action and the corresponding payoffs.

Alternatives	**Strong Market**	**Poor Market**
Large plant	200,000	-180,000
Small Plant	100,000	-20,000
Do nothing	0	0
Probabilities	*0.6*	*0.4*

XYZ Ltd has contacted *Kasiita,* a consultancy firm, which is willing to charge Shs. 72,000 for the cost of provision of perfect information about the states of nature.
Required;

(a) Compute
- *(i) The Expected monetary value (EMV) for each alternative and show the EMV_{Max}*
- *(ii) The Expected value with perfect information(EV_wPI)*
- (iii) *The Expected value of perfect information(EV_oPI)*

(b) Should XYZ obtain the perfect information from Kasiita? Give reason why

Solution

(a)

(i) EMV for each alternative; *from the formula*;

EMV (alternative k) = *(payoff of SN_1) × $P_r(SN_1)$ + payoff of SN_2 × $P_r(SN_2)$ + ……+ payoff last SN × P_r(last SN).*

Where $\mathbf{SN_1}$ = Strong market and $\mathbf{SN_2}$ = poor market

$$EMV(large\ plant) = (200{,}000 \times 0.6) + (-180{,}000 \times 0.4)$$
$$= 120{,}000 + (-72{,}000) = \mathbf{48{,}000}$$

$$EMV(Small\ plant) = (100{,}000 \times 0.6) + (-20{,}000 \times 0.4)$$
$$= 60{,}000 + (-8{,}000) = \mathbf{52{,}000} *** \boldsymbol{EMV_{Max}}$$

$$EMV(Do\ nothing) = (0 \times 0.6) + (0 \times 0.4) = \mathbf{0}$$

(ii) EV_wPI ; from the formula

EV_wPI = (Best outcome of SN_1) × (Probability of SN_1) +
(Best outcome of SN_2) × (Probability of SN_2) + ……. +
(Best outcome of SN_n) × (Probability of SN_n

Where

- the best outcome of SN_1(Strong Market) = **120,000**
- the best outcome of SN_2(Poor market) = **0**

We shall then multiply these with the corresponding probabilities and sum the results to give the *EV_wPI* as shown below;
Hence;

EV_wPI = (120,000 × 0.6) + (0 × 0.4)
= **Shs. 120.000**

(iii) EV_oPI ; from the formula

EV_oPI = 120,000 – 52,000
= **Shs. 68,000**/-

(b) From the above computations, *XYZ* can only pay a maximum of Shs. 68,000, meaning that there is no need to look for perfect information about the strong markets and poor markets in Uganda from *Kasiita* since their charge of Shs. 72,000 is significantly higher than the EV_oPI.

(ii) Opportunity loss

For a given event, the difference between the payoff of the most favourable act and some other act may be termed as loss due to losing the opportunity of choosing the best alternative.

To use the opportunity loss approach to make a decision, one must calculate the Expected opportunity loss (EOL), and pick the alternative that gives the minimum EOL.

The opportunity loss table is computed and from which the EOL for each alternative is then obtained.

Steps in computing the Expected Opportunity Loss

There are basically two steps in computing the EOL;

(i) For each SN, subtract each outcome in the column from the best outcome in that column. Draw the opportunity loss table.
(ii) Compute the EOL for each alternative by multiplying the probability of each SN by the appropriate opportunity loss value, then select the alternative that gives the minimum EOL.

Example 23.2
Rework example 23.1 and determine;
(i) The EOL for each alternative
(ii) Advise XYZ on the best alternative using the EOL.

Solution

	States of Nature	
Alternatives	**Strong Market**	**Poor Market**
Large plant	200,000-200,000	0-(-180,000)
Small Plant	200,000-100,000	0-(-20,000)
Do nothing	200,000-0	0-(0)

- *Now construct the opportunity loss table*;

	States of Nature	
Alternatives	**Strong Market**	**Poor Market**
Large plant	0	180,000
Small Plant	100,000	20,000
Do nothing	200,000	0
Probabilities	*0.6*	*0.4*

- *The EOL Workings*;

$EOL(large\ plant) = (0.6 \times 0) + (0.4 \times 180{,}000) \quad = \mathbf{72{,}000}$

$EOL(Small\ plant) = (0.6 \times 100{,}000) + (0.4 \times 20{,}000) = \mathbf{68{,}000}$*** *minimum EOL*

$EOL(Do\ nothing) = (0.6 \times 200{,}000) + (0.4 \times 0) \quad = \mathbf{120{,}000}$

The minimum **EOL** is Shs. **68,000/-** corresponding to the alternative of building a small plant. Hence XYZ can launch their new product in Uganda by constructing a small production facility.

Point to note

Decision making under the EOL technique is always the same as that from the EMV approach and that the EV_oPI is the same as the minimum EOL value.

Decision making under uncertainty

Conditions of uncertainty as described before, imply that the decision maker does not know what conditions will prevail for the working out of his chosen strategy.
For example, a pricing decision on a new product would be eased if the intentions of competitors were known.

The major criterion usually adapted while making decisions under conditions of uncertainty include but not limited to;

- **Maximin criteria**

This is a pessimistic criterion based upon the conservative approach to assume that the worst possible is going to happen.

Using this method, we identify the minimum of the payoffs for every alternative and then select the maximum amongst them; this will be the alternative that maximises the minimum outcome / payoff for every alternative.

Therefore, the decision maker considers each strategy and locates the minimum pay for each and then selects that alternative which maximizes the minimum payoff.

Example 23.3:
Looking at example 23.1, a decision maker would choose the "do nothing" strategy since it gives the maximum value of the smallest gains/losses.

- **The Maximax criterion**

The Maximax criterion is based upon extreme optimism. The decision maker selects the particular strategy which corresponds to the maximum of the maximum payoff for each strategy.

Therefore, the decision maker will proceed by identifying the maximum payoff for each alternative and then follow by selecting the maximum (greatest) among them; this will be the alternative that maximises the outcome of every alternative.

Example 23.4:
Looking at example 23.1, a decision maker would choose the "large plant" strategy since it gives the maximum value of the maximum payoffs.

- **The Hurwicz criterion**

In order overcome the disadvantages of extreme pessimism of maximin and extreme optimism of maximax criterion, K. Hurwicz introduced the concept of coefficient of optimism or pessimism.

This allows the decision maker to take into account both the maximum and minimum payoff to the degree of optimism or pessimism. The alternative which maximizes the sum of the weighted payoffs is then selected.

- **The Laplace criterion**

The above three decision rules ignore most of the information in the payoff matrix, as they focus on either the best or the worst that could happen. The Laplace criterion uses all the information by assigning equal probabilities to the possible payoffs for each action and then selecting that alternative which corresponds to the maximum expected payoff.

- **The Minimax regret criterion**

It is based upon the assumption that the decision maker might experience 'regret' after he has made the decision and the events have occurred. The criterion follows the opportunity loss approach and proceeds by identifying the alternative which minimizes the maximum opportunity loss within each alternative.

After computing the opportunity loss values, pick out the maximum per row and the minimax is selected from the alternative with the minimum value from the column of maximum (row) values.

***Example 23.5**:*
Looking at example 23.2, the opportunity loss table for minimax decision will appear as below;

	States of Nature		Maximum in the
Alternatives	**Strong Market**	**Poor Market**	**row**
Large plant	0	180,000	180,000
Small Plant	100,000	20,000	100,000
Do nothing	200,000	0	200,000

The minimum from the column of maximum values is thus; Shs. 100,000/=. The alternative corresponding to this value is selected. Therefore, the decision maker will choose the "*small plant*" strategy.

23.6: DECISION TREES

A decision tree is a pictorial display of the various decision alternatives and their corresponding outcomes. All the possible choices are shown on the tree as branches and the possible outcomes as subsidiary branches.

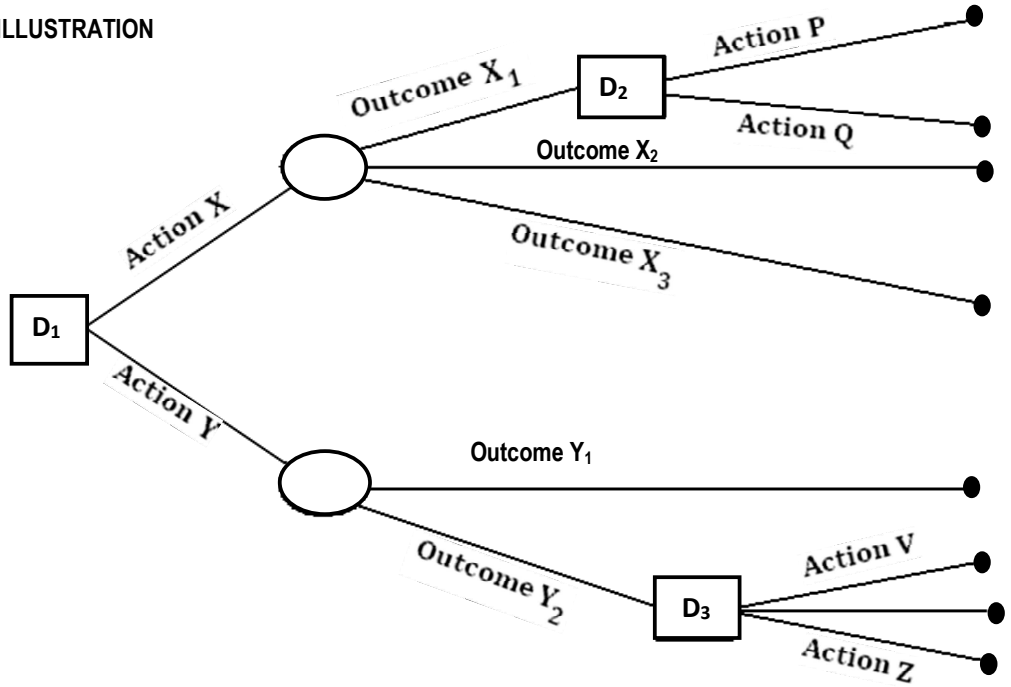
The difference between a "probability tree" and a "decision tree" is that the probability tree will only show the probability of an event occurring and will not take into account the outcome (actions) in decision making.

Structure of decision trees

It is a convention to use the symbol ' □ 'to indicate the decision point and ' ○ ' to denote the situation of uncertainty or event or outcome.
Branches coming out of a decision point are immediate mutually exclusive alternative acts (action) open to the decision maker.

Branches coming from the event point ○ represent all possible outcomes or situation or events.

ILLUSTRATION



When the probabilities of the various events are known, they are written along the corresponding branches.

A decision tree diagram is simply a probability tree with the chosen alternatives indicated at the respective branches.

Steps in drawing up a decision tree

Step 1; - Draw the tree showing the various decision points and outcome nodes.
The tree is drawn from left to right (*i.e. forward pass*)

Step 2; - Fill the outcomes values and probabilities

Step 3; - Use the probabilities and outcome values to calculate the expected values at various points so that the correct decisions are highlighted.
This stage works from right to left (*i.e. backward pass*)
At this stage the tree shows the expected values at the outcome nodes from which the highest expected values from the subsequent part of the tree indicate the best decision that can be taken.

