

BASIC CONCEPTS OF STATISTICS

SESSION 1-1: NATURE OF STATISTICS

1-1.1 Introduction

Statistics is that body of knowledge (or field of study) used in making sense of data. It is basically concerned with the collection and analysis of data in order to obtain better understanding of phenomena for effective decision-making. Statistical methods required to achieve the objectives of a statistical investigation are classified into two, namely, *Descriptive Statistics* and *Inferential Statistics*. Descriptive Statistics deals with methods for summarizing and presenting data in tabular and graphical forms as well as using numerical measures such as percentages, mean, standard deviation, etc. in an informative way. Inferential Statistics is concerned with procedures used to make generalization about the characteristics of a population using the information contained in a sample, randomly selected from the population under investigation. Another important tool, in the study of Statistics is *Probability Theory*. Probability Theory provides good analysis of any situation (in Science, Business or in everyday life), which in some way involves an element of uncertainty or chance.

It is the study of measure of uncertainty and the risk associated with it. It provides the basis for the methods involved in inferential analysis of data.

The use of Statistics has permeated almost every facet of our lives. Everyone, both in professional careers and in everyday life through contact with newspapers, television and other media, is presented with information in the form of data which we often need to draw some conclusion from; hence some understanding of Statistics would be helpful to anyone. Since scientists, engineers, and business professionals routinely engage in obtaining and analysing data, knowledge of Statistics is especially important in these fields. Specifically, knowledge of Statistics and Probability Theory can be a powerful tool to help professionals in Science, Engineering, Industry and Business in designing new products and systems, improving existing designs and designing, developing and improving product processes. Most universities all over the world require at least an introductory study of Statistics in all their academic programmes for students to be able acquire knowledge in statistical methods to evaluate published

numerical facts and to believe or reject them as well as employ such scientific methods to help interpret the results of surveys or take decisions and understand them more effectively.

The amount of statistical information that is collected, processed and disseminated to the public for one reason or the other has increased almost beyond comprehension and what part of is “good” statistics and part is “bad” statistics is anybody’s guess. To act as watchdogs more and more persons with some knowledge of Statistics are needed to take an active part in the collection, analysis and, more importantly, in the preliminary planning of data. Persons employed in this area of study need to know the basic concepts, strength and limitations of Statistics. Statisticians are indeed of great assistance in scientific research works. They, specifically,

- Design surveys and experiments to minimize the cost of obtaining a specified quantity of information,
- Seek the best method for analysing the data and making an inference for a given sampling situation, and,
- Analyse the data collected and provide a measure of goodness of an inference.

1-1.2 Applications of Statistics

Data, as earlier noted, are numerical facts and figures from which conclusions can be drawn. Such conclusions are important to the decision-making processes of many professions and organizations. Governments, businesses and individuals collect required statistical data to carry out their activities efficiently and effectively. The rate at which statistical data are being collected is staggering and is primarily due to the realisation that better decisions are possible with more information and, perhaps more importantly, to technological advances that have enabled the efficient collection and analysis of large bodies of data. The most important technological advance in this area has, of course, been the development of the electronic digital computer. Statistical concepts and methods, and the use of computers in statistical analyses, have affected virtually all disciplines in Science, Engineering, Business and others. In Business and Economics, the development and application of statistical methods have led to greater production efficiency, to better forecasting techniques, and to better management

practices. To appreciate the extensive applications of Statistics to wide range of problems, a whole lot of examples can be cited:

- *Product Quality Design/Improvement:* The importance of Statistics in Science, Engineering and Management has been underscored by the involvement of industries in quality improvement. Many companies worldwide have realized that poor product quality in the form of manufacturing defects and /or unsatisfactory product reliability and field performance dramatically affects their overall productivity, market share and competitive position and ultimately the profitability. Improving these aspects of quality can eliminate waste, reduce scrap and rework, the requirements for inspection and test, and warranty losses, enhance customer satisfaction and enable the company to become the high-quality, low cost producer in its market. Businesses and other organizations often employ statistical analysis of data to help in improving their processes. Production supervisors use manufacturing data to evaluate, control and improve product quality. Businesses decide which products to develop and market by using data that reveal consumer preferences. In particular, statistical methods help to demonstrate the need for improvements, identify ways to make improvements, assess whether or not improvement activities have been successful, and estimate the benefits of improvement strategies.
- *Insurance Premiums:* Insurance companies use statistical analyses to set rates for home, automobile, life and health insurance. Tables are available, determining the probabilities of survival of persons of years ahead of them. On the basis of these probabilities, life insurance premiums can be established.
- *Water Quality:* The Environmental Protection Agency (EPA) is always interested in the water quality of rivers/lakes. They periodically take water samples to establish the level of contamination and maintain the level of quality.
- *Potency of Drugs:* Medical Researchers study the cure rates for diseases based on the use of different drugs and different forms of treatment. For example, what is the effect of treating a certain type of knee injury surgically or with physical therapy? If you take an aspirin each day, does that reduce your risk of a heart attack? Physicians and hospitals use data on the effectiveness of drugs and surgical procedures to provide patients with the best possible treatment.

- *Opinion Polls:* Politicians and their supporters rely immensely on data from public opinion polls to formulate legislation and devise campaign strategies about their prospects of winning an election. The percentage of a candidate winning an election from a random sample of about 1,000 registered voters prior to the election may be used to estimate the percentage of votes they are likely to receive in the election.
- *Unemployment/Inflation:* Government officials use conclusions drawn from the latest data on unemployment and inflation from a survey to make policy decisions.
- *Investment Decisions:* Financial planners use recent trends in stock market prices to make investment decisions.

The use of computers now play important role providing statistical summaries of data arising from the above situations. They are used for variety of purposes, such as word-processing, record-keeping, accounting, etc. There are various statistical packages notably among these are *Microsoft Excel*, *MINITAB*, *GENSTAT*, *Statistical analysis System (SAS)*, *Statistical Package for Social Sciences (SPSS)*, *Biomedical Statistics Packages (BMDP)*, *R* and *Strata*.

SESSION 2-1: COLLECTION OF DATA

2-1.1 Definitions

- *Population and Sample:* A *population* is a collection of all possible individual units (persons, objects, experimental outcomes, etc.), whose characteristics are to be studied. A *sample* is a part of a population that is studied to learn more about the entire population. That is, to infer about a population, we usually take a sample from the population.
- *Parameters and Statistics:* Numerical values computed from a given set of data are quantitative measures. A quantitative measure that describes a characteristic of a population is called a *parameter*. Such a measure is always computed from the population data. A quantitative measure that describes a characteristic of the sample is called a *statistic*. A statistic is computed from a sample data and used to estimate a parameter or make an inference about a certain characteristic of the population under study.

- *Types of Data:* Data are classified as either *quantitative* or *qualitative*. *quantitative data* assume numerical values, which are as a result of measurements. They indicate “how much or many” of something. Quantitative data are always numeric and are obtained from either an *interval* or a *ratio scale* of measurement. For example, age, height, number of brother/sisters, number of accidents occurring, etc.

Qualitative data (also known as *categorical data* or *attributes*) are data whose values fall into one or another of a set of mutually exclusive and exhaustive classes or categories. For example, *sex* (male or female), *marital Status* (single, married, widower, divorced, etc.) or *performance* (excellent, very good, good, fair, fail). They are obtained from either a *nominal* or an *ordinal* scale of measurement.

- *Variables:* Statistical data or information gathered are obtained by conducting interview, inspecting items and in many other ways. The characteristic that is being studied is called a *variable*. That is, a variable is a characteristic observed on sample or population units and that can vary from unit to unit. For example, heights of people, ages of persons, weights of newly born babies, grades obtained in an examination etc. vary from person to person. There are two kinds of variables: *quantitative* and *qualitative variable*. Quantitative variables are further classified as either *discrete* or *continuous*. Discrete variables assume only discrete or integral values while continuous variables assume any value within a specific range.
- *Univariate Data:* These are data obtained on a single characteristic of the individuals under study. For example, ages of students in a school.

Multivariate Data: They are data obtained on more than one variable of the individual units under study. A special case of it is bivariate *data*, which involve two variables of the individuals under study. For example, when data collected are on two attributes of individuals under study, they are summarized by cross-tabulation in a table called *contingency table*. A contingency table summarizes categorical data on, say, *sex* and *performance*, *sex* and *smoking habits*, *ethnicity* and *party affiliation*, etc.

2-1.2 Scales of Measurement

The scale of measurement of data determines the amount of information contained in the data. It indicates the data summarization and statistical analysis that are most appropriate. There are four scales of measurement, namely *Nominal*, *Ordinal*, *Interval* and *Ratio*.

2-1.2.1 Nominal Scale:

It simply uses labels or codes to identify an attribute of an element or element or individual. Data measured on this scale, called nominal data, may be numeric or non-numeric. Arithmetic operations for nominal data are inappropriate. For example we have:

- Sex: male, female
- Marital Status: single, married, divorced, etc.
- Employment Status: employed, unemployed
- Firms: (services, industries, manufacturing)

2-1.2.2 Ordinal Scale:

This is a scale of measurement for a variable that has the properties of nominal scale and can be used to rank or order the data. Ordinal data may be non-numeric or numeric. Arithmetic operations do not make sense for ordinal data, even if data are numeric such operations and averaging are inappropriate. For example we may have:

- Academic performance: excellent, very good, good, fair, poor
- Award winners (winner (1), first runner-up (2), second runner-up (3), etc.)
- Condition of patience: much better, better, bad
- Socio-economic status: low, middle, high

2-1.2.3 Interval Scale:

It has the properties of ordinal scale with interval or difference between data values indicating how much more or less of a variable one element possesses when compared to another element. Interval data are always numeric and have arithmetic operations and averaging is meaningful. For example, the difference between two temperatures indicates that one is warmer than the other.