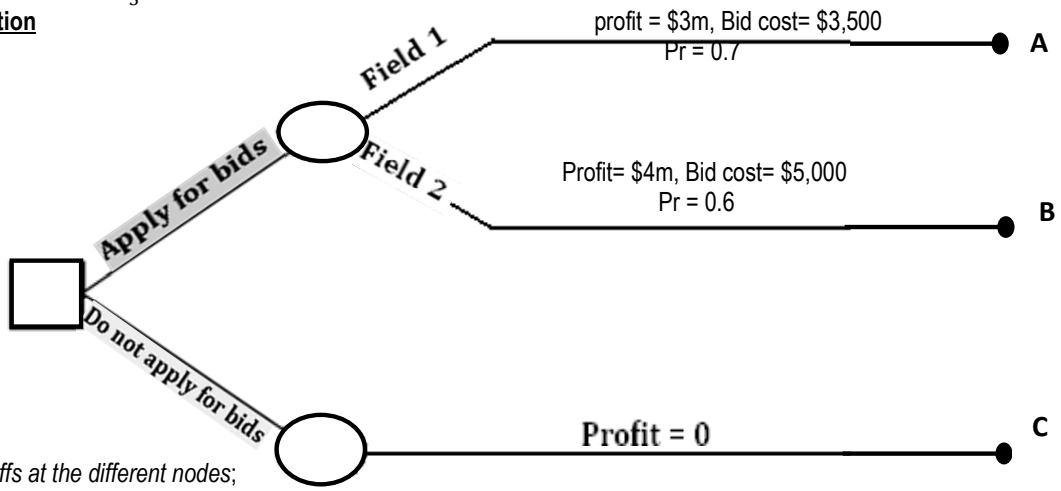
The above steps are illustrated in the examples below;

Example 23.6

An oil company may bid for drilling one of the two contracts for drilling in two different areas in Western Uganda. It is estimated that the profit of $3 million per month would be realized in the first field and $4 million in the second field. These profit figures were determined ignoring the cost of bidding which amounts to $3,500 for the first field and $5,000 for the second field.

Required:
Advise the oil company on which field they should bid for if the probability of winning a contract for the field is $\frac{7}{10}$ *and the second field is* $\frac{3}{5}$.

Solution



Payoffs at the different nodes;

$\boldsymbol{Node\ A} = (3m - 3{,}500) \times 0.7 = \$2.1m$

$\boldsymbol{Node\ B} = (4m - 5{,}000) \times 0.6 = \$2.4m$

$\boldsymbol{Node\ C} = 0$

The payoffs are obtained by multiplying the respective probabilities with the net estimated profit. The net estimated profit is obtained by subtracting the bidding cost from the gross estimated profit.
The company should bid for oil drilling in the second field since it has the highest payoff (expected monetary value).

Example 23.7

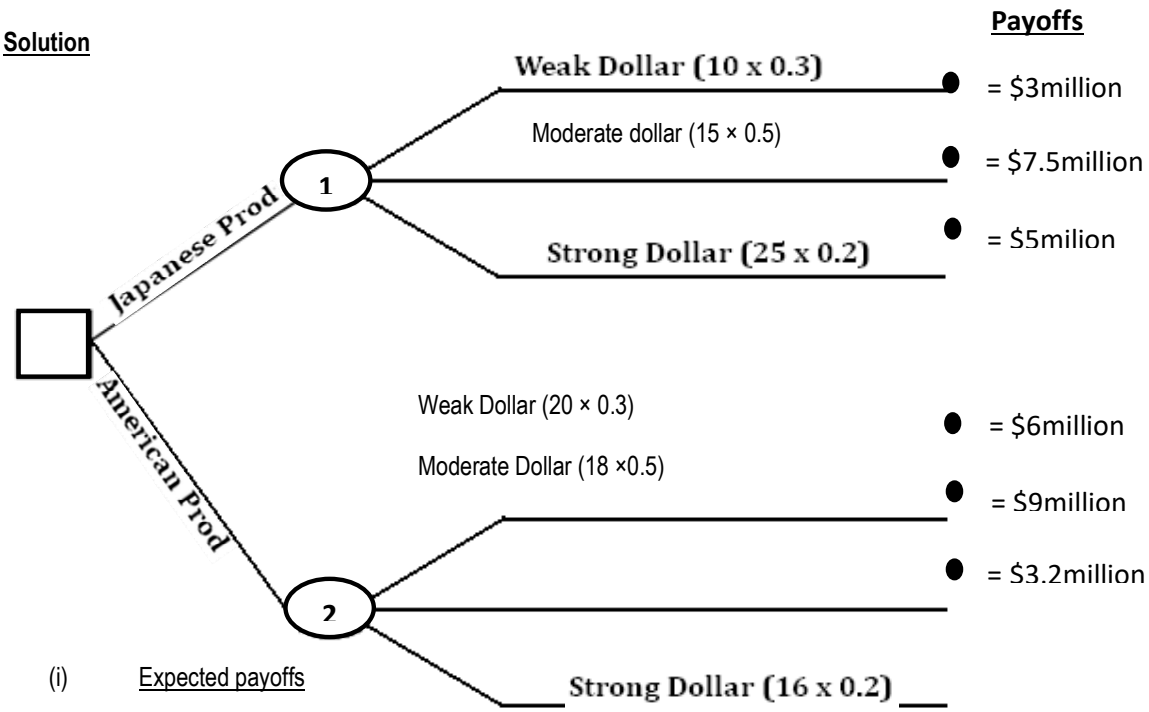
A Japanese automobile manufacture currently produces its best selling U.S model in Japan, but the relative strength of the Japanese Yen verses the U.S dollar has been making the car very expensive for the U.S market. To ensure lower and more stable prices, the company is considering the possibility of manufacturing cars for U.S consumers at one of its American plants. Its payoff table, in millions of dollars, is given below;

		States of Nature	
Alternatives	Weak dollar Probability = 0.3	Moderate Dollar Probability = 0.5	Strong Dollar Probability = 0.2
Japanese Production	10	15	25
American Production	20	18	16

Required

(i) *Make a decision tree summary of the table above.*
(ii) *Compute the expected payoff for each production.*
(iii) *Compute the expected payoff with perfect information.*
(iv) *Compute the expected value of perfect information.*

Solution



(i) Expected payoffs

$\boldsymbol{Node\ 1}$: $Japanese\ Production = 3 + 7.5 + 5 = \$\mathbf{15.5\ million}$

$\boldsymbol{Node\ 2}$: $American\ production = 6 + 9 + 3.2 = \$\mathbf{18.2\ million} \ *** \ EMV_{max}$

(ii) Expected payoff with perfect information

EV_wPI = (Best outcome of SN_1) × (Pr SN_1) + (Best outcome of SN_2) × (Pr SN_2) +(Best outcome of SN_3) × (Pr SN_3)

Where; N_1 = Weak Dollar, N_2 = Moderate Dollar and N_2 = Strong Dollar

Hence;

= (20 × 0.3) + (18 × 0.5) + (25 × 0.2)

= 6 + 9 + 5 = **$20 million**

(iii) Expected value of perfect information

EV_oPI = $EV_wPI - EMV_{max}$
= \$20 million - \$18.2million
= **\$1.8million**

Example 23.8

ABC electronics specializes in the production of modern electronics and consequent sale. The managers recently met to discuss the prospects of a new equipment and they came up with the following concerning construction of a new facility.

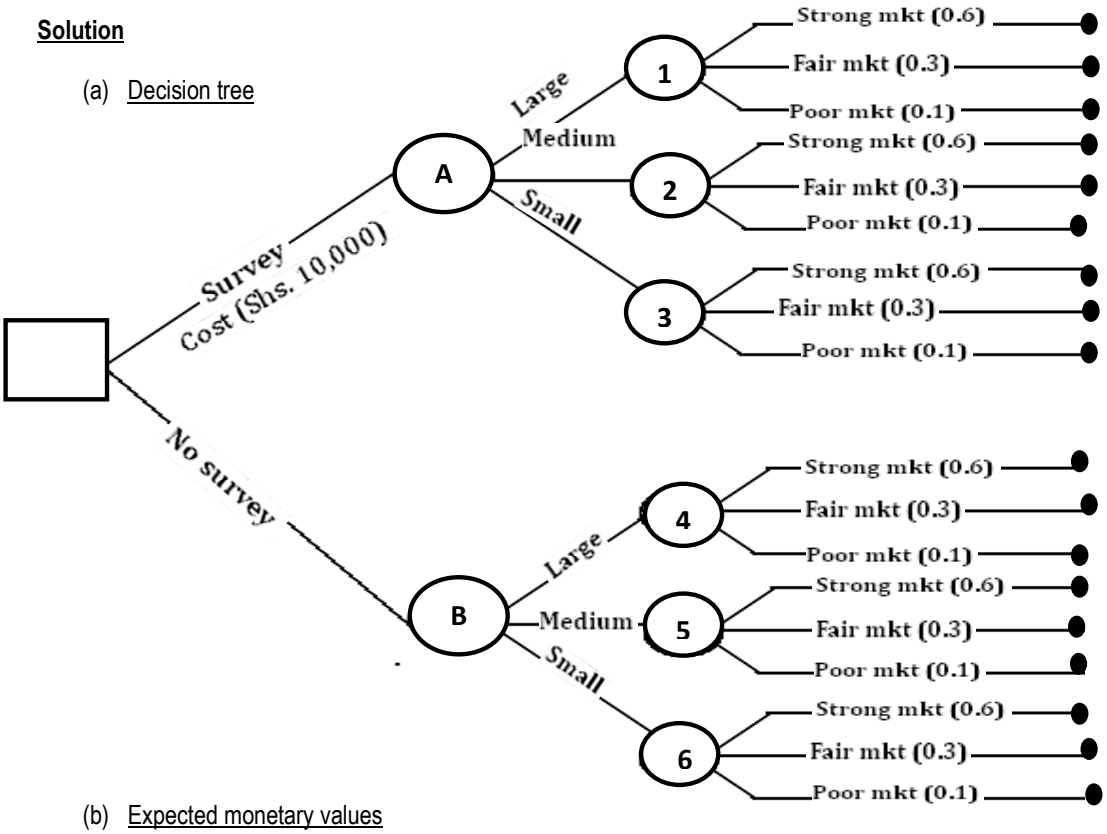
	Strong market Shs	Fair market Shs	Poor market Shs
Large size	550,000	110,000	-310,000
Medium size	300,000	129,000	-100,000
Small size	200,000	100,000	-32,000
Probability	0.6	0.3	0.1

Before taking on the construction, ABC electronics may carry out a market survey which will cost Shs. 100,000.

(a) Develop a decision tree for the table above.
(b) Calculate the expected monetary value at respective nodes.
(c) Advise ABC electronics on which alternative to take.

Solution

(a) Decision tree



(b) Expected monetary values

$\boldsymbol{No.1}$: $0.6(550{,}000 - 10{,}000) + 0.3(110{,}000 - 10{,}000) + 0.1(-310{,}000 - 10{,}000)$

$$= 0.6 \times 500{,}000 + 0.3 \times 100{,}000 - 0.1 \times 320{,}000 = \boldsymbol{Shs.322,000}$$

$$\boldsymbol{No.2}: 0.6(300{,}000 - 10{,}000) + 0.3(129{,}000 - 10{,}000) + 0.1(-100{,}000 - 10{,}000)$$

$$= 0.6 \times 290{,}000 + 0.3 \times 119{,}000 - 0.1 \times 110{,}000 = \boldsymbol{Shs.198,700}$$

$$\boldsymbol{No.3}: 0.6(200{,}000 - 10{,}000) + 0.3(100{,}000 - 10{,}000) + 0.1(-32{,}000 - 10{,}000)$$

$$= 0.6 \times 190{,}000 + 0.3 \times 90{,}000 - 0.1 \times 42{,}000 = \boldsymbol{Shs.136,800}$$

$$\boldsymbol{No.4}: = 0.6 \times 550{,}000 + 0.3 \times 110{,}000 - 0.1 \times 310{,}000 = \boldsymbol{Shs.332,000}$$
$$\boldsymbol{No.5}: = 0.6 \times 300{,}000 + 0.3 \times 129{,}000 - 0.1 \times 100{,}000 = \boldsymbol{Shs.208,700}$$

$$\boldsymbol{No.6}: = 0.6 \times 200{,}000 + 0.3 \times 100{,}000 - 0.1 \times 32{,}000 = \boldsymbol{Shs.146,800}$$

(c) ABC should take the alternative with maximum EMV i.e. the alternative of Node 4 = "Do no survey and construct a large facility.

Advantages of decision trees

(i) It clearly brings out implicit assumptions and calculation for all to see, question and revise.
(ii) It allows one to understand, simplify by inspection, various assumptions and alternatives in a graphical from, which is much easier to understand than the abstract analytical form.

Disadvantages of decision trees

(i) Decision tree diagrams become more and more complicated as the number of decision alternatives increases and as more variables are increased.
(ii) It becomes highly complicated when interdependent alternatives and dependent variables are present in the problem.
(iii) It assumes that the utility of money is linear.

23.7: THE EXPECTED VALUE

The expected value of an event is the product of its probability and the outcome or value of the event over a series of trails.

The probability distribution of a random variable is a list of the values of random variables with their corresponding probabilities of occurrence.
If $\boldsymbol{x}$ is a random variable which can take any one of the values $\boldsymbol{x_1, x_2, \dots, x_n}$ with respective probabilities $\boldsymbol{P_1, P_2, \dots, P_n}$, then the expected value of X, denoted by ***E(X)*** is defined as;

$$E(X) = P_1X_1 + P_2X_2 + \cdots + P_nX_n$$

Where; each $P_i > 0$ and $P_1 + P_2 + \cdots + P_n = 1$

Example 23.9

The revenue curve , R, of Basaza and Sons is known to be $R = 20{,}000x$ while the cost curve , C, is known to be $C = 10{,}000 + bx$ where x represents the number of units produced while b represents the variable costs with the following probabilities;

B	probabilities
10,000	0.1
20,000	0.3
28,000	0.2

Required;

(a) If Basaza expects a production of 1,000 units, compute the expected profit.

(b) Given that the probability of selling 1,000 units is 0.5 and the probability of producing zero units is also 0.5, compute the expected profit.

Solution

(a)

$C = 10{,}000 + bx$

$R = 20{,}000x$

Profit = Revenue – Cost

$Profit = \ 20{,}000x - (10{,}000 + bx)$

$E(Profit) = \ E(20{,}0000x) - E(10{,}000 + bx)$

$E(Profit) = \ E(20{,}000)E(x) - E(10{,}000) - E(bx)$

$E(Profit) = \ 20{,}0000\ E(1{,}000) - 10{,}000 - E(b)E(x)$

$E(Profit) = \ 20{,}0000\ \times 1{,}000 - 10{,}000 - E(b)E(1{,}000)$

But E(b) is given as in the table below;

b	Prob	E(b)
10,000	0.1	1,000
20,000	0.3	6,000
28,000	0.2	5,600
		12,600

Hence;

$E(Profit) = \ 20{,}0000\ \times 1{,}000 - 10{,}000 - E(12{,}600)E(1{,}000)$

$E(Profit) = \ 20{,}000{,}000 - 10{,}000 - (12{,}600)(1{,}000) =$ **Shs. 7,390,000**

$E(Profit) = \mathbf{Shs.\,7{,}390{,}000}$

(b)

Units sold	prob	Expected sales
1,000	0.5	500
0	0.5	0
		$\sum \boldsymbol{E(x) = 500}$

$E(Profit) = E(20{,}0000x) - E(10{,}000 + bx)$

$E(Profit) = E(20{,}000)E(x) - E(10{,}000) - E(bx)$

$E(Profit) = 20{,}0000\,E(x) - E(10{,}000) - E(b)E(x)$

E(b) = 12,600, E(x)= 500

Hence;

$E(Profit) = (20{,}0000 \times 500) - 10{,}000 - (12600 \times 500)$

$= \underline{\textbf{Shs. 3,690,000}}$

Properties of expected values

(i) The expected value of a constant is a constant itself i.e; $\boldsymbol{E(k) = k}$.

(ii) The expected value of a product of a constant and a random variable is the product of the constant and the expected value of the random variable. i.e;

$$\boldsymbol{E(kX) = kE(X)}$$

(iii) The expected value of the sum or difference of two random variables is equal to the sum or difference of the expected values of the individual random variables;

$$\boldsymbol{E(X + Y) = E(X) + E(Y)}$$
$$\boldsymbol{E(X - Y) = E(X) - E(Y)}$$

(iv) The expected value of a product of two independent random variables is equal to the product of their individual expected values. i.e;

$$\boldsymbol{E(XY) = E(X).E(Y)}$$

(v) The expected value of a random variable is also equal to its mean;

$$\boldsymbol{E(X) = \mu = \sum P_i X_i}$$

(vi) The variance of a random variable is given by;

$$Var(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

$$= \sum x^2 p(x) - [\sum xp(x)]^2$$

$$= \boldsymbol{\sum x^2 p(x) - \mu^2}$$

(vii) The standard deviation of a random variable is given by;

$$\delta = \sqrt{variance}$$

Example 23.10

ABC Ltd is in the process of introducing its newly developed shaving machine. Data on sales results has been obtained plus the corresponding probabilities.

Sales	50	100	150	200	250	300
Probabilities	0.10	0.30	0.30	0.15	0.10	0.05

The cost of introduction is Shs. 1,000,000 and unit selling price Shs. 10,000.

Required

(i) Compute the expected sales of ABC Ltd
(ii) Compute the expected profit from the above data.
(iii) Compute the sales variance and sales standard deviation

Solution

Sales (x)	Probabilities	Expected Sales
50	0.10	5
100	0.30	30
150	0.30	45
200	0.15	30
250	0.10	25
300	0.05	15
		$\sum E(x) = 150$

(i) Expected sales = **150 units**.

(ii) Expected profit = Expected sales – cost of introduction
= (150 × 10,000) – 1,000,000
= 1,500,000 – 1,000,000
= **Shs. 500,000**

(iii) Sales variance and Standard deviation

Sales (x)	$Pr(x)$	$xPr(x)$	$x^2P(x)$
50	0.10	5	250
100	0.30	30	3,000
150	0.30	45	6,750
200	0.15	30	6,000
250	0.10	25	6,250
300	0.05	15	4,500
		$\sum E(x) = 150$	$\sum x^2 p(x) = 26{,}750$

- Sales variance = 26,750 – $(150)^2$ = **4,250 units**
- Sales standard deviation = $\sqrt{4{,}250}$ =**65.2 units**

Example 23.11

Patel, an Asian investor is in the process of considering to invest in either of the two projects A and B. The table below shows data that was collected and forwarded to him about each project.

	PROJECT A		PROJECT B	
	Value Shs "000"	Probability	Value Shs "000"	Probability
Optimistic outcome	6,000	0.2	7,000	0.1
Most likely outcome	4,000	0.5	5,000	0.6
Pessimistic outcome	3,000	0.3	2,000	0.2

Required
Compute the expected value for each project and advise Patel on which project to invest in order ensure value for his money.

Solution

We shall begin by computing the expected value for each project as below;

	PROJECT A			PROJECT B		
	Value Shs "000"	Prob	Expected value	Value Shs "000"	Prob	Expected value
Optimistic outcome	6,000	0.2	1200	7,000	0.1	700
Most likely outcome	4,000	0.5	2000	5,000	0.6	300
Pessimistic outcome	3,000	0.3	900	2,000	0.3	600
Project EV			**4,100**			**4,300**

Basing on the computed expected value, Project B would be preferred as it has the highest value.

Arguments in favour of Expected values

(i) Simple to understand and calculate.
(ii) Represents whole distribution by a simple figure.
(iii) It arithmetically takes into account the expectation of all outcomes.

Arguments against the use of Expected values

(a) By representing the whole distribution by a simple figure it ignores other characteristics possessed by the distribution such as skewness and range.
(b) It is based on the assumption that the decision maker is risk averse which may not be the case.

EXERCISES 19

19.1
Metaballs manufacturing company Ltd is considering introducing two new products in the market. The company has the following options;

Option 1: Introduce both products.
Option 2: Introduce either of the products.
Option 3: Introduce none of the products, depending on their performance in the market.

An analysis of the product's likely performance indicates the probability of a good performance as 30%, fair performance as 50% and poor performance as 20% and the sales revenue as shown below in the table;

Decision	States of nature		
	Good performance Shs "millions"	Fair performance Shs "millions"	Poor performance Shs "millions"
Neither	0	0	0
Product 1 only	10.0	5.2	2.4
Product 2 only	8.4	4.8	2.4
Both	7.6	8.8	3.2

Required;

(i) Compute the expected monetary value for each decision.
(ii) What decision would you recommend?
(iii) Formulate the opportunity loss table.
(iv) Compute the expected opportunity loss for each decision.
(v) *Determine the expected value of perfect information.*

19.2 The sales people at Jomayi properties sell up to 9 houses per month. The probability distribution of a sales person selling **x** houses in a month is as follows

Sales (x)	0	1	2	3	4	5	6	7	8	9
Prob f(x)	0.05	0.10	0.15	0.20	0.15	0.10	0.10	0.05	0.05	0.05

Required;
Any sales person selling more houses than the amount equal to the mean plus two standard deviation receives a bonus. Determine the number of houses per month that a sales person should sell to receive a bonus. (Variance = 5.49)

19.3 Homeland Computer company is considering a plant expansion that will enable the company to begin production of a new computer product. The company's Executive Director must determine whether to make the expansion a medium or large scale project. Uncertainty exists in the demand for the new product, which for planning purposes may be low, medium or high demand. The probability estimates for demand are 0.20, 0.50, and 0.30 respectively. The firm's accountants have developed the following annual profit (in thousands of shillings) forecasts for the medium and large scale expansion projects.

		Medium scale expansion		Large scale expansion	
		Profit	Probability	Profit	Probability
Demand	Low	50	0.20	0	0.20
	Medium	150	0.50	100	0.50
	High	200	0.30	300	0.30

Required

(i) *Which decision is preferred for the objective of maximizing the expected profit?*
(ii) *Which decision is preferred for the objective of minimising the risk or uncertainty?*
(iii) *From your answers in (i) and (ii) above, should the company go for medium scale expansion or the large scale expansion? Explain.*

19.4 Jama investments (U) Ltd recently acquired an option to purchase the shares of Uganda Clays Ltd. The company can either sell the option to Summit Ltd who has offered a profit of Shs. 22 million or it can exercise the option and sell the shares in the stock market in three months time. The return on the shares will depend on the state of the stock market at the time. The stock market could be bullish, constant or bearish. Jama Ltd estimates to make a profit of shs. 320 million if the market is bullish, Shs. 80 million if the market is constant and a loss of Shs. 120 million if the market is bearish. Jama Ltd also estimates that the probability of the stock market being bullish is 0.1. However the company is unable to reach a conclusive decision about the probability oF the market remaining constant or being bearish.
Required;

(i) *Draw a decision tree for this problem.*
(ii) *If the probability of the stock market being constant is p, calculate Jama's expected profit in terms of p.*
(iii) *For what values of p would it be advisable for Jama Ltd to sell the option?*

19.5 Farm products Ltd produces and sells fresh milk. The milk has to be sold within a day otherwise it will get spoilt. Each distribution crate carries 20 packets of milk. Each crate costs Shs. 300 to produce and sells for Shs. 450. If demand exceeds supply, a special production of milk is made at a cost of Shs. 400 per crate leaving the selling price constant while unsold milk at the end of the day is given free of charge to the workers. FPL sells milk in multiples of hundred of crates. *The table below shows the distribution of sale of crates of milk over the last 80 days*.