2-1.2.4 Ratio Scale:

It has the properties of interval scale. The ratio of two data values is meaningful. Ratio data are always numeric and arithmetic operations are possible. For example, variables such as distance, height, weight and time, use ratio scale of measurement. A requirement of this scale is that a zero value is inherently defined on the scale.

The amount of information in data as seen from above, varies with the scale of measurement. The nominal data contain the least amount of information while the ratio data contain the highest.

2-1.3 Sources of Data

The data needed for a statistical investigation are either readily available or must be collected. Data that are already available are known as *secondary data* and that must be collected are known as *primary data*. Primary data are original data that has been collected directly from source for the purpose required or is response to a problem that has arisen. For example, data collected during population census. They are very useful for statistical analysis because the exact information required is obtained directly. However, their collection may be too expensive, time-consuming and cumbersome. Secondary data are already compiled data for statistical analysis. They are not collected especially for the investigation at hand but have been collected for some other purpose(s). Secondary data are cheaper and easier to obtain. They are, however, more generalized in nature and less reliable as it is removed at least one stage from its original source.

Statistical investigations can use either primary, secondary data or combination of the two. Suppose that a national company is planning to introduce a new range of products.

It might refer to use secondary data on rail and road transport, areas of relevant skilled labour and information on production and distribution of similar goods from data compiled by the Ghana Statistical Services to site their new factory. The company might also have carried out a survey to produce their own primary data on prospective customer attitudes and the availability of distribution through wholesalers.

2-1.4 Data Collection Methods:

Data are the basic raw material needed for any statistical work. Their collection, processing and analysis are therefore very important for decision-making since wrongly collected data would lead to wrong decision-taking. Statistical data may be obtained through the methods below. Each method has its advantages and disadvantages. The statistician decides on one which is best for that particular investigation. Sometimes it may be necessary to try different methods to see which one actually works best in practice. The basic methods are presented as follows:

2-1.4.1 Personal Interview

In most situations, the best method of eliciting information from individuals is by a personal interview. The interviewer personally contacts individuals selected to participate in the survey or experiment. The responses are then recorded on the schedule (the questionnaire form to be completed). This method produces a higher response rate and further allows the interviewer to clear up any misunderstandings about any of the questions on the schedule. However, personal interviews are very expensive. Interviewers must be carefully selected and trained, and sufficient remuneration must be provided to ensure that the interviewer is competent and dedicated to the chore. To ascertain the responses already gathered and the interviewers' demeanor, it is always prudent in this method to call some of the respondents to ensure that they were actually contacted to ascertain the accuracy of the responses.

2-1.4.2 Self-Administered Questionnaire:

This is probably the most common method of acquiring data from people in a survey or an experiment. The questionnaire is usually distributed to the selected individuals by mail or delivered personally. The use of this type of method suffers from two main serious drawbacks. First, the respondents usually have difficulty in interpreting the questions since no one is available for assistance. If this situation arises, the information received may contain a high degree of non-sampling error or the respondents become frustrated and not bother completing or returning the questionnaire. Second, the response (or return) rate of questionnaire is extremely low. The principal advantage of the method is the relatively low cost of obtaining information.

2-1.4.3 Telephone Interview:

Occasionally, it is possible to conduct an interview over the telephone with the interviewer working from a schedule as in a personal interview. Polls to determine the most popular programme on television or radio are frequently conducted in this manner. Telephone interviews are usually less expensive than personal interviews, but the responses rate is lower and fewer questions are often asked since respondents soon get fed up and abandon the proceedings. This method is restricted to the urban centres where telephone facilities are often located (although not everyone owns one).

2-1.4.4 Observation and Experimentation:

Data for a statistical investigation can also be obtained by direct observation or performing the necessary experiment. This can be used for examining items sampled from production line, in traffic surveys or in work study. It is normally considered to be the most accurate form of data collection, but is very labour-intensive and cannot be used in many situations.

2-1.4.5 Extraction from Administrative Records:

This method is solely used to collect secondary data from published sources such as administrative files, libraries, print/electronic media, internet, etc. For example, a study on births/deaths in Ghana, data can easily be obtained from Births and Deaths Department, Ghana Statistical Services and Ministry of Health.

2-1.5 Design of Questionnaire

The design of a questionnaire in a statistical study requires careful consideration. A badly designed questionnaire can cause many administrative problems and may cause incorrect deductions to be made from the statistical analysis of the results.

There are three basic steps involved in designing a questionnaire or schedule, namely, *Designing the Instrument, Pretesting and Editing Results.*

- (a) *Designing the Instrument:* The questionnaire must be short as much as possible. The items (questions) should be simple and unambiguous, not involve tests of memory, not personal and offensive or leading. The questions must also be asked in a logical order.
- (b) *Pretesting in Pilot Survey:* This step is very essential in constructing a questionnaire or schedule instrument. The instrument is usually given to a small

number of respondents in a pilot survey to determine its adequacy. The information gathered during the pretesting exercise may be used to estimate statistics required for the proper planning of the statistical design of the study.

- (c) *Editing Results:* The completed questionnaire forms or schedules are carefully checked and edited to eliminate or reduce errors, if not completely. Nowadays, computers are used extensively to edit data. Various computer assisted techniques have been developed to identify *outliers* – responses which are greatly different from the majority of responses. Many outliers result from recording, transcription, or clerical errors, or from false information provided by the respondent.

SESSION 3-1: STAGES OF STATISTICAL INVESTIGATIONS

If the investigation is to optimize the use of the available resources, expertise and time it is essential to carefully examine all aspects of the design and application of statistical investigations (experiments and surveys) at the planning level. The main stages involved in the planning and execution of a sample survey may be grouped somewhat arbitrarily under the following headings.

3-1.1 Statement of Problem and Objectives

An investigation cannot be launched in general terms. We must identify the cause for concern and state explicitly what the problem is. The objective is then translated into a set of:

- Definitions of the characteristic for which data are to be collected, and
- Specifications for collecting, processing and publishing.

Hence define in clear and precise terms the objective of the investigation. ‘Clear and precise’ are intended to mean that the statement is not ambiguous and is concrete in defining what is to be achieved.

Stating the objective carefully gives those conducting the investigation terms of reference from which they can start to collect relevant data for analysis.

3-1.2 Target Population and the Use of Sample or Entire Population

Define in clear and unambiguous terms the population of interest. The decision to use a sample or the entire population is based of the following:

- Definition of Sampling Units: The population must be capable of division into sampling units for purposes of sample selection. The sampling units must be current, cover the entire population and be defined in such a way that they will be distinct, recognizable without ambiguity and non-overlapping in the sense that every element of the population belongs to one and only one sampling units. Also, it must also be located at the time of sampling for a mobile population.
- Selection of Appropriate Sampling Design: The key factors in selecting a sampling design are variability, cost and time involved. The choice of a design usually requires the involvement of an expert.

3-1.3 Design of Questionnaire or Schedule

The construction of questionnaire or schedule of enquiry is an extremely difficult task since the respondent or the data collector must interpret them. It requires skill, special techniques as well as familiarity with the subject-matter under study. Where possible use a set of questions, which have been designed by an expert and have been tested.

3-1.4 Method of Data Collection and Organization of Fieldwork

Whether data should be collected by personal interview, telephone, mail questionnaire method, by physical observation or by abstraction from available sources has to be decided keeping in view the costs involved and the accuracy aimed at.

It is absolutely essential that the personnel should be thoroughly trained in locating the sample units, recording the measurements, in the method of collection of required data before starting the fieldwork. The success of an investigation to a great extent depends upon the reliable fieldwork. It is necessary to make provisions for adequate supervisory staff for inspection after fieldwork.

3-1.5 Required Data

The data to be collected should be guided by the objective of the investigation. It is essential not to collect too many data some of which are never subsequently examined and analyzed. A practical method is to make an outline of the tables that the investigation should produce. This would help in eliminating the collection of irrelevant information and ensure that no essential data are omitted.

3-1.6 List of Available Resources

A wide variety of resources is likely to be required for the operation of the investigation and the analysis of the results. These includes the following,

(a) Physical Resources:

- Sampling frame (lists of the sampling units, maps, identifying positions of sampling units, etc)
- Provision of field Manuals and records:
- Computer facilities: Data collected in a survey or experiments are generally stored on a computer. Consequently there must be a computer with a computer programme to input, summarize and analyze the data. There must be sufficient time on the computer and sufficient money to pay the computer time (if a payment is required)

(b) Human resources:

- Expertise in survey (experimental) design,
- Data collectors and data processors,
- Expertise in the processing, analyzing and interpretation of results.

(c) Financial Resources: Money would be needed for:

- Planning, implementation and analysis,
- Payment of computer time (if a payment is required).

3-1.7 Conducting Pilot Survey

A pilot survey (also known as pre-testing) is survey carried out before the main. It is conducted to test the techniques of the survey and not collect viable data. It is used to:

- test the questionnaire for its clarity and its length,
- estimate cost of the main survey,
- estimate the time needed for responding to the survey,
- detect the sources of error,
- identify problems which may be encountered in the main survey:

A pilot survey is necessary if:

- the survey to be conducted is large,
- the results of the survey is important, or
- enough resources are available.

Once the researcher have digested the results of the pilot survey, changes are made, and, if time and budget permit, a second pilot survey can be undertaken on a fresh sampling of subject to further improve the final document.

3-1.8 Collection, Editing, Storage and Organization of Data

- (i) Collection, Security and Editing of Data: This stage is the most time-consuming and costly component of the whole statistical process.
- (ii) Data Storage, Organization and Analysis: Storage of data: Commonly it is necessary to store the information collected on a computer, making sure first that the computer has the capacity to meet the requirement, or alternatively, has facilities that will enable that data to be transferred to a more powerful or larger computer.