Number of milk Crates sold	100	200	300	400	500
Frequency of sales (days)	8	20	32	12	8

(i) *Compute the pay-off table for the above.*
(ii) *Determine the optimal number of crates of milk that FPL should produce per day using (a) the Maximax criterion (b) Maximin criterion (c) the expected monetary vary criterion.*
(iii) *What is the optimal number of crates of milk FPL should produce in order to minimize risk?*

PART J
LINEAR PROGRAMMING

Chapter 24
LINEAR PROGRAMMING

24.1 Definition

Linear programming is a mathematical technique used in determining the maximum or minimum value of a linear expression. It is mainly concerned with the allocation of scarce resources expressed in mathematical terms.
Linear programming seeks to optimize the value of some objectives such as profit maximisation or minimisation of costs.

In accounting, linear programming is applied in marginal costing and profit –volume analysis.

24.2 Linear programming problems

A linear programming problem is one in which we are to find the maximum or minimum value of a linear expression such as $ax + by + cz + \cdots$;(called the objective function) subject to a number of linear constraints of the form $Ax + By + Cz + \cdots \leq N$ (maximisation inequaity) or
$x + By + Cz + \cdots \geq N$ (minimisation inequality)

24.3 Features of linear programming problems

Linear programming can be used to solve problems with the following features;

(i) The problem must be capable of being stated in mathematical terms.
(ii) All factors involved in the problem must have linear relationships e.g. to double output, we have to double labour input.
(iii) The problem must permit choices between alternative courses of action.
(iv) There must be one or more restrictions on the factors involved, e.g. restrictions on labour hours and materials.

24.4 Solving linear programming problems

Linear programming problems are mathematically expressed by two distinctive equations; i.e.

(i) The objective function
(ii) The equation of constraints / limitations

(i) The objective function

The objective function is an equation which expresses in mathematical terms what the linear program seeks to find.
The objective function for most LP problems seeks to miximise profits or constraints or minimize cost / time and some other appropriate measure.

(ii) The equation of limitations / constraints

The equation of limitations or constraints seeks to express in mathematical terms the minimum or maximum utilization levels for the factors used to achieve the objective function.

Constraints or limitations involve resources, such as labour, materials e.t.c.
Therefore, the equation of limitations seeks to express such factors mathematically.

18.5 Minimisation problems

Minimisation problems are normally concerned with minimizing the cost of fulfilling some objective subject to limitations. The limitations are generally of the GREATER THAN or EQUAL TO type i.e. denoted as ≥, for the logical reason that the best way is to minimize costs.

24.6 Methods of solving linear programming problems

There are two methods of solving linear programming problems, and these include;

(i) The graphical method, and
(ii) The Simplex method

The graphical method is used to solve problems with two unknowns or decision variables while the simplex method is used to solve problems with three or more unknowns. The simplex method can also solve equations with 2 unknowns as well.

24.7 The graphical method

(a) Simple Mathematical Expressions

Here, we shall begin by demonstrating how one can sketch a simple inequality mathematical expression.

Example 24.1

Sketch the linear inequality $\mathbf{3x - 4y \leq 12}$

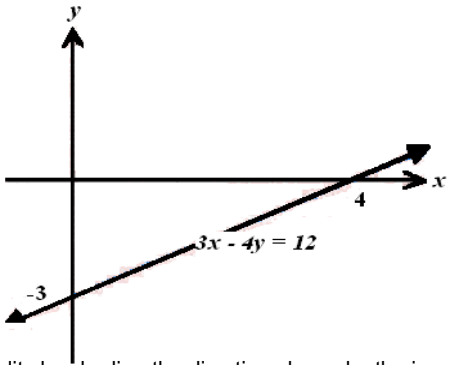
Solution

To sketch the above inequality, we shall need to follow the steps below; i.e;

- Replace the inequality with an equality to obtain the line shown below;

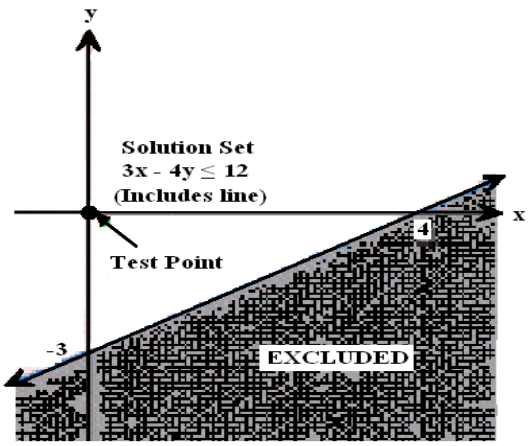
$$\mathbf{3x - 4y = 12}$$

- Sketch the straight line obtained by getting two coordinates when x=0 and y=0. These are the points of intersection on the either axis of the graph.



- Lastly sketch the inequality by shading the direction shown by the inequality. This is done by choosing a test point not on the line. (0,0) is a good choice if the line does not pass through the origin, and if the line does pass through the origin a point on one of the axes would be a good choice.

- If the test point satisfies the inequality, then the set of solutions is the entire region on the same side of the line as the test point. Otherwise it is the region on the other side of the line. In either case, shade out the side that does not contain the solutions, leaving the solution region showing.



In our illustration, we substitute x=0 and y=0 for the test point (0,0) in the inequality and this gives;

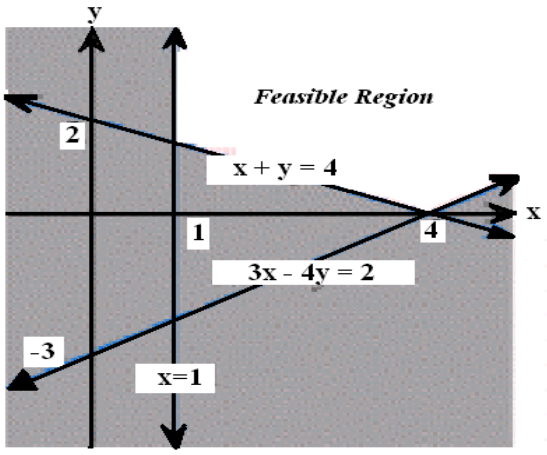
$$\mathbf{3(0) + 4(0) \leq 12;} \qquad \textit{which satisfies the solution}!$$

Since this is a true statement, (0,0) is in the solution set, so the solution set consists of all points on the same side as (0,0). This region is left unshaded (feasible region), while the shaded region is blocked out.

Example 24.2
By shading show the feasible region for the following collection of inequalities;

$$3x - 4y \leq 12$$
$$4x + 2y \geq 4$$
$$x \geq 1$$
$$y \geq 0$$

Solution
We shall still follow the same steps as described in the above example. Note that this time we have more than one inequality. We shall therefore sketch the inequality $4x + 2y \geq 4$ and $x \geq 1$ by shading off the unwanted region on the appropriate side of the lines $4x + 2y = 4$ and $x = 1$ respectively.
The inequality $y \geq 0$ represents the line y = 0 and this is the x-axis shaded with the appropriate feasible region.
We therefore obtain the feasible region shown below (including the boundary).



A test point can be obtained to ascertain if the feasible region indeed satisfies the set of inequalities.

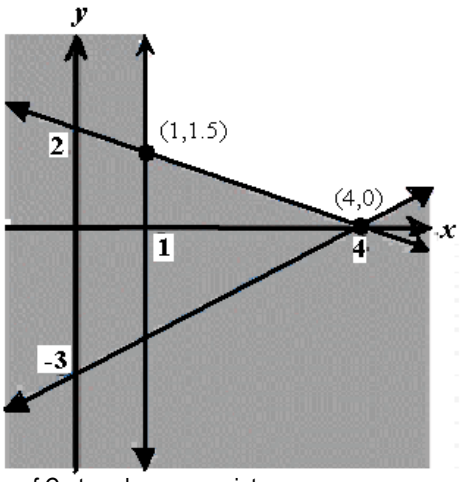
Example 24.3
Given the objective function $C = 3x + 4y$, subject to the constraints

$$3x - 4y \leq 12$$
$$4x + 2y \geq 4$$
$$x \geq 1 \quad and \quad y \geq 0$$

Determine the points that minimize C.

Solution

The feasible region for above set of constraints was shown in example 24.2, here it is again with the corner points shown.



The following table shows the value of C at each corner point:

Point	C = 3x + 4y	
(1, 1.5)	3(1)+4(1.5) = 9	✔ minimum
(4, 0)	3(4)+4(0) = 12	

Therefore, the solution is x = 1, y = 1.5, giving the minimum value C = 9

(b) Linear programming problems involving wording

1. **Formulation of the objective and equations of limitations / constraints.**

Step 1: Formulate the objective function noting how many unknowns appear.

Step 2: Formulate equations of limitations / Constraints ensuring that inequalities of ≥ and ≤ are correctly spelt out.

Step3: Ensure all relationships established are linear or approximately linear.

Step 4: Ensure that non-negativity constraints i.e. x ≥ 0 and y ≥ 0 are stated amongst the equations of limitations / constraints.

2. **Graphical Procedure for plotting the linear relationships established.**

(i) Draw the axes of the graph which represent the unknowns in the objective equation such as X and Y, A and B, P and Q, X_1 and X_2 etc as the question may state.

(ii) Draw equations of constraints that have only one unknown first. These are actually straight lines.

(iii) Equations of constraints with two unknowns are drawn next by determining their respective points of intercept on the either axis of the graph i.e. Points of intersection are obtained at points where x = 0 and y = 0.

(iv) Indicate the feasible regions for each equation of limitations / constraints. Feasible regions are those that don't violet the relationship depicted by the respective equation of limitations / constraints i.e. these are areas that contain all the possible relational combinations.

(v) On defining the feasible region, we locate coordinate points on the vertex of the feasible region which give the maximum relational combination that is substituted in the objective function to give the maximum profit attainable and the percentage resource utilization.

Example 24.4

A company produces 2 products x and y. Product x requires 8 minutes while y requires 10 minutes of labour. Product x consumes 2kgs of raw-materials and y consumes 4kgs of raw-materials. In any week, only 800 hours of labour and 280kgs of raw-materials are available.
In whatever situation, at least 20 units of each product must be produced. Each unit of x generates a profit of 12,000/= and y generates 9,000/= per unit.

Required;

(i) Formulate the objective function for the above company.
(ii) Formulate the constraints to the above function.
(iii) Determine the weekly production that maximizes profits and calculate the profit at this level.
(iv) Determine the percentage utilization of labour when profits are maximized.
(v) Compute the labour spare capacity (if any) available.

Solution

(i) Objective function

Since this is concerned with profit contribution the objective function will definitely be of a profit maximizing nature and expressed in terms of profit contribution per unit hence we have the objective function as;

$$12{,}000x + 9{,}000y$$

(ii) Constraints to the objective function

At this stage, we shall formulate inequalities for labour and raw materials. However, its preferred to summarise wordy questions in form of a table as shown below;

PRODUCTS	Labour (hrs)	Materials (Kgs)
X	8	2
Y	10	4
Available per week	800	280

Inequalities;

$Labour\ constraint$: $8x + 10y \leq 800$
$Raw - materials\ constraint$: $2x + 4y \leq 280$
$x - Production\ constraint$: $x \geq 20$
$y - Production\ constraint$: $y \geq 20$

Plotting;

To plot the above inequalities, we shall need to obtain coordinates that give line intercepts at each respective axis. These will be points where x = 0 and y=0. However, we shall need to replace the inequality signs with the equality signs for each constraint just as we had discussed before.