3-1.9 Interpretation and Presentation of Results

This is the last stage of the survey where a report on the whole study is prepared and presented in a very simple style. In this report

- The technical aspect of the design is reported;
- The terms of reference is quoted;
- Tables, charts and diagrams are presented to show the findings;
- The results of the study should be interpreted in simple language and accurately and concisely be presented,
- Some recommendations are made to resolve the problem studied. The future direction of investigations should be indicated.

3-1.10 Cost Effectiveness

A survey is usually conducted to solve a problem. The subject matter may be on quality of a product, congestion at the market centre, provision of inadequate service at reasonable cost or the improvement of the environment. Data should not be collected for their mere sake or demonstrate one's skill of writing a report. Conducting surveys are very expensive. The cost for all stages of the survey must be identified and detailed budget prepared to determine the cost-effectiveness of the study.

DESCRIPTIVE STATISTICS

The observations (data) obtained from a survey or experiment may represent a sample selected from a population or the entire population, as in a national census. These observations are usually too many to gain an insight into the nature of the acquired information and generally making it impossible for one to convey much information about the characteristics of the population under study. It therefore becomes necessary to organize and reduce the data into meaningful forms. This unit gives the various descriptive methods for summarizing survey or experimental data. These are mainly categorized as tabular, graphical and numerical representation of data.

Learning Objectives

After studying the unit, students will be able to know the various methods of descriptive analysis of data and how to apply them. Specifically, they should be able to:

- Construct the various frequency distributions and some graphical representations of data such as pie/bar chart, histogram, cumulative frequency curve, etc.
- Distinguish between the measures of central tendency and dispersion and compute these measures.
- State the properties of numerical measures and interpret their computed values.
- Describe the tabular, graphical, and numerical representation of data in a descriptive analysis of data.

SESSION 1-2: TABULAR AND GRAPHICAL REPRESENTATION OF DATA

1-2.1 Tabular Representation of Data

The data gathered from a survey/experiment are usually summarized or organized numerically in tabular form using a *frequency distribution table* and its related forms. The frequency distribution table indicates the occurrence of the observations or values in the data obtained. The distribution is said to be *ungrouped* if it shows the distinct observations and their corresponding occurrences, called *frequencies*. If the number of

observations is too large then they are put into groups, called *classes* or *categories*. The number of classes is usually chosen between 5 and 20, inclusive.

The general rule is to use small number of classes for small amount of data and large number of classes for large amount of data. The best choice of number of classes (k) is suggested by the following:

- The number of classes is the smallest integer value, k such that $2^k \geq n$, or
- By Sturges (Approximation) Rule, the number of classes,

$$k = 1 + 3.322 \log_{10} n, \text{ and class width, } C = \frac{\text{Range}(R)}{k},$$

where n is the total number of observations.

Another useful technique for summarizing data is *relative frequency* or *cumulative frequency distribution table*. The relative frequency distribution indicates the proportion of occurrence of the observations while cumulative frequency distribution shows the total number of occurrences above or below certain key observations or classes. The frequency distribution table is obtained by first putting the observations in ordered array (that is, arranging the observations in order of magnitude). Depending on nature of study being conducted other tables can also be adopted to summarize the measurements made.

Example 2.1:

The data given below are the number of children per family sampled from a community some time ago. The data given below are the number of children per family sampled from a community some time ago.

0	1	4	4	3	1	2	3	1	2
2	4	3	0	2	5	0	2	2	1
3	2	1	1	3	2	3	4	5	2
1	0	5	4	2	0	3	5	1	2
4	3	0	2	5	1	1	2	2	4

The frequencies, relative and cumulative frequencies for the above data are shown in the distribution below.

No. of children(x)	Tally	No. of families(f)	Relative frequency	Cumulative frequency
0	/// /	6	0.12	6
1	/// ///	10	0.20	16
2	/// - /// ///	15	0.30	31
3	//////	8	0.16	39
4	/// //	7	0.14	46
5	///	4	0.08	50
<i>Total</i>	-	$n = 50$	1.00	

It is observed from the distribution that a greater number (15) or proportion (30%) of the families have two (2) children and fewer (4) families have 5 children.

Example 2.2 :

Suppose we conduct a sample survey to find the shoe sizes of students in a department of Faculty of Science of KNUST and obtain the following responses:

7	6	6	7	6	8	9	10	10	11
7	7	8	7	8	9	10	10	7	8
8	7	6	9	5	6	5	8	7	8
7	6	7	8	8	5	10	9	8	9
8	8	7	8	8	7	6	5	9	8
12	11	5	6	10	8	8	9	9	11

The various sizes of shoes are: 5, 6, 7, 6, 8, 9, 10, 11 and 12 giving the distribution as shown below:

Shoe Size	Tally	Frequency	Relative frequency.	Cumulative frequency
5	///	5	0.08	5
6	/// ///	8	0.13	13
7	/// /// //	12	0.20	25
8	/// - /// /// //	17	0.28	42
9	/// ///	8	0.13	50
10	/// /	6	0.10	56
11	///	3	0.05	59
12	/	1	0.02	60
<i>Total</i>	-	$n = 50$	1.00	-

The proportions of students in the various shoe size categories are determined by computing the relative frequencies. For example, the relative frequency of shoe size, 8 is 0.28. That is, 28% of the students in the department wear that size of shoe.

Example 2.3:

The data below show the weights (in ounces) of malignant tumours removed from the abdomens of 65 patients:

68	63	42	27	30	57	28	32	48	27
23	24	25	44	51	36	12	45	25	28
28	42	36	51	74	25	43	65	12	32
38	42	27	31	50	38	21	16	24	59
23	22	43	27	49	38	23	19	49	30
49	12	22	31	49	47	43	80	63	35
55	41	54	11	38					

The given data is grouped into a number of classes in a frequency distribution as follows:

- (a) The number classes, k (since it is not given) using the Sturges' (Approximation) Rule,

$$\begin{aligned} k &= 1 + 3.322 \log_{10} n \\ &= 1 + 3.322 \log_{10} 65 = 7.0225 \approx 7 \end{aligned}$$

Rounded up to the nearest or desired whole number.

- (b) The range (R) and class width (C) are computed using the Sturges' (Approximation) Rule, as

$$R = \text{maximum} - \text{minimum observation (weight)}$$

$$= 80 - 11 = 69$$

$$C = \frac{\text{Range}}{k} = \frac{69}{7} = 9.857 \approx 10$$

Rounded up to 10 to include all the observations.

- (c) The class boundaries: We determine the first $(LB_1 - UB_1)$ as follows:

$$\begin{aligned} LB_1 &= (\text{minimum observation or lesser}) - \frac{1}{2}(\text{smallest unit of the measurements}) \\ &= 11 - \frac{1}{2}(1) = 10.5 \quad \text{or} \quad \left\{ 10 - \frac{1}{2}(1) = 9.5 \right\} \end{aligned}$$

$$UB_i = LB_i + C$$

$$= 10.5 + 10 = 20.5 \text{ or } (9.5 + 10 = 19.5)$$

The subsequent class boundaries are obtained by adding C to the class limits or boundaries as shown in the distribution below.

- (e) The grouped frequency distribution is obtained by finding the number of times the observations fall within each class by tallying.

<i>Weight of Tumour</i>	<i>Tally</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>
10.5 – 20.5	/// /	6	6
20.5 – 30.5	/// / / / / /	20	26
30.5 – 40.5	/// / / / /	11	37
40.5 – 50.5	/// / / / / /	16	53
50.5 – 60.5	/// /	6	59
60.5 – 70.5	/// /	4	63
70.5 – 80.5	//	2	65
<i>Total</i>	-	65	-

Using a minimum observation of 10 for the class boundaries would require a larger class width of about 10.1 to obtain the required number of class boundaries: 9.5– 19.6, 19.6 – 29.7, 29.7 – 39.8, 39.8 – 49.9, 49.9 – 60.0, 60.0 – 71.1, 71.1 – 81.2.

Example 2.4:

The data below are the average sulphur dioxide (SO_2) emission rates (in lb/million btn) from utility and industrial boilers from 50 states.

2.3	2.7	1.5	1.7	0.3	0.6	4.2	0.9	1.2	0.4
0.5	2.2	4.5	3.8	1.2	0.2	1.0	0.7	0.3	1.4
0.7	3.6	1.0	0.7	1.7	0.5	0.2	0.6	2.5	2.7
1.5	1.4	2.9	1.0	3.4	2.1	0.9	1.9	1.0	1.7
1.8	0.6	1.7	2.9	1.8	1.4	3.7	5.0	3.8	2.1

- (a) Summarize the data by constructing a grouped relative frequency distribution.
- (b) Find the approximate proportion of states with the following sulphur dioxide emission rates.
- (c)
 - (i) between 0.9 and 2.2 lb/million btn
 - (ii) at least 3.6 lb/million btn

Solution:

- (a) By Sturges' Rule, the required number of classes and class width are:

$$k = 1 + 3.322 \log_{10} 50$$

$$= 6.64 \approx 7, \text{ and}$$

$$C = \frac{5.0 - 0.2}{7} = 0.69 \approx 0.7$$

The limits of first class boundary:

$$LB_1 = 0.2 - \frac{1}{2}(0.1) = 0.15$$

$$UB_1 = LB_1 + C = 0.15 + 0.7 = 0.85$$

The other class boundaries are shown in the following required grouped relative frequency distribution.

Emission rate (lb/million btn)	Tally	Frequency	Relative frequency
0.15 – 0.85	/// /// / //	13	0.26
0.85 – 1.55	/// /// / //	13	0.26
1.55 – 2.25	/// / //	10	0.20
2.25 – 2.95	/// /	6	0.12
2.95 – 3.65	//	2	0.04
3.65 – 4.35	///	4	0.08
4.35 – 5.05	//	2	0.04
<i>Total</i>	-	50	1.00

- (b) The approximate proportion of states whose emission rate is

- (i) between 0.9 and 2.2 lb/million btn

$$= \frac{13 + 10}{50} = \frac{23}{50} = 0.46$$

- (ii) at least 3.6 lb/million Btn

$$= \frac{4 + 2}{50} = \frac{6}{50} = 0.12$$

1-2.2 Graphical Representation of Data

The data represented on frequency distribution and its related forms are further summarized using graphs or charts for stronger visual impact. These diagrams are very useful in interpreting data when quick analysis of data is needed. The diagrams are categorized for quantitative and qualitative data.