Therefore;

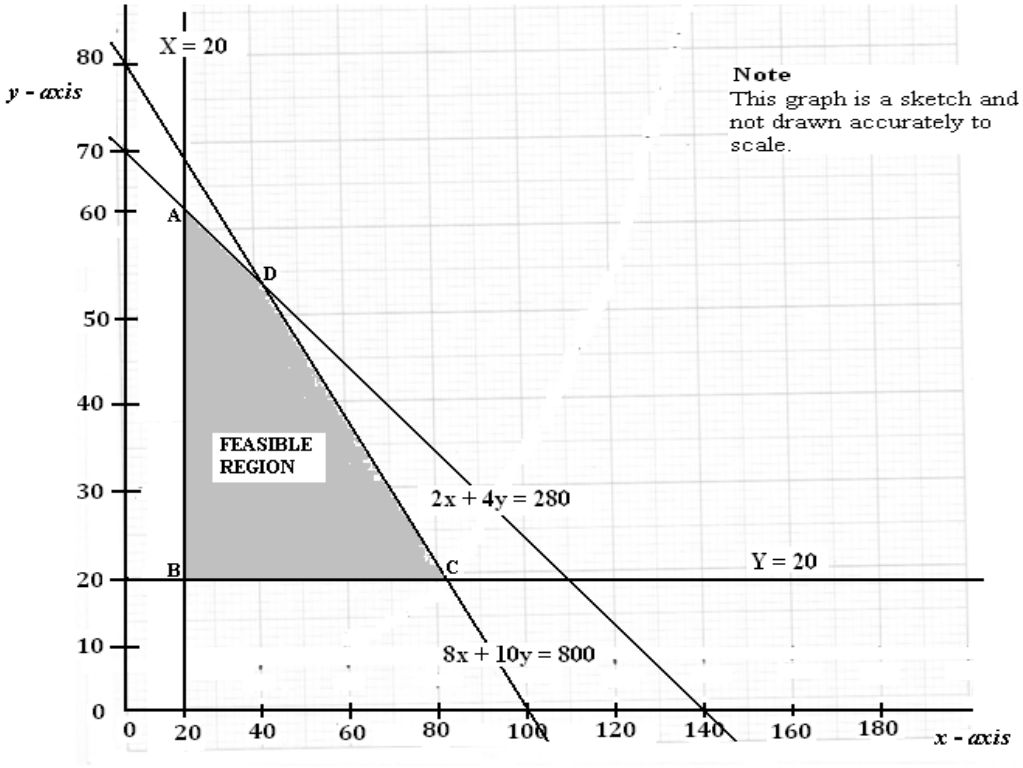
- Labour constraint: $\mathbf{8x + 10y = 800}$

When x = 0, y = 80 and when y = 0, x = 100
hence the coordinates to be plotted are **(0,80)** and **(100,0)**

- Raw Materials constraint: $\mathbf{2x + 4y = 280}$

When x = 0, y = 70 and when y = 0, x = 140
hence the coordinates to be plotted are **(0,70)** and **(140,0)**

- X - production constrain is the line $\boldsymbol{x = 20}$
- Y- production constrain is the line $\boldsymbol{y = 20}$



(iii) *Weekly production that maximises the profit and the profit level*

To obtain the weekly production that maximises profit, we shall read off from the graph the vertex points forming the feasible region and then substitute them in the objective function to establish the pair that gives the greatest profit margin, this will then form the weekly production that maximises profit.

Hence;

Vertex points are A(20,60), B(20,20), C(71,20) and D(34,53)

We then substitute these in the objective function;

Point A: 12,000(20) + 9,000(60) = Shs. 780,000
Point B: 12,000(20) + 9,000(20) = Shs. 420,000
Point C: 12,000(71) + 9,000(20) = Shs. 1,032,000 **** *Profit maximizing pair*
Point D: 12,000(34) + 9,000(53) = Shs. 885,000

Therefore the weekly production that maximises profit is **71 units** of product **X** and **20 units** of product **Y**.

Profit at this level = **Shs. 1,032,000/=**

(iv) *Labour utilization*

$$\%\ Utilisation\ of\ labour = \frac{hours\ utilised}{Available\ per\ week} \times 100$$

Hours utilized will be obtained by substituting x = 71 units and y = 20 units in the objective function; i.e;

8(71)+ 10(20) = 768 hours

$$\%\ Utilisation\ of\ labour = \frac{768}{800} \times 100 = \mathbf{96\%}\ (\boldsymbol{Under\ utilisation})$$

(v) *Labour spare capacity*

Spare capacity = Available per week – hours utilized
= 800 – 768
= **32 labour hours**

Example 24.5

Mukwano industries produces two products P and Q, where P requires 4 hours of labour and Q requires 6 hours of labour. Both products P and Q consume 1kg of materials. Product P requires 4 hours of processing while product Q requires 2 hours of processing. In any given week, only 100 hours of processing, 180 labour hours and 40kgs of materials are available. Product P generates a profit of Shs. 3,000 per unit while Q generates a profit of Shs. 4,000 per unit.

Because of a trade agreement, sales of P are limited to a weekly maximum of 20 units and to honour an agreement with an old established customer at least 10units of Q must be sold per week

Required;

(i) *Formulate the objective function for the above company.*
(ii) *Formulate the constraints to the above function.*
(iii) *Determine the weekly production that maximizes profits and calculate the profit at this level.*
(iv) *Determine the degree of utilization for each constraint at the profit maximisation level*

Solution

(i) *The objective function*

$$3{,}000P + 4{,}000Q$$

(ii) *Constraints to the objective function*

These will be obtained easily by rearranging the given information in the table as shown below;

PRODUCTS	Labour (hrs)	Materials (Kgs)	Processing (hrs)
P	4	1	4
Q	6	1	2
Available per week	180	40	100

Inequalities;

$Labour\ constraint$: $4P + 6Q \leq 180$
$Raw - materials\ constraint$: $P + Q \leq 40$
$Processing\ constraint$: $4P + 2Q \leq 100$
$P - Production\ constraint$: $P \leq 20$
$Q - sales\ constraint$: $Q \leq 10$

Plotting;- Coordinates

- Labour constraint: $\boldsymbol{4P + 6Q = 180}$

When P = 0, Q = 30 and when Q = 0, P = 45
hence the coordinates to be plotted are **(0,30)** and **(45,0)**

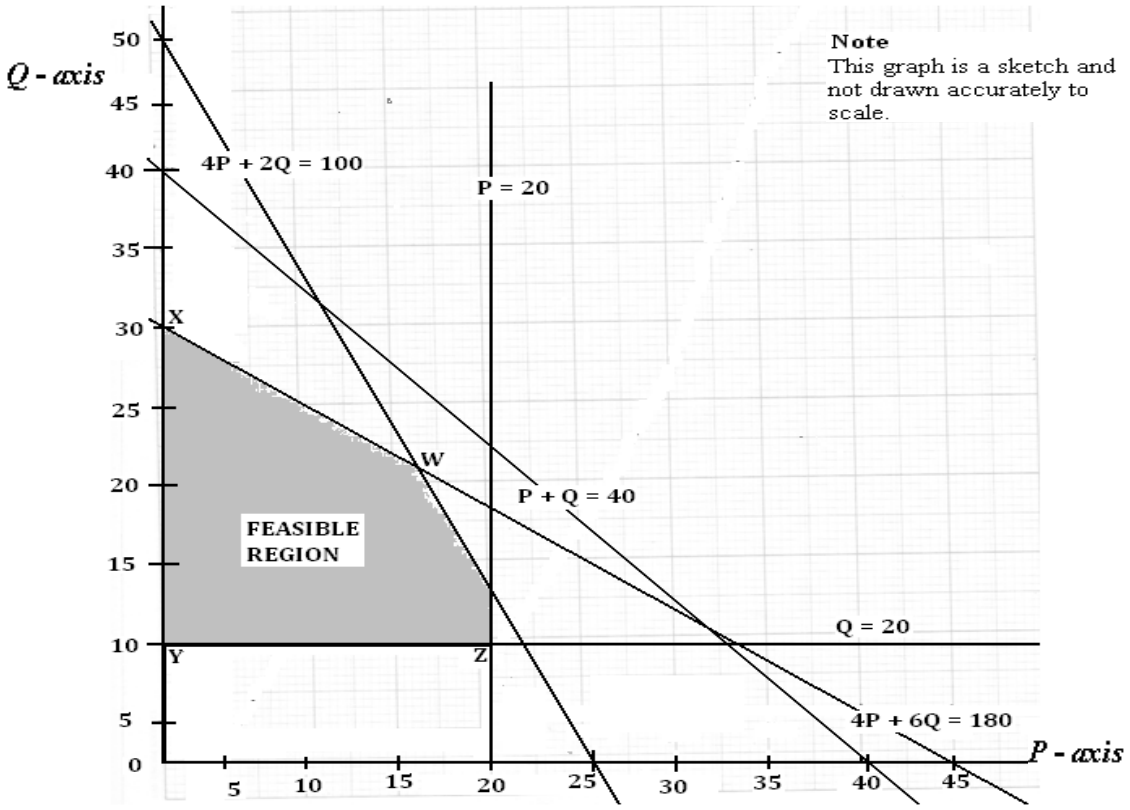
- Materials constraint: $\boldsymbol{P + Q = 40}$

When P = 0, Q = 40 and when Q = 0, P = 40
hence the coordinates to be plotted are **(0,40)** and **(40,0)**

- Processing constraint: $\boldsymbol{4P + 2Q = 100}$

When P = 0, Q = 50 and when Q = 0, P = 25
hence the coordinates to be plotted are **(0,50)** and **(25,0)**

- P - Production constrain is the line $\boldsymbol{P = 20}$
- Q- sales constrain is the line $\boldsymbol{Q = 10}$



(iii) *Weekly production that maximises profit*

Vertex points include; X(0,30), Y(0,10), Z(20,10) and W(15,20)
We substitute these in the objective function to establish the pair that gives the greatest profit margin (profit maximizing basket); i.e;

Point X: 3,000(0) + 4,000(30) = Shs. 120,000
Point Y: 3,000(0) + 4,000(10) = Shs. 40,000
Point Z: 3,000(20) + 4,000(10) = Shs. 100,000
Point W: 3,000(15) + 4,000(20) = Shs. 125,000*** *Profit maximizing pair*

Therefore the production that maximises profit will comprise of 15 units of P and 20 units of Q.

Profit maximised = **Shs. 125,000/=**

(iv) Utilization levels for each constraint

Here, we shall insert the values P = 15 units and Q = 20 units in each of the equation of constraints. i.e;

$Labour\ constraint: \quad 4(15) + 6(20) = 180 \ \ Full\ utilisation, no\ spare$

$Raw - materials\ constraint: \quad 15 + 20 = 35 \ \ Utilisation\ 5kgs\ below\ maximum$

$Processing\ constraint: \quad 4(15) + 2(20) = 100\ Full\ utilisation, no\ spare$

$P - Production\ constraint: \quad P = 15, production\ 5\ units\ below\ maximum$

$Q - sales\ constraint: \quad Q = 10\ \ , sales\ 10\ units\ above\ maximum$

24.8: Shadow and dual prices

Equations without spare capacity are known as "binding equations" and represent constraints that are at full capacity whereby shortage of such resource will halt further production.

Dual prices arise from the amount of increase (or decrease) in contribution (or profit) that would arise if one or more (or less) of a unit of a scare resource was availed.

The shadow price of binding constraint provides valuable guidance because it indicates to management the extra contribution they would gain from increasing by one unit the amount of the scarce resource.

Note:-
Only resources at full capacity utilization are used to compute dual prices.

Example 24.6

Rework example 24.5 and determine;
(i) Shadow price per processing hour
(ii) Shadow price per labour hour
(iii) Amount of overtime pay for a person working an extra 20 hours assuming the current labour cost is Shs. 8,000 per hour

Solution

In example 24.5, labour and processing hours are at full capacity; these therefore form the binding constraints for this particular example.
We shall hence follow to compute the shadow / dual prices as follows;

(i) Shadow price per Processing hour

We assume 1 more processing hour is available (leaving the labour hours constant at 180), then compute the resulting difference in contribution; i.e;

Processing hours: 4P + 2Q = 101 (i.e. Original 100 + 1) ……..(i)
Labour hours: 4P + 6Q = 180 (Unchanged) ……..(ii)

By solving equations (i) and (ii) Simultaneously new values of P and Q are obtained.
Hence **P= 15.375** and **Q = 19.75**

By substituting in the objective function we obtain;

3(15.375) + 4(19.75) = Shs. 125,125
Original contribution = Shs. 125,000
Difference **Shs. 125/=**

Thus 1 extra machine hour has resulted in a extra increase in contribution of Shs. 125/- which is the shadow price per processing hour.