1-2.2.1 Graphical Representation of Quantitative Data:

Quantitative data are represented graphically using *Histogram/Frequency Polygon*, (the most widely used form of data presentation), *Cumulative Frequency Curve* or techniques of *Exploratory Data Analysis (EDA)*.

1-2.2.1.1 Histogram and Frequency Polygon:

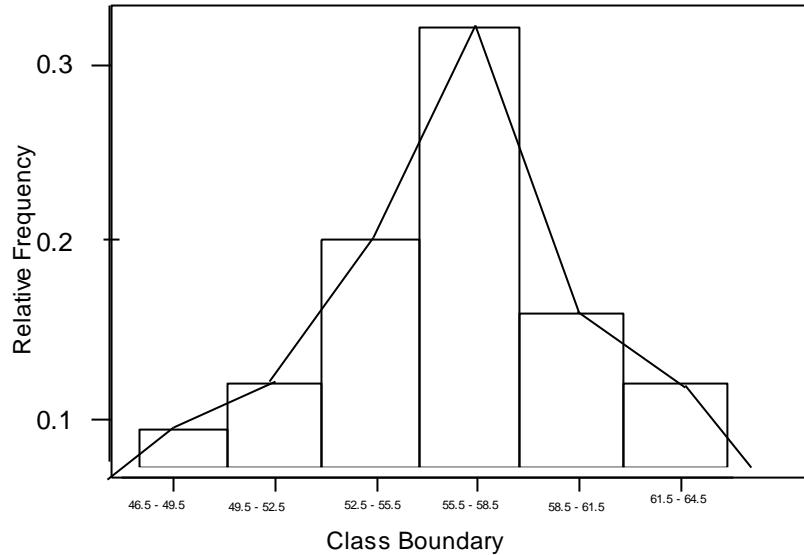
The histogram is the most widely used form of data presentation. It is a graph of frequency or relative frequency distribution where the (relative) frequencies are represented vertically by rectangular bars with no gaps in between them. The horizontal axis takes the observed data using the class boundaries while the vertical axis is labelled as (relative) frequency. For unequal class intervals, we plot the class boundaries against frequency densities, where the *frequency density is defined as frequency divided by class width*.

Another diagram closely associated with histogram is the *frequency polygon*. It is drawn by joining the mid-points of tops of rectangular bars in a histogram. As an illustrative example, we consider the distribution given below:

Class Boundary	Frequency	Relative Frequency
46.5 – 49.5	2	$2/25 = 0.08$
49.5 – 52.5	3	$3/25 = 0.12$
52.5 – 55.5	5	$5/25 = 0.20$
55.5 – 58.5	8	$8/25 = 0.32$
58.5 – 61.5	4	$4/25 = 0.16$
61.5 – 64.5	3	$3/25 = 0.12$
Total	25	1.00

The histogram/frequency polygon for the given distribution is as drawn below.

Histogram/frequency polygon



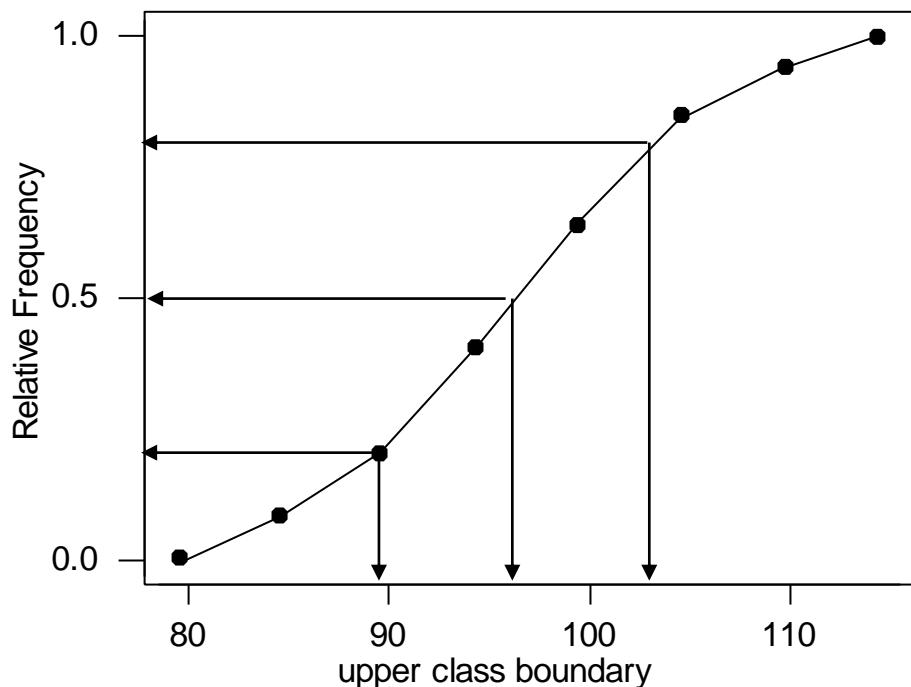
1-2.2.1.2 Cumulative Frequency Curve (or Ogive):

The cumulative frequency distribution shows the number of observations that fall above or below a specified value of observation. The cumulative frequency of a class is observed by cumulating (or summing) all frequencies up to the class. A graph obtained by plotting the cumulative points by smooth curve is called *cumulative frequency curve or Ogive*.

For example, we consider the following distribution:

Class boundary	Frequency	Cumulative frequency	Relative Cum. Frequency
79.5 – 84.5	5	5 + 0 = 5	0.0625
84.5 – 89.5	10	5 + 10 = 15	0.1875
89.5 – 94.5	15	15 + 15 = 30	0.3750
94.5 – 99.5	26	30 + 26 = 56	0.7000
99.5 – 104.5	13	56 + 13 = 69	0.8625
104.5 – 109.5	7	69 + 7 = 76	0.9500
109.5 – 114.5	4	76 + 4 = 80	1.0000
<i>Total</i>	60	-	-

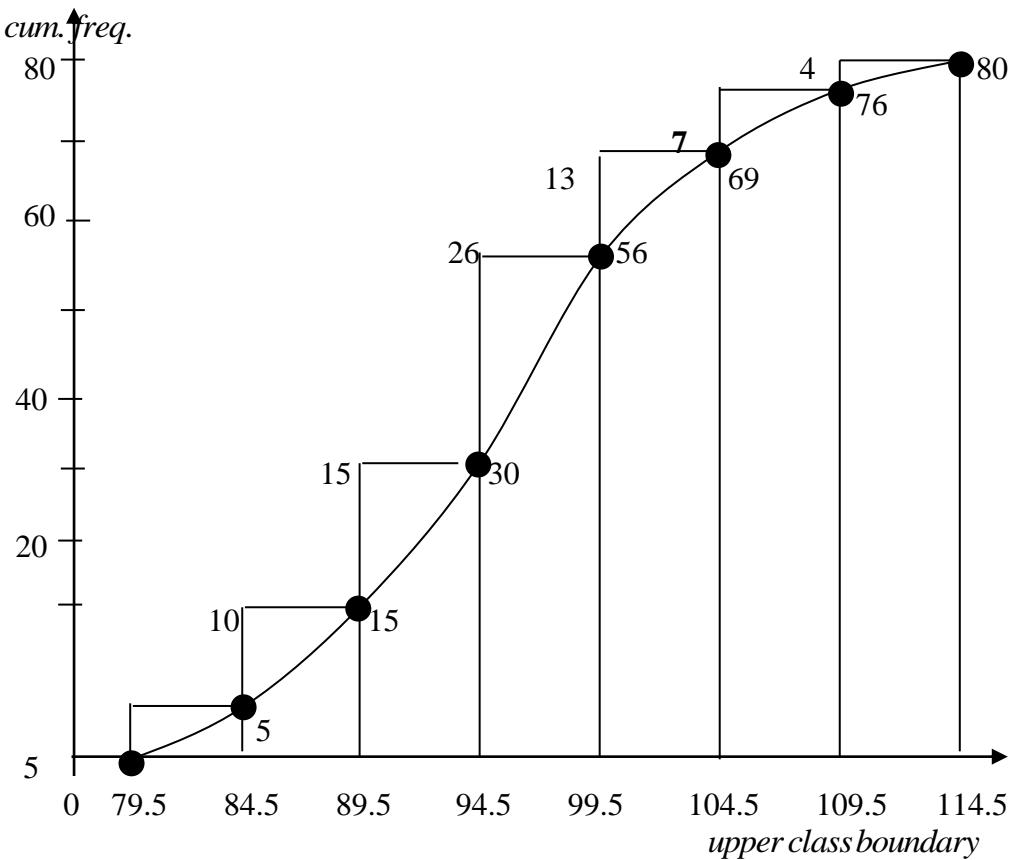
The relative cumulative frequency indicates the proportion of observations that are located above or below a given point or observation. It is defined as cumulative frequency divided by total frequency. The cumulative frequency curve is very useful for finding how many data points or observations are located above or below a given point.



From the curve we have, approximately,

- (i) 20% of the observations fall below 89.5
- (ii) 50% of the observations fall below 96.5
- (iii) 80% of the observations fall below 101.5

Plotting the class cumulative frequency as a rectangle over the corresponding class boundary gives a histogram with an appearance of a *staircase* or *step function*. A smooth curve joining the co-ordinates (upper class boundaries, cumulative frequencies) produces the cumulative frequency curve. This is shown below.



Example 2.5:

Many people experience allergic reactions to insect stings. These reactions differ from patient to patient not only in severity but also in time of reaction. The following data (measured in minutes) are on 40 patients who experience a systematic reaction to bee stings.