(ii) Shadow price per Labour hour

We assume 1 more labour hour is available (leaving the processing hours constant at 100), then compute the resulting difference in contribution; i.e;

Processing hours: 4P + 2Q = 100
Labour hours: 4P + 6Q = 181

where **P= 14.875** and **Q = 20.25**

By substituting in the objective function we obtain;

3(14.875) + 4(20.625) = Shs. 125,625
Original contribution = Shs. 125,000
Difference **Shs. 625/=**

Therefore shadow price per labour hour = **Shs. 625/-.**

The above implies that management would be prepared to pay up to Shs. 625/- per hour in order to gain more labour hours.

(iii) Overtime pay = (8,000 + 625) × 20 hours

= Shs. 8,625 × 20 hours

= **Shs. 172,500/=**

24.9: THE SIMPLEX METHOD

24.9.1: INTRODUCTION

The simplex method is another linear programming mathematical technique that is used to maximize or minimize a quantity by choosing appropriate values for variables involved.

The simplex method unlike the graphical method can be used to solve both equations having only two variables and those with three or more variables.

The main feature that charaterises the simplex method is the procedure converting inequalities (less-than or equal-to) into equality equations which involves the introduction of "*Slack Variables*" in each of the inequality equation.
The slack variable in this case represents the spare capacity or what we may call the unused units.

24.9.2: STANDARD MAXIMISATION PROBLEM

As noted before, a standard maximization problem is a linear programming problem for which the objective function is to be maximized and all the constraints are "less-than or equal-to" inequalities.

A typical example of a standard maximisation problem is *profit maximisation*.

Main steps followed in the simplex method-Standard maximisation

Step 1:- The first step of the simplex method requires that you express the linear programming into a standardized format.

Step 2:- The second step requires that each inequality (in constraints) be converted into an equation by adding a *slack variable*.

Step 3:- At this stage, formulate an *augmented matrix* of the system of linear equations and use it to set up the *initial tableau*.

Step 4- Select the *pivot column* in the initial tableau. This is the column with the "*highest contribution*" in the objective function row.
Divide the positive numbers in the pivot column with the corresponding elements in the solution quantity column.
Step 5- Select the *pivot row*. This is the row with the *smallest non-negative* quotient resulting from step 4 above.
Identify the *pivot number* which is the number that appears in *both* the pivot column and the pivot row.

Step 6:- Calculate the new values for the pivot row by dividing every number in the row by the pivot number.

Step 7:- Replace the slack variable in the pivot row with the basic variable ***x*** in the pivot column. Use row operations to make all numbers in the pivot column equal to 0 except for the pivot number which remains as **1**.

Example 24.6

Footsteps Furniture Company produces chairs and tables. Each table takes four hours of labour from the carpentry department and two hours of labour from the finishing department.
Each chair requires three hours of carpentry and one hour of finishing. During the current week, 240 hours of carpentry time are available and 100 hours of finishing time. Each table produced gives a profit of Shs. 70 and each chair a profit of Shs. 50.

Required;
Determine the number of chairs and tables to be produced in a week in order to maximize profits

Solution:
We first choose the variables to use;
Let ***x*** represent tables and ***y*** chairs that are produced in the week.

The information is then summarized as follows;

	Tables	Chairs	Constraints	
Number produced per week	x	y	Cannot be negative	$x \geq 0\ ,\ y \geq 0$
carpentry	4 h/table	3 h/chair	Maximum of 240 hours for the week	$4x+3y \leq 240$
finishing	2 h/table	1 h/chair	Maximum of 100 hours for the week	$2x+y \leq 100$

The total profit for the week is given by the objective function as P = 70x + 50y.

Step 1:- Rearranging the system of inequalities;

Carpentry hrs constraint: 4x + 3y ≤ 240
Finishing hrs constraint: 2x + y ≤ 100
Objective function: 70x + 5y = P

Step 2:- Convert inequalities to equations by adding slack variables i.e;

4x + 3y + 1a + 0b = 240
2x + y + 0a + 1b = 100

As unused hours result in no profit, the slack variables can be included in the objective function with zero coefficients.

P = 70x + 5y + 0a + 0b

The problem can now be considered as solving a system of 3 linear equations involving the 5 variables x, y, a, b and P where a and b are slack variables with variable P having the maximum value.

Step 3:- The system of equations can be written in matrix form or a 3 x 5 *augmented* matrix i.e;

$$\begin{pmatrix} 4 & 3 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ 70 & 50 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ a \\ b \end{pmatrix} = \begin{pmatrix} 240 \\ 100 \\ 0 \end{pmatrix} \; or \; \left[\begin{array}{ccccc|c} 4 & 3 & 1 & 0 & 0 & 240 \\ 2 & 1 & 0 & 1 & 0 & 100 \\ 70 & 50 & 0 & 0 & 1 & 0 \end{array}\right]$$

In simplex method, the *augmented* matrix is referred to as the "*tableau*".

The initial tableau is;

Solution Variables	x	y	A	b	Solution Quantity	
a	4	3	1	0	240	240÷4 = 60
b	2	1	0	1	100	100÷2 = 50
	70	50	0	0	0	

Step 4:- Column 1 forms the pivot column at this stage since it contains 70 which is the highest contribution in the objective function row.

Step 5:- The pivot row is row 2 since 50 is the smallest non-negative result when elements in the solution quantity are divided by the corresponding elements in the pivot column.

Step 6:- Calculate the new values for the pivot row by dividing every number in the row with the corresponding elements in the pivot column i.e; $R_2 / 2$.

Solution Variables	x	y	A	b	Solution Quantity
a	4	3	1	0	240
b	1	$\frac{1}{2}$	0	$\frac{1}{2}$	50
	70	50	0	0	0

Step 7:- Replace the slack variable **b** (in the pivot row) with the basic variable ***x*** in the pivot column. Use row operations to make all numbers in the pivot column 0 except for the pivot number which remains as 1. i.e; $(R_1 – 4R_2)$ and $R_3 – 70R_2$.

Solution Variables	x	y	A	b	Solution Quantity
a	0	1	1	-2	40
x	1	$\frac{1}{2}$	0	$\frac{1}{2}$	50
	0	15	0	-35	-3,500

Now repeat the steps until there are no positive contributions in the objective function row.

Additional Steps;

- Set up a new tableau by selecting the next pivot column. Y should replace slack variable **a** in the solution variable column. All other steps are followed. Hence we have the next initial tableau as;

Solution Variables	x	y	A	b	Solution quantity	
a	0	1	1	-2	40	40÷1 = 40
x	1	$\frac{1}{2}$	0	$\frac{1}{2}$	50	50÷½ = 100
	0	15	0	-35	-3,500	

- Calculate new values for the pivot row. As the pivot number is already 1, there is no need to calculate new values for the pivot row.

- Use row operations to make all numbers in the pivot column equal to 0 except for the pivot number. i.e. $R_2 - \frac{1}{2}R_1$ and $R_3 - 15R_1$

Solution Variables	x	y	A	b	Solution quantity
y	0	1	1	-2	40
x	1	0	$-\frac{1}{2}$	$\frac{3}{2}$	30
Row 3	0	0	-15	-5	-4,100

Since there are no positive contributions in the objective function column, the optimum solution has been reached.

Therefore, maximum profit of Shs. 4,100 occurs when 30 tables and 40 chairs are made. There are no unused hours.

The values of Row 3 for the slack variables are of great importance. These are the valuations of resources and are known as *shadow prices*.

***a* = -15** means that for every extra carpentry labour hour available, Shs. 15 extra overall contribution would be gained.
***b* = -5** means that for every extra finishing labour hour availed, the overall contribution would increase by Shs. 5.

24.9.3: INEQUALITIES WITH MIXED CONSTRAINTS

Example 24.7

Consider the problem of finding the maximum value of the objective function
$\boldsymbol{P = 2x + y}$, subject to the constraints;

$$x + y \geq 10$$

$$3x + y \geq 15$$

$$y \leq 10$$

$$x \leq 8$$

$$x \geq 10, y \geq 0$$

This is not a standard maximisation problem because two of the constraints are not "les-than or equal-to" inequalities. Consider the first inequality $\boldsymbol{x + y \geq 10}$, we shall convert this into an equation by introducing a surplus variable s_1 such that $\boldsymbol{x + y - s_1 = 10}$

Using the same approach for the second inequality, the constraints become;

$$1x + 1y - 1s_1 + 0s_2 + 0s_3 + 0s_4 = 10$$
$$3x + 1y + 0s_1 - 1s_2 + 0s_3 + 0s_4 = 15$$
$$0x + 1y + 0s_1 + 0s_2 + 1s_3 + 0s_4 = 12$$
$$1x + 0y + 0s_1 + 0s_2 + 0s_3 + 1s_4 = 8$$

The objective function becomes;

$$P = 2x + 1y + 0s_1 + 0s_2 + 0s_3 + 1s_4$$
$$-2x - 1y + 0s_1 + 0s_2 + 0s_3 + 1s_4 + 1P = 0$$

The initial tableau becomes;

	x	y	s1	s2	s3	s4	p	rhs
s1	1	1	-1	0	0	0	0	10
s2	3	1	0	-1	0	0	0	15
s3	0	1	0	0	1	0	0	12
s4	1	0	0	0	0	1	0	8
	-2	-1	0	0	0	0	1	0

In the above case, Linear programming requires that all the variables should be non-negative. To make this happen, consider only those rows for variables with negative values (1st and 2nd rows) and perform the following;

(a) Select the pivot row (the row with the largest value in the last column).
(b) Select the pivot column (the column with the largest value in the pivot row, ignoring the last column).
(c) Calculate new values for the pivot row (divide every number in the row by the pivot number).
(d) Use row operations to make all numbers in the pivot column equal to 0 except for the pivot number.

Repeat the above steps until all the variables are non-negative.

Therefore; the pivot row in the above example is row 2 since 15>10 and the pivot column is column 1 (3 is the largest value in the pivot row, ignoring the last column)

Below is the final tableau for the above example which gives the maximum value of *P* as *P* = 28, occurring when x = 8 and y = 12.

	x	y	s1	s2	s3	s4	p	rhs
y	0	1	0	0	1	0	0	12
x	1	0	0	0	0	1	0	8
s1	0	0	1	0	1	1	0	10
s2	0	0	0	1	1	3	0	21
	0	0	0	0	1	2	1	28

Students are encouraged to verify the above solution for purposes of obtaining practice in this area!

24.9.4: SIMPLEX METHOD-MINIMISATION PROBLEMS

A typical example of a Simplex method-minimisation problem is *cost-minimisation*.

Example 24.8

Minimise the objective function $40x_1 + 50x_2$ subject to the constraints;

$$3x_1 + 5x_2 \geq 150$$

$$5x_1 + 5x_2 \geq 200$$

$$3x_1 + x_2 \geq 60$$

and the non-negative constraints

Solution

- In simplex method-minimisation problems, we begin by changing the minimisation problem into a maximisation problem. This is done by formulating the inverse or dual of the above inequalities. i.e;

  Minimize $150a + 200b + 60c$
  Subject to: $3a + 5b + 3c \leq 150$
  $5a + 5b + c \leq 150$
  Where a, b and c are inequality inverse variable / constants.

The above inverse relationship is possible since for every maximizing problem there is an equal and opposite minimising problem. The reverse is true.