5.9	10.5	9.9	14.4	16.5	12.7	11.6	7.9	10.9	13.4
8.6	3.8	11.7	12.5	9.1	9.1	12.3	11.5	7.4	8.8
11.5	13.6	11.5	10.9	12.9	11.2	15.0	12.7	10.1	14.7
9.9	11.4	6.2	8.3	8.1	10.5	8.4	11.2	10.4	9.8

- (a) Group the data into six classes and obtain a relative frequency distribution.
- (b) Draw a histogram for the distribution and use it to find the mode.
- (c) Plot the cumulative frequency curve. Use it to estimate percentage of patients who have experienced a reaction within 10 minutes and the median.

Solution:

- (a) Given the number of classes, $k = 6$ we find the class width, using the *Sturges' Rule*,

$$C = \frac{16.5 - 3.8}{6} = 2.12 \approx 2.2, \text{ and the first class boundary:}$$

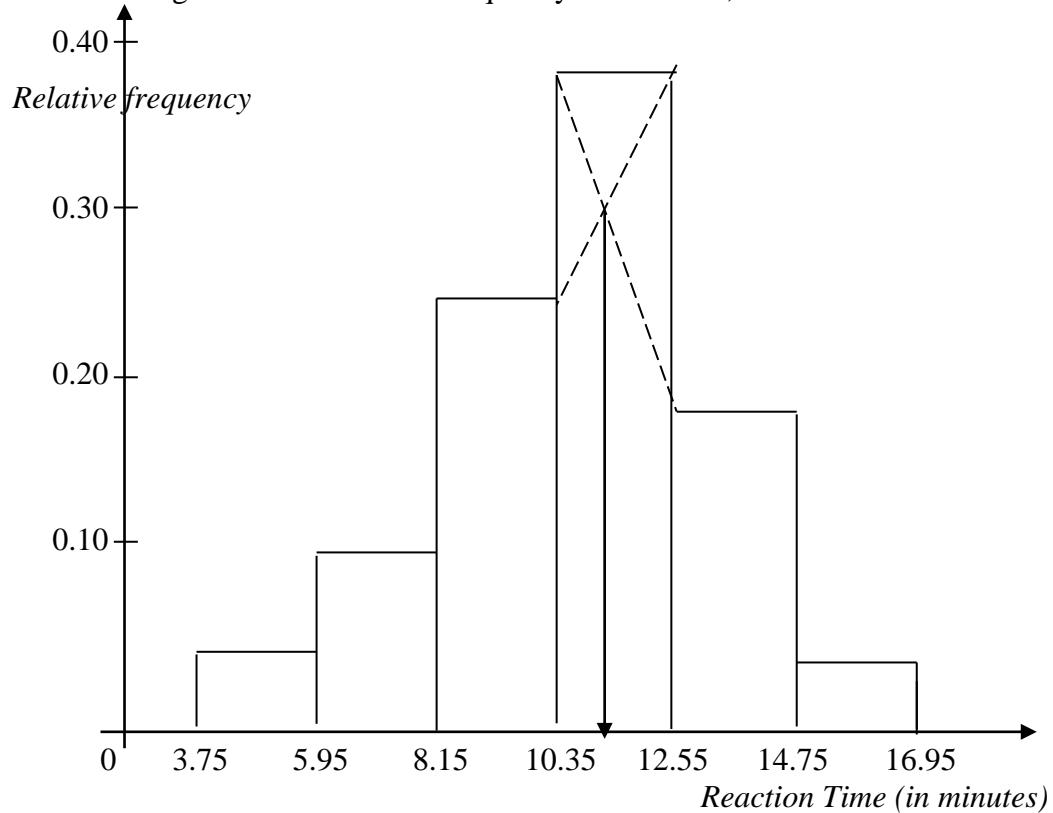
$$LB_1 = 3.8 - \frac{1}{2}(0.1) = 3.75$$

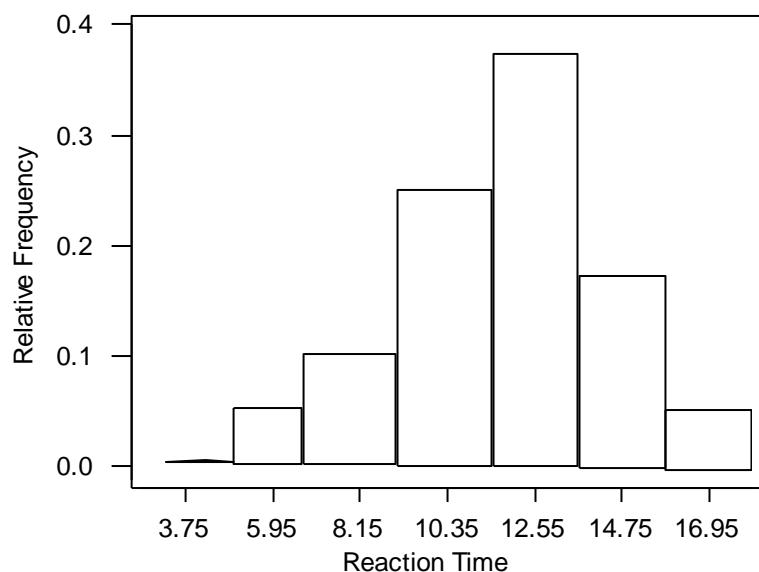
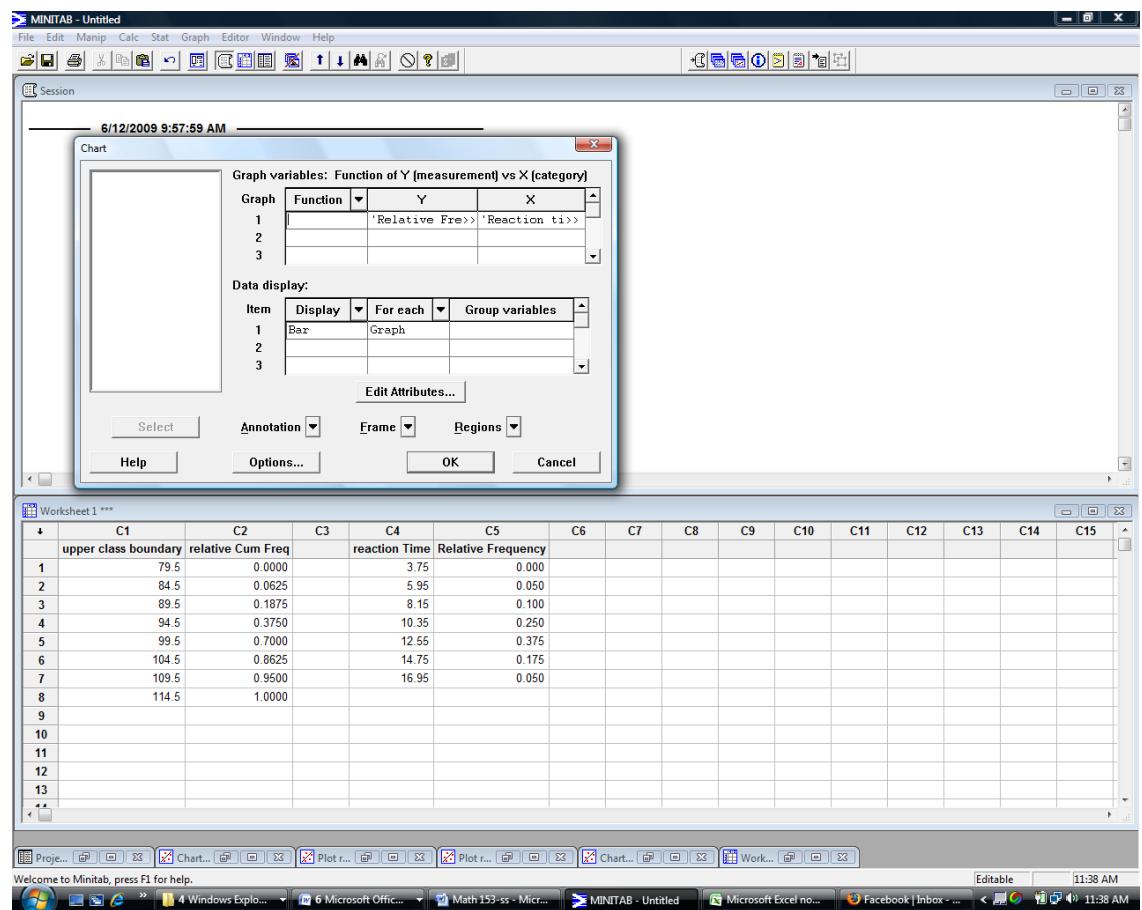
$$\text{and } UB_1 = 3.75 + 2.2 = 5.95$$

Hence the relative frequency distribution,

Time	Tally	Frequency	Relative Frequency
3.75 – 5.95	//	2	0.050
5.95 – 8.15	///	4	0.100
8.15 – 10.35	### ###	10	0.250
10.35 – 12.55	### ### ###	15	0.375
12.55 – 14.75	### //	7	0.175
14.75 – 16.95	//	2	0.050
<i>Total</i>	-	40	1.000

- (b) The histogram for the relative frequency distribution,



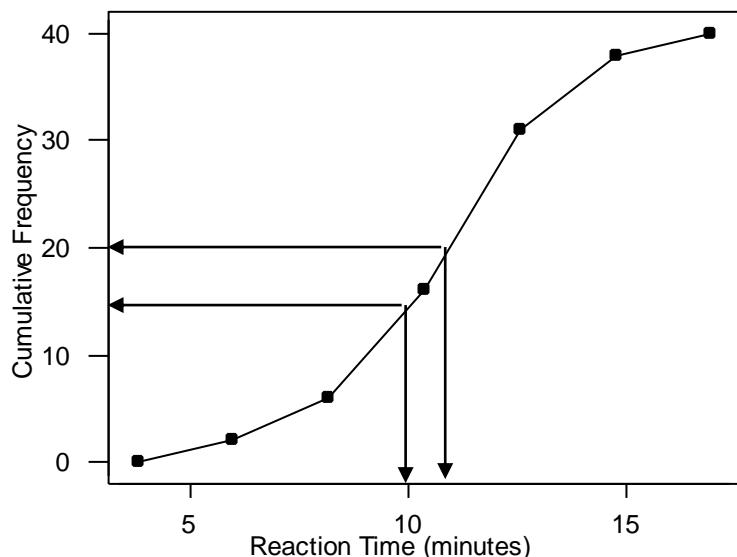


From the histogram, the *mode* = $10.35 + 1 = 11.35$ minutes

- (b) The relative/cumulative frequency distribution.

Reaction Time up to	Cumulative frequency	<i>Relative Cum. Frequency</i>
3.75	0	0.000
5.95	2	0.050
8.15	6	0.150
10.35	16	0.400
12.55	31	0.775
14.75	38	0.950
16.95	40	1.000
<i>Total</i>	-	-

We now draw cumulative frequency curve.



From the curve,

- (i) The percentage of patients who experience a reaction within 10 minutes

$$= \frac{15}{40} \times 100\% = 37.5\%$$

- (ii) The median is the time that a reaction occurred in half of the patients

$$= 10.35 + 1.00 = 11.35 \text{ minutes}$$

2. The following data are on the amount of time (in hours) 80 college students spent their leisure time during a typical school week.

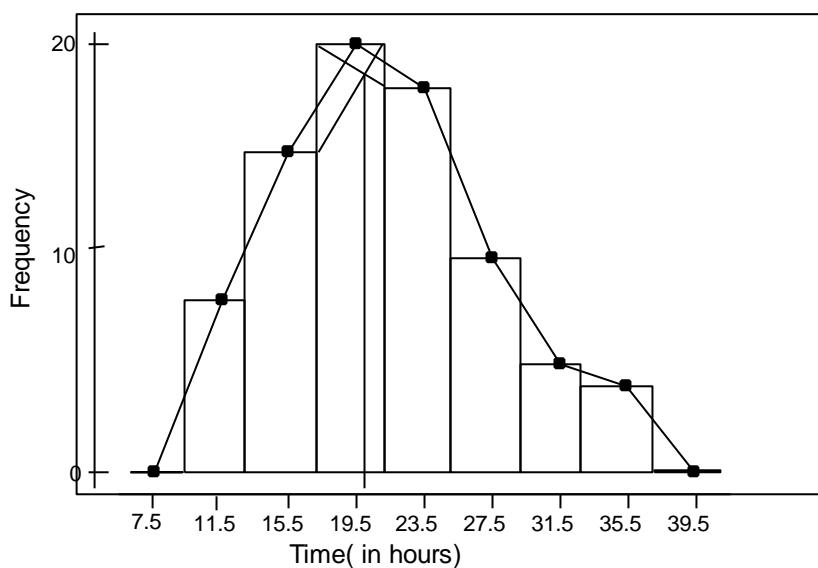
11	23	24	18	14	20	24	26	24	23	21
17	16	12	15	19	26	16	20	22	30	13
20	35	27	13	18	29	22	37	28	34	32
23	22	21	23	19	21	31	20	27	16	28
19	18	12	27	15	21	25	32	10	23	17
12	15	24	25	37	22	17	18	15	19	20
23	18	17	15							

Construct a frequency distribution and histogram, given classes; 10 – 13, 14 – 17, 18 – 21, etc. and use it to construct a histogram.

Solution:

- (a) The (relative) frequency distribution and histogram/frequency polygon:

Time (in hours)	Tally	Frequency	Relative frequency
10 – 13	/// / //	8	0.0100
14 - 17	/// / /// - / //	15	0.1875
18 - 21	/// - / / / - / / / - / / /	20	0.2500
22 – 25	/// / / / - / / / / /	18	0.2250
26 – 29	/// / / /	10	0.1250
30 – 33	/ / /	5	0.0625
34 – 37	/ / /	4	0.0500
<i>Total</i>	-	80	1.0000



From the histogram, modal leisure time = $17.5+2.5 = 30.0$ hours

1-2.2.1.3 Exploratory Data Analysis (EDA):

Exploratory data analysis is a process of using statistical tools (such as graphs and numerical measures) to investigate data sets in order to understand their important characteristics. It is a recently developed technique for providing easy-to-construct diagrams that summarize and describe a set of data. The five important characteristics for describing, exploring, and comparing data sets which need to be noted are as listed below:

- *Centre*, a representative or average value that indicates where the middle of the data set is located.
- *Variation*, a measure of the amount that the data values vary among themselves.
- *Distribution*, the nature or shape of the distribution of the data (such as bell-shaped, uniform, or skewed)
- *Outliers*, data values that lie very far away from the vast majority of the values.
- *Time*, changing characteristics of the data over time.

These characteristics are well-remembered by the phrase “Computer Virus Destroy or Terminate (CVDOT)”

The techniques or diagrams of EDA discussed in this section are the *Dotplots*, *Boxplots* and *Stem and Leaf*.

- *Dotplots*: A dot plot is a plot that displays a dot for each value in a data set along a number line. If there are multiple occurrences of a specific value, then the dots will be stacked vertically.
- *Boxplots*: Boxplots are useful for revealing the centre and spread of the data as well as the outliers of the data. The construction of boxplot requires that we first obtain the minimum value, maximum value, and the quartiles. The boxplot graph consists of a line and a box indicating the five-number summary. The five-number summary consists of the minimum value, first quartile, median, upper quartile, and the maximum value.
- *The Stem-and-Leaf Plots*: The stem-and-leaf plot was originally developed by John Tukey. It is extremely useful in summarizing reasonably sized data sets (usually under 100), and unlike histograms, results in no loss of information. The stem-and-leaf plot is constructed by first separating each observed value in

the data set into two parts, called *stem* and *leaf*. The stems are then arranged vertically in ascending order of magnitude and the leaves are recorded against their corresponding stems. A stem-and-leaf plot has an advantage over a grouped frequency distribution, since a stem-and-leaf plot retains the actual data by showing them in a graphic form.

Example 2.6:

The weights of 33 students in Department of Mathematics, KNUST are given below. Construct a stem-and-leaf, boxplots, and dotplots diagrams to summarize the data.

143	158	136	127	132	132	126	138	119	104	113
90	126	123	121	133	104	99	112	120	107	139
122	137	112	121	140	134	133	123	150	115	141

Solution

- (a) Stem-and-Leaf Diagram: The stems for the data are 9, 10, 11, 12, 13, 14, and 15. We arrange them vertically and each leaf is recorded against its corresponding stem. The R program out for the diagram is as shown below:

```
> p=c(data)
> stem(p, scale=1)
```

The decimal point is 1 digit(s) to the right of the |

9 | 0 9

10 | 4 4 7

11 | 2 2 3 5 9

12 | 0 1 1 2 3 3 6 6 7

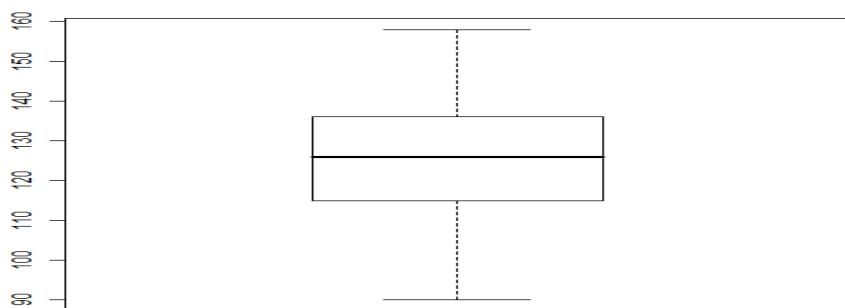
13 | 2 2 3 3 4 6 7 8 9

14 | 0 1 3

15 | 0 8

- (b) The boxplot diagram (by the R program) is given by

```
> boxplot(p)
```



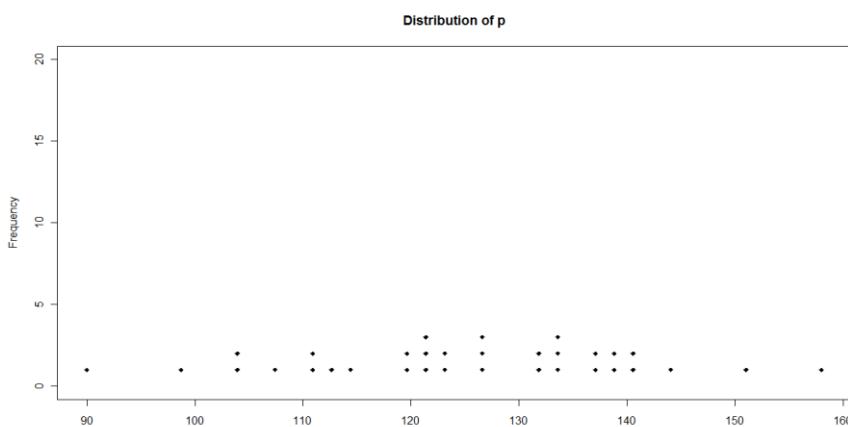
The five-number summary is given by

Minimum	1 st Quartile	Median	3 rd Quartile	Maximum
90.0	115.0	126.0	136.0	158.0

These results also be obtained by the R command,
> summary (p).

- (c) The dotplots diagram is given by

> boxplot (p)



1-2.2.2 Graphical Representation of Qualitative Data:

The most commonly used graphical representation of qualitative data is *bar charts, pie Charts and line or time series graphs*.