- The initial simplex tableau will therefore be as;

Solution Variables	a	b	c	s_1	s_2	Solution quantity
s_1	3	5	3	1	0	40
s_2	5	5	1	0	1	50
Row 3	150	200	60	0	0	0

- Proceeding with a step by step reduction of all positive contributions in Row 3 into negative contributions (as described before) gives the final tableau as;

Solution Variables	a	b	c	s_1	s_2	Solution quantity
b	0	1	$\frac{6}{5}$	$-\frac{3}{10}$	$-\frac{3}{10}$	5
a	1	0	-1	$-\frac{1}{2}$	$\frac{1}{2}$	5
Row 3	0	0	-30	-25	-15	-1,750

- The tableau above is the normal result of a simplex maximizing routine. However, to obtain the solutions to the original minimising problem some differences in the usual interpretations are necessary as follows;

  (i) The figures under the slack variable columns represent the quantities that minimize the objective function (e.g. cost minimisation basket); therefore to minimize the objective function, we need **25** units of **a** (representing x_1) and **15** units of b (representing x_2)

  (ii) The figure under the solution quantity column (i.e 1,750) represents the total value that can be minimized e.g. in the case of minimising costs, it would represent the total cost that can be minimized.

  (iii) Row titled ***b*** and ***a*** indicate the shadow prices e.g. an extra unit of ***b*** or ***a*** will change by 5.

  (iv) -30 under column ***c*** indicates an overproduction of 30 units.

EXERCISE 20

20.1 A firm produces two products, *X* and *Y* with a profit contribution of Shs. **8** and Shs. **10** per unit respectively. Production data are shown in the table below (per unit);

	Labour hours	Material A	Material B
X	3	4	6
Y	5	2	8
Total available	500	350	800

Required;

(i) Formulate the objective function for the above company.
(ii) Formulate the constraints to the above function.
(iii) Determine the weekly production that maximizes profits and calculate the profit at this level.

20.2 A portable computer corporation manufactures two types of portable computers, type X and type Y, on which it earns a profit of Shs. 300 per unit, and type Y, on which it earns a profit of Shs. 400 per unit. In order to produce each unit of computer X, the company uses 1 unit of input A, ½ unit of input B, and 1 unit of input C. To produce each unit of computer Y, the company uses 1 units of input A, 1 unit of input B and no input C. The firm can use only 12 units of input A and only 10 units of inputs B and C per time period.

Required;
(a) Determine how many computers of type X and Y the firm should produce in order to maximize its profits.
(b) How many of each input should the firm use in producing the product mix that maximizes total profits?

20.3 Use the following information to formulate a linear programming problem.

	Units for one unit of product		Total units available
	x	y	
Labour hours	1	6	120
Machine hours	1	3	66
Material units	3	2	86
overheads	10	3	250

Given that a unit of product X carries a profit of $50 and each unit of Y earns $60.

(a) Write down the maximization problem with all the above information.

(b) Determine the optimal solution using the graphical technique.

20.4 A bicycle factory is to be built. There are two types of assembly lines which can be used for assembling the bicycles. An assembly line of type A can produce 5 bicycles per day. It occupies 500m^2 of floor space ad needs 9 skilled workmen to operate. An assembly line of type B can produce 3 bicycles per day. It occupies 600m^2 of floor space but needs only 4 skilled men to operate. In the factory there is 6000m^2 of floor space available for assembly lines. Only 72 skilled workmen are available.

Required

(i) Write the maximisation problem.

(ii) Determine all the constraints

(iii) Use the graphical method to determine how many of each type of lines the factory manager should install.

(iv) Find the maximum possible number of bicycles to be produced in a day.

(v) *Find the maximum value of output for the day if each bicycle costs Shs. 130,000.*

20.5 A linear programming problem with three variables x_1, x_2 and x_3 has the objective function $P = 100x_1 + 300x_2 + 200x_3$ and the following constraints;

$$x_1 + x_2 + x_3 \leq 100$$
$$40x_1 + 20x_2 + 30x_3 \leq 3{,}200$$
$$x_1 + 2x_2 + x_3 \leq 160$$

and the non-negative constraints.

Required;

Solve the linear programming problem using the simplex tableau method and interpret the results.

20.6 A company can produce three products A, B and C. The products yield a contribution of Shs. 8, Shs. 5 and Shs. 10 respectively. The products use a machine which has 400 hours capacity in the next period. Each unit of the products uses 2, 3 and 1 hour respectively of the machine's capacity. There are only 150 units available in the period of a special component which is used singly in production of A and C. 200Kgs only of a special alloy is available in the period where Product A uses 2Kgs per unit and Product C uses 4Kgs per unit. There is an agreement with a trade union to produce not more than 50 units of product B in the period.

The company wishes to find out the production plan which maximizes contribution.

20.7 A firm produces three products X, Y and Z with a contribution of Shs. 20, Shs. 18 and Shs. 16 respectively. Production dates are as follows;

Products	Machine Hours	Labour Hours	Materials (Kgs)
X	5	2	8
Y	3	5	10
Z	6	3	3
Availability	3,000 hours	2,500 hours	10,000 Kgs

Required

Determine the production of X, Y and Z that will maximize profits and the profit maximized under such a production contribution.

PART K

STATISTICAL QUALITY CONTROL

Chapter 25
CONTROL CHARTS

25.1 INTRODUCTION

Chapter 25 has been designed to help in understanding and learning the use, design and analysis of control charts, which is the most important tool of statistical quality control.

Control Charting is one of the tools of statistical Quality control (SQC). It is the most technically sophisticated tool of SQC. It was developed in the 1920s by Dr. Walter A. Shewhart of the Bell Telephone labs.

Dr. Shewhart developed the control charts as a statistical approach to the study of manufacturing process variation for the purpose of improving the economic effectiveness of the process. These methods are based on continuous monitoring process variation.

25.2 The statistical Quality Control Process (SQCP)

The SQCP involves taking samples from the parent production whereby they are tested, weighed or measured and statistical tests are applied to see whether the process is performing satisfactorily or not. Its another form of hypothesis testing.

Normally all production methods and processes are subject to variability

Illustration

A Machine filling 300mls of soda will be set to work an average of 300mls but inevitably some bottles will contain slightly more than 300mls of soda, others slightly less. Such variability may arise due to two main causes namely;

(i) *Inherent causes*
(ii) *Special causes*

These are explained as below;

(i) Inherent Causes

These are unavoidable causes due to the nature of the process, the machines used, the material etc. Such causes produce a distribution of values (sizes, weights, volumes etc) which is approximately normal and predictable.
If all the variability of the samples were due to inherent causes, the process would be said to be "*in control*."

(ii) Special Causes

These are variations caused by specific changes in the process or environment. Such changes may arise as a result of wrong material being used, incorrect adjustment on production machinery, failure of a component among others.
These causes produce instability in the distribution which is no longer predictable and the process is said to be "*out of control*".
Presence of special causes in the production process normally requires immediate investigation.

25.3 The Importance of Control charts in the SQCP

A Control Chart is a device used to indicate whether a process is "within control" (i.e. Variability due to inherent causes) or "out of Control" (i.e. variability due to specific causes). Besides being a proven technique for improving productivity some of its other importance include;

(i) Effective prevention of defect in production processes.
(ii) Prevention of unnecessary process adjustments
(iii) Provision of diagnostic information
(iv) Provision of information about process capability.

25.4 The Structure/ Build up of a Control Chart

Control Charts normally contain control limits usually computed from statistical data available and the sample values are checked against the control limits.

They normally have two levels of control limits i.e. a warning limit and the action limit, whereby if a sample plotting is outside the control limits action would be taken to trace the specific cause of such a variation.

Example 25.1

At Mukwano industries an oil filling tap is known to fill 10,000 Jerricans with
20 litres of cooking oil per hour, with standard deviation 0.24 litres. Random samples of 9 jerricans are taken and each jerrican measured. The results for 6 consecutive samples are given below;

Sample No.	1	2	3	4	5	6
Sample Mean(in Litres)	20.06	19.96	20.03	20.11	20.16	20.08

Required
(a) Compute the upper and lower warning and action control lines at;
(i) 95% confidence level as warning limit
(ii) 99% confidence level as action limit
(b) Plot the control lines and samples results on a mean control chart for the 95% confidence limit.
(c) Interpret your results.

Solution

(a)
(i) The confidence interval at 95% is given by;

$$\boldsymbol{\mu \pm \frac{1.96\sigma}{\sqrt{n}}}$$

$$\text{Hence the warning limits at 95\% level } = 20 \pm \frac{1.96 \times 0.24}{\sqrt{9}}$$

$$= 20 \pm 0.157$$

The Upper warning limit = 20 + 0.157 = ***20.157 litres*** and the lower warning limit =
20 - 0.157 = ***19.843 litres.*** We conclude therefore that if the process is under control, 95% of the time, the capacity of each jerrican produced must be between 19.843 and 20.157 litres. Anything outside these limits serves as a warning that something may well be wrong.

(ii) The confidence interval at 99% is given by;

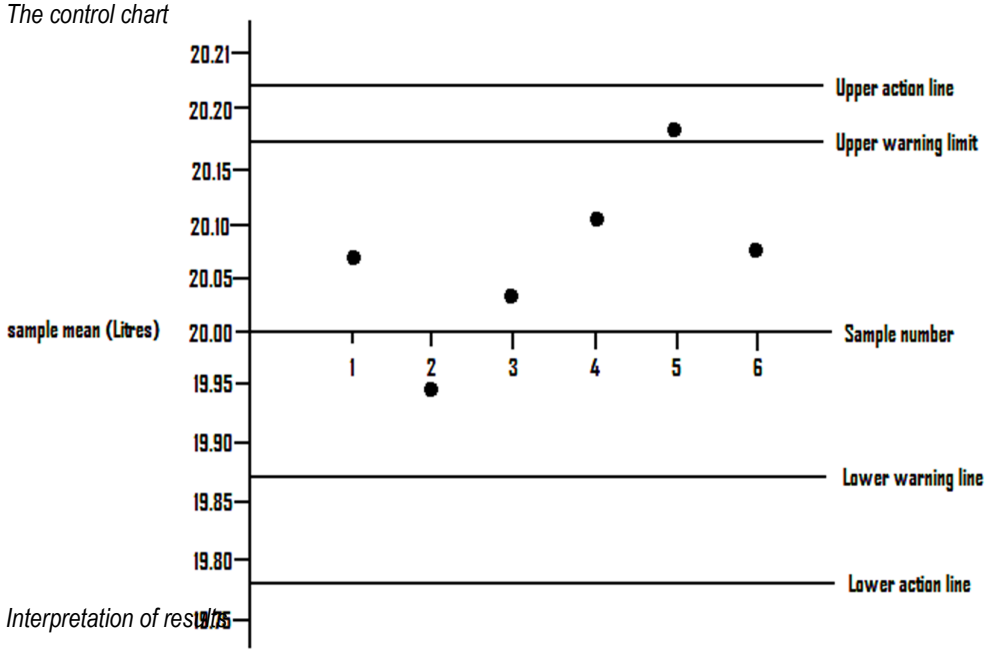
$$\boldsymbol{\mu \pm \frac{2.58\sigma}{\sqrt{n}}}$$

$$\text{Action limits at 95\% level } = 20 \pm \frac{2.58 \times 0.24}{\sqrt{9}}$$

$$= 20 \pm 0.206$$

The Upper action limit = 20 + 0.206 = ***20.206 litres*** and the lower action limit = 20 - 0.206 = ***19.794 litres.*** We conclude therefore that any production outside these limits serves as a warning that certainly something is wrong and action needs to be taken.

(b) The control chart



(c) Interpretation of results

The mean for sample 5 is just above the upper warning line, but the mean of the next sample is well within the control lines, so we assume that there is no problem. It is a normal practice to take a further sample immediately a sample is found to be on or beyond the control lines.