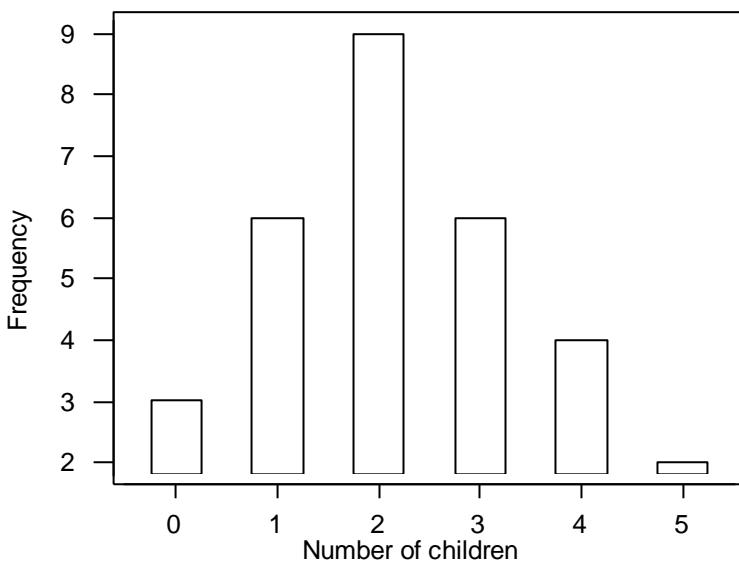
- *Bar Charts:* A bar chart consists of rectangular bars with equal widths and separated by gaps. Then length or heights of the bars are proportional to the (relative) frequencies of the categorized data. The bars are separated by gaps to emphasize the fact each class is a separate category. They might be used to compare, for example, one year or place with others. The length of the bars is the basis of the comparison. A bar chart is classified as either being simple, *multiple/compound* or *component* depending on sets of data being compared. The simple bar chart represents a set of data while the multiple/compound compares a number of items over a period of time. The compound bar chart displays the various categories of data as components of the whole set of data.

Example 2.7 (Simple Bar Charts):

The bar Consider the data given in *Example 2.1* with the frequency distribution.

No. of children	frequency	Relative frequency
0	3	0.1
1	6	0.2
2	9	0.3
3	6	0.2
4	4	0.15
5	2	0.05
<i>Total</i>	30	1.00

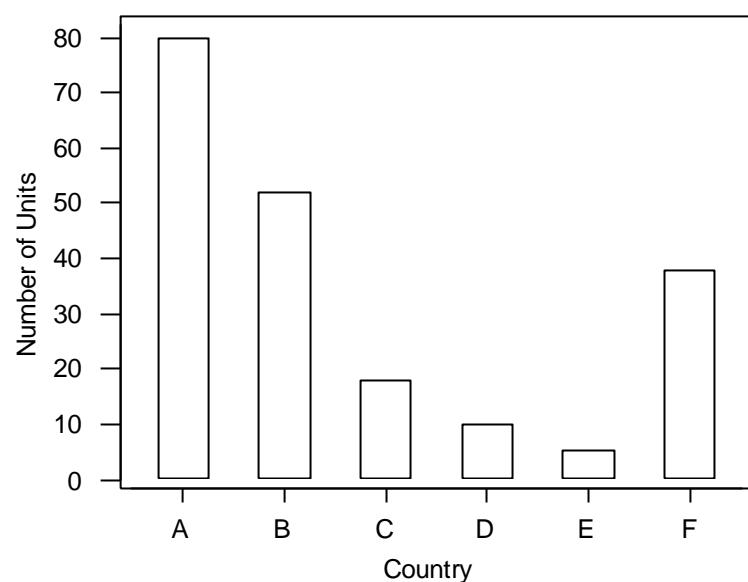
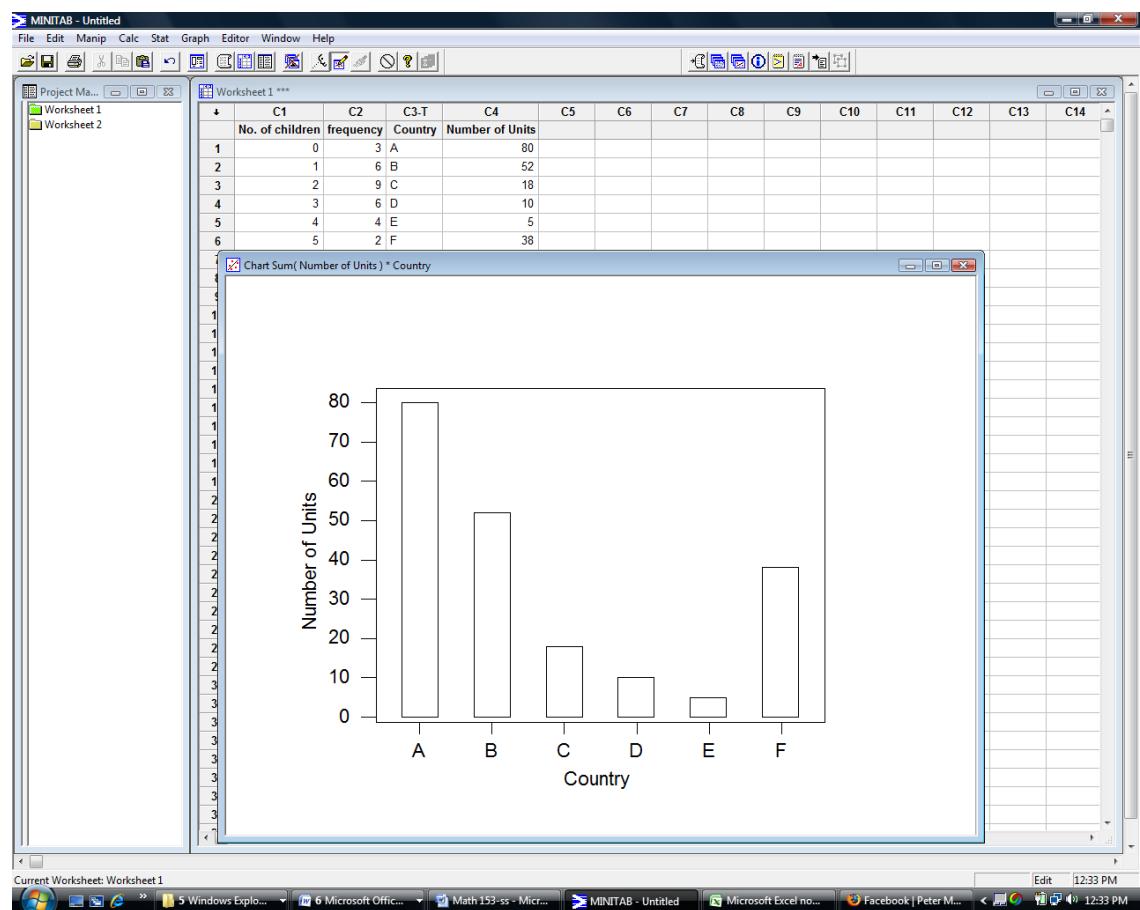
The bar chart for the above distribution show below



- (b) The number of units of nuclear reactors in some countries in 1984 is given in the table below.

Country	A	B	C	D	E	F
No. of Units	80	52	18	10	5	38

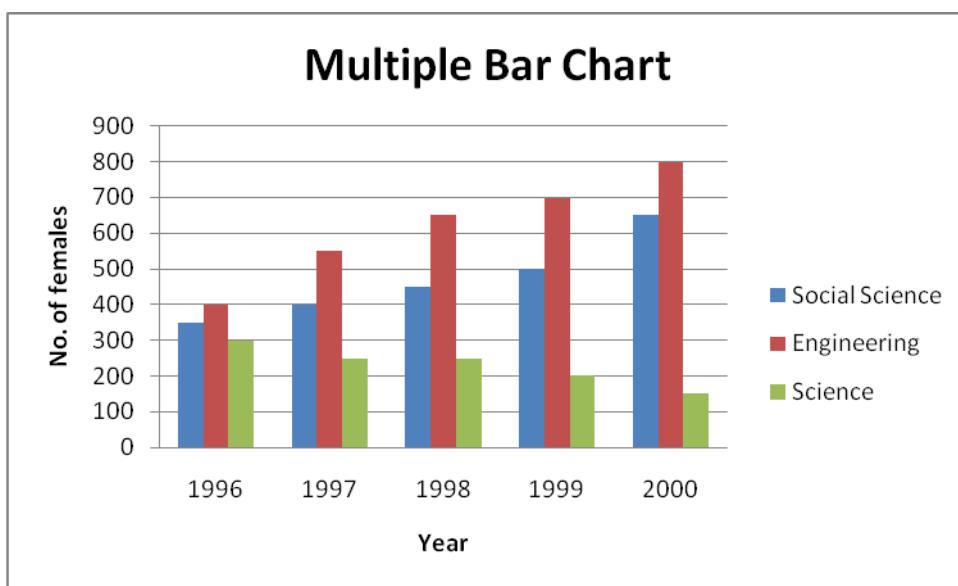
This can be displayed using the bar chart shown below:



Example 2.8 (Multiple Bar Charts):

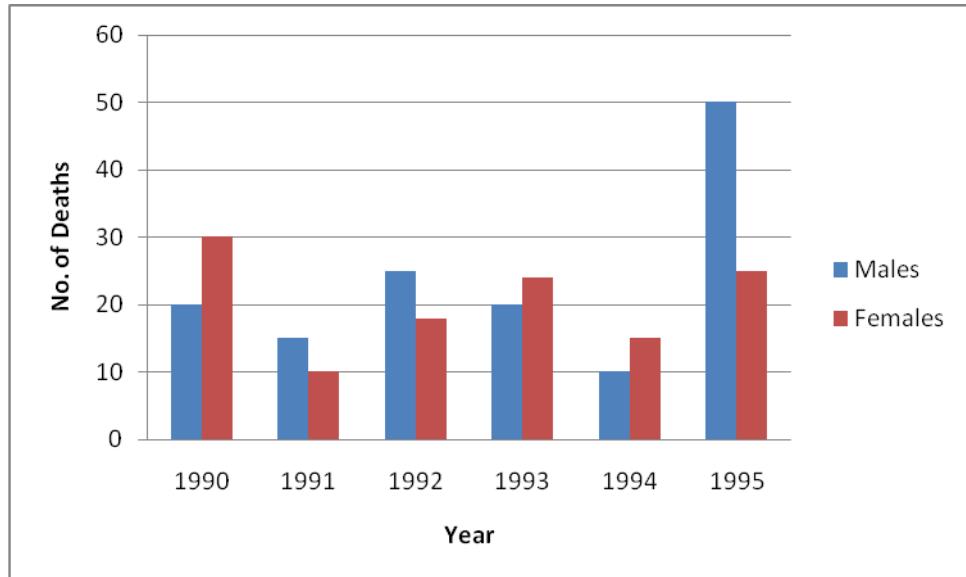
The number of females opted to offer programmes in Social Sciences, Engineering and Science for the period 1996–2000 in KNUST is as in the table below. The given data are displayed in the multiple bar charts, also shown below.

Year	Social Science	Engineering	Science
1996	350	400	300
1997	400	550	250
1998	450	650	250
1999	500	700	200
2000	650	800	150



The death rate (per 1000) in a year of males and females of a disease in community over a period of 6 years is given as follows:

Year	1990	1991	1992	1993	1994	1995
Males	20	15	25	20	10	50
Females	30	10	18	24	15	25



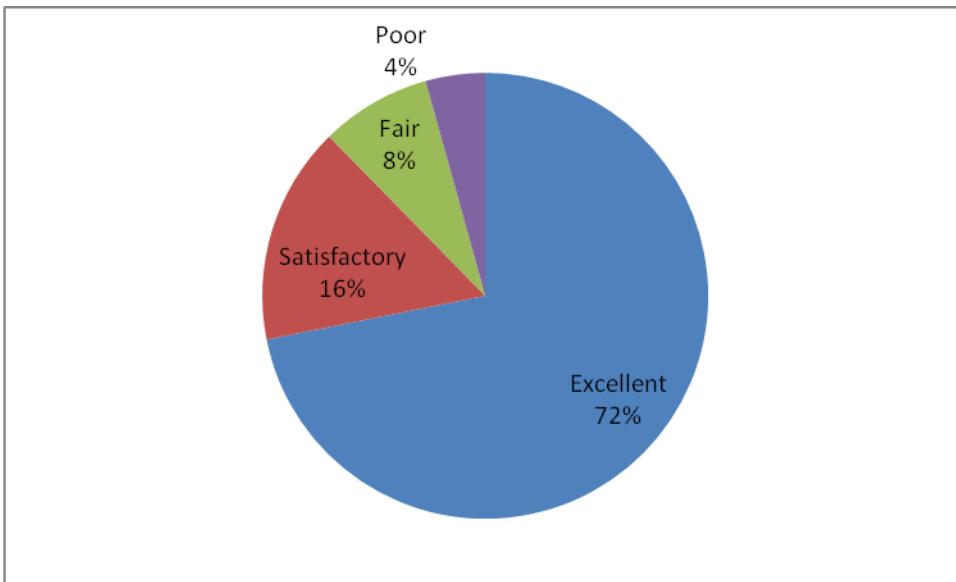
- **Pie Charts:** A pie chart is a circular diagram giving various fractions of section of a given data. The total number of observations of the data is represented by a *pie* which is denoted by a circle. The pie is then cut into slices (sectors) where each slice represents a category of the data. The size of a slice is proportional to the relative frequency of a category. A pie chart is often used in newspapers, magazines and articles to depict budgets and other economic information. In constructing a pie chart, we represent the total number of observations by a circle of an angle of 360^0 . The angle of a slice (sector) at centre of a pie (circle) is given by the product: $\text{Relative Frequency} \times 360^0$.

Example 2.9:

- (a) Consider the responses regarding the relief provided by a pain-killing drug.

Response	Frequency	Relative frequency	Angle of sector
Excellent	30	0.20	$0.20 \times 360^0 = 72^0$
Satisfactory	66	0.44	$0.44 \times 360^0 = 158.4^0$
Fair	36	0.24	$0.24 \times 360^0 = 80.4^0$
Poor	18	0.12	$0.12 \times 360^0 = 43.2^0$
Females	150	1.00	360^0

The pie chart for the given data is as shown below.



- (b) Consumers spend their incomes on a vast array of goods and services. The data below provide a guide summary of how the average consumer dollar is spent.

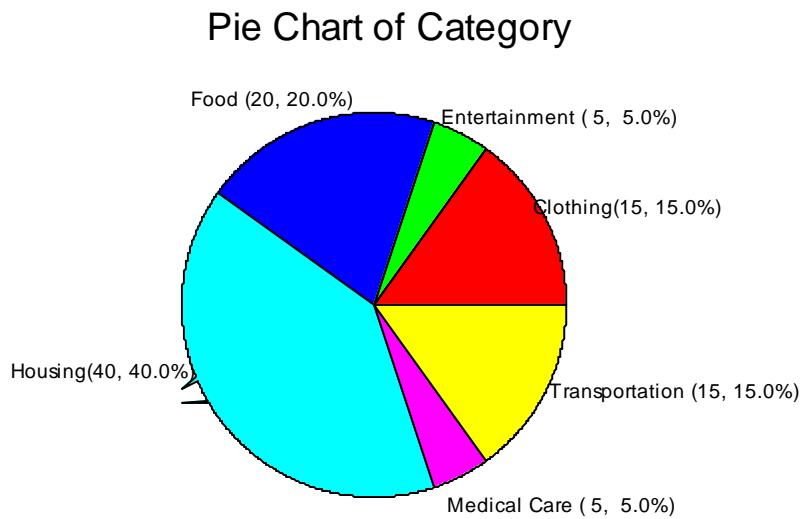
<i>Category</i>	<i>Percentage of income</i>
Medical Care	5
Clothing	15
Entertainment	5
Housing	40
Food	20
Transportation	15

- (i) Summarize the information in the form of pie chart.
(ii) What area represents the largest piece of the pie?

Solution:

<i>Category</i>	<i>Percentage of income</i>	<i>Angle of Sector</i>
Medical Care	5	$0.05 \times 360^\circ = 18^\circ$
Clothing	15	$0.15 \times 360^\circ = 54^\circ$
Entertainment	5	$0.05 \times 360^\circ = 18^\circ$
Housing	40	$0.05 \times 360^\circ = 144^\circ$
Food	20	$0.05 \times 360^\circ = 72^\circ$
Transportation	15	$0.05 \times 360^\circ = 54^\circ$
<i>Total</i>	100	360°

(i) The required pie chart is as

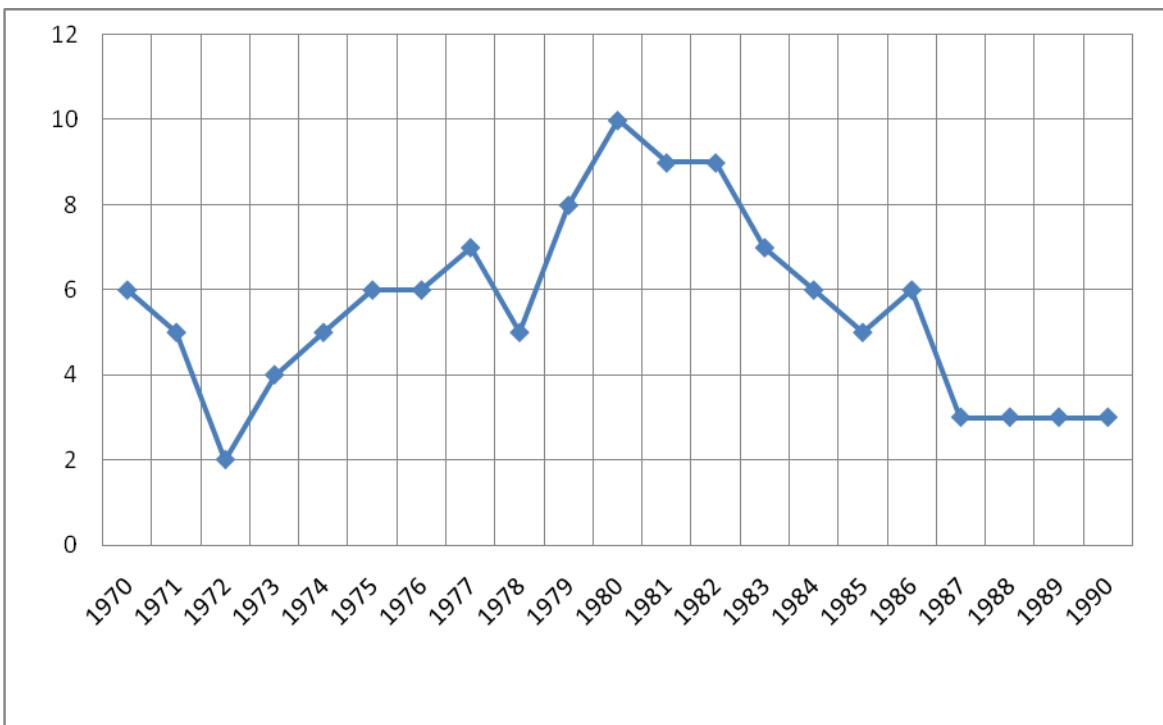


(ii) The largest piece of pie is “*housing*”

- *Time Series Graphs:* Time is an important factor that contributes to variability in data. Data collected over time can be displayed using a line chart (better known as time series graph). A time series graph is useful for describing data over a period of time. The graph is obtained by plotting the values of the observations (vertical axis) against time (horizontal axis) which could be days, weeks, months, etc. From the graph we see trends, cycles or other broad features of the data.
- A