NOTE:

(i) In this example we have used the 95% level as the warning limit and the 98% as the action limit; however other limits such as 98% (as action limit) and 90% (as warning limit) may be chosen.

(ii) There are other types of control charts such as *attribute control charts* besides the sample control charts demonstrated above. Students are encouraged to learn all these. However the principles remain as illustrated in example 25.1.

ANSWERS TO EXERCISES

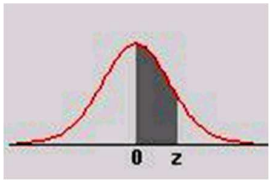
	EXERCISE 3		**EXERCISE 4**
3.1	Mode 6, Median 6, Mean 2.7	**4.3**	Mean 12.13, Standard deviation1.87
3.2	Mode 3, Median 3, Mean 2.7	**4.4**	Mean 3.24, Standard deviation 0.244
3.3	Median 148 ½ lb	**4.5**	Mean 42.74, Standard Deviation 4.04
3.4	Median 48 in.		
3.5	Arithmetic Mean of the salaries $7,000		
3.6	Average grade = 85		
3.7	Mean Salary = $2.75		
3.8	Arithmetic mean = 8.25		
3.9	Mean 39.69, Median 39.90, Mode 41		
3.10	Mean 79.08, Median 79.5, Mode 79.7		

	EXERCISE 6		**EXERCISE 7**
6.1	(i) $\frac{11}{12}$ (ii) $\frac{3}{4}$	**7.3**	$\mu = \frac{1}{2},\ \sigma^2 = \frac{1}{12}$
6.3	$\frac{1}{9}$		
6.4	0.24		**(a)** (x) p(x) **(b)** p(y)
6.5	(i) $\frac{1}{3}$ (ii) $\frac{7}{15}$ (iii) $\frac{8}{15}$		0 $\frac{1}{16}$ 0 $\frac{2}{16}$
5.6	$\frac{11}{24}$		1 $\frac{4}{16}$ 1 $\frac{6}{16}$
6.7	(i) 0.23 (ii) 0.21		2 $\frac{6}{16}$ 2 $\frac{2}{16}$
6.8	(i) $\frac{8}{13}$ (ii) $\frac{4}{13}$ (iii) $\frac{12}{13}$		3 $\frac{4}{16}$ 3 $\frac{6}{16}$
6.9	$\frac{38}{63}$		4 $\frac{1}{16}$
6.11	$\frac{21}{64}$		
6.12	(i) $\frac{7}{9}$ (ii) $\frac{2}{9}$ (iii) $\frac{3}{5}$ (iv) $\frac{2}{45}$	**7.5**	**x** **p(x)**
6.13	(i) $\frac{1}{15}$ (ii) $\frac{17}{40}$		0 $\frac{6}{36}$
6.16	(i) $\frac{2}{3}$ (ii) $\frac{11}{15}$		1 $\frac{10}{36}$
6.19	0.216		2 $\frac{8}{36}$
6.20	(i) 0.46 (ii) 0.16 (iii) 0.13 (iv) 0.83		3 $\frac{6}{36}$
			4 $\frac{4}{36}$
			5 $\frac{2}{36}$

	EXERCISE 8		**EXERCISE 9**
8.1		**9.1**	(i) $\frac{1}{2}$ (ii) $\frac{3}{4}$ (iii) $\frac{\pi}{2}$
8.6	1	**9.2**	(i) $\frac{4}{5}$ (ii) 1.2 (iii) 1.22 (iv) $\frac{7}{12}$
8.7	2		
8.8	3.041		
8.9	$1.08		
8.10	$420		
	EXERCISE 10		**EXERCISE 11**
10.1	(a) 0.234 (b) n = 400 , p =1/10	**11.1**	(a) 0.6147 (b) 0.4822 (c) 0.7008 (d) 0.9973
		11.3	(a) 0.6247 (b) 0.9772 (c) 0.0228
		11.4	(a) 0.0913 (b) 0.2523 (c) 0.6564
		11.7	(i) 0.2029 (ii) 6.68% (iii) 114Kg
		11.9	(i) 0.2751 (ii) 0.0228 (iii) 0.7249
	EXERCISE 12		**EXERCISE 13**
12.1	0.9987	**13.1**	v = 2 degrees of freedom, $x^2 = 67.11$, H_0 is rejected.
12.2	0.0154		
12.4	60 ± 8.58g		
12.5	(i) 7.613 to 8.387 (ii) 7.706 to 8.294	**13.2**	v = 1 degrees of freedom, $x^2 = 2.9$, critical value = 3.84, H_0 is accepted.
12.6	Sample size n = 1937		
12.8	(a) 50 ± 5.877g (b) 50 ± 8.444g		

TABLES

Area Between 0 and Z

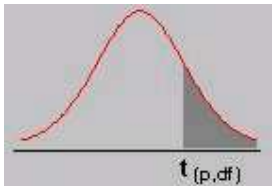


CUMULATIVE NORMAL DISTRIBUTION P(z)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.5000

The Standard Normal distribution is used in various hypothesis tests including tests on single means, the difference between two means, and tests on proportions. The Standard Normal distribution has a ***mean*** *of* ***0*** *and a* ***standard deviation*** *of* ***1****.*

As shown in the illustration above, the values inside the given table represent the areas under the standard normal curve for values between 0 and the relative z-score. For example, to determine the area under the curve between 0 and 2.36, look in the intersecting cell for the row labeled 2.30 and the column labeled 0.06. The area under the curve is .4909. To determine the area between 0 and a negative value, look in the intersecting cell of the row and column which sums to the absolute value of the number in question. For example, the area under the curve between -1.3 and 0 is equal to the area under the curve between 1.3 and 0, so look at the cell on the 1.3 row and the 0.00 column (the area is 0.4032).



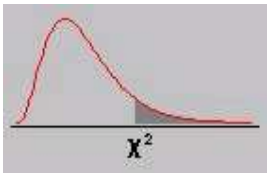
t - DISTRIBUTION TABLE WITH RIGHT TAIL PROBABILITIES

Degrees of freedom	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460

*The Shape of the Student's **t**-distribution is determined by the degrees of freedom. Its shape changes as the degrees of freedom increases.*

*As indicated by the chart above, the areas given at the top of this table are the right tail areas for the t-value inside the table. To determine the **0.05** critical value from the t-distribution with 6 degrees of freedom, look in the **0.05** column at the 6 row: $\mathbf{t_{(.05,6)} = 1.943180.}$*

Right tail areas for the *Chi-square* Distribution



PERCENTAGE POINTS OF THE CHI-SQUARE (χ^2) DISTRIBUTION χ^2_Q

V	.995	.990	.975	.950	.100	.050	.025	.010	.005	.001
1	0.00004	0.00016	0.00098	0.00393	2.70554	3.84146	5.02389	6.63490	7.87944	10.8300
2	0.01003	0.02010	0.05064	0.10259	4.60517	5.99146	7.37776	9.21034	10.59663	13.8200
3	0.07172	0.11483	0.21580	0.35185	6.25139	7.81473	9.34840	11.34487	12.83816	16.2700
4	0.20699	0.29711	0.48442	0.71072	7.77944	9.48773	11.14329	13.27670	14.86026	18.4700
5	0.41174	0.55430	0.83121	1.14548	9.23636	11.07050	12.83250	15.08627	16.74960	20.5200
6	0.67573	0.87209	1.23734	1.63538	10.64464	12.59159	14.44938	16.81189	18.54758	22.4600
7	0.98926	1.23904	1.68987	2.16735	12.01704	14.06714	16.01276	18.47531	20.27774	24.3200
8	1.34441	1.64650	2.17973	2.73264	13.36157	15.50731	17.53455	20.09024	21.95495	26.1200
9	1.73493	2.08790	2.70039	3.32511	14.68366	16.91898	19.02277	21.66599	23.58935	27.8800
10	2.15586	2.55821	3.24697	3.94030	15.98718	18.30704	20.48318	23.20925	25.18818	29.5900
11	2.60322	3.05348	3.81575	4.57481	17.27501	19.67514	21.92005	24.72497	26.75685	31.2600
12	3.07382	3.57057	4.40379	5.22603	18.54935	21.02607	23.33666	26.21697	28.29952	32.9100
13	3.56503	4.10692	5.00875	5.89186	19.81193	22.36203	24.73560	27.68825	29.81947	34.5300
14	4.07467	4.66043	5.62873	6.57063	21.06414	23.68479	26.11895	29.14124	31.31935	36.1200
15	4.60092	5.22935	6.26214	7.26094	22.30713	24.99579	27.48839	30.57791	32.80132	37.7000
16	5.14221	5.81221	6.90766	7.96165	23.54183	26.29623	28.84535	31.99993	34.26719	39.2500
17	5.69722	6.40776	7.56419	8.67176	24.76904	27.58711	30.19101	33.40866	35.71847	40.7900
18	6.26480	7.01491	8.23075	9.39046	25.98942	28.86930	31.52638	34.80531	37.15645	42.3100
19	6.84397	7.63273	8.90652	10.11701	27.20357	30.14353	32.85233	36.19087	38.58226	43.8200
20	7.43384	8.26040	9.59078	10.85081	28.41198	31.41043	34.16961	37.56623	39.99685	45.3100
21	8.03365	8.89720	10.28290	11.59131	29.61509	32.67057	35.47888	38.93217	41.40106	46.8000
22	8.64272	9.54249	10.98232	12.33801	30.81328	33.92444	36.78071	40.28936	42.79565	48.2700
23	9.26042	10.19572	11.68855	13.09051	32.00690	35.17246	38.07563	41.63840	44.18128	49.7300
24	9.88623	10.85636	12.40115	13.84843	33.19624	36.41503	39.36408	42.97982	45.55851	51.1800
25	10.51965	11.52398	13.11972	14.61141	34.38159	37.65248	40.64647	44.31410	46.92789	52.6200
26	11.16024	12.19815	13.84390	15.37916	35.56317	38.88514	41.92317	45.64168	48.28988	54.0500
27	11.80759	12.87850	14.57338	16.15140	36.74122	40.11327	43.19451	46.96294	49.64492	55.4800
28	12.46134	13.56471	15.30786	16.92788	37.91592	41.33714	44.46079	48.27824	50.99338	56.8900
29	13.12115	14.25645	16.04707	17.70837	39.08747	42.55697	45.72229	49.58788	52.33562	58.3000
30	13.78672	14.95346	16.79077	18.49266	40.25602	43.77297	46.97924	50.89218	53.67196	59.7000
40	20.71000	22.16000	24.40000	26.51000	51.81000	55.76000	59.34000	63.69000	66.77000	73.4000
50	27.71000	29.71000	32.3600 0	34.76000	63.17000	67.50000	71.42000	76.15000	79.49000	86.6600
60	35.53000	37.48000	40.48000	43.19000	74.40000	79.08000	83.30000	88.38000	91.95000	99.6100
70	43.28000	45.44000	48.76000	51.74000	85.53000	90.53000	95.02000	100.4000	104.2000	112.300
80	51.17000	53.54000	57.15000	60.39000	96.58000	101.9000	106.6000	112.3000	116.3000	124.800
90	59.20000	61.75000	65.65000	69.13000	107.6000	113.1000	118.1000	124.1000	128.3000	137.200
100	67.33000	70.06000	74.22000	77.93000	118.5000	124.3000	129.6000	135.8000	140.2000	149.400

Like the Student's t-Distribution, the Chi-square distribution's shape is determined by its degrees of freedom. As shown in the illustration above, the values inside this table are critical values of the Chi-square distribution with the corresponding degrees of freedom. To determine the value from a Chi-square distribution (with a specific degree of freedom) which has a given area above it, go to the given area column and the desired degree of freedom row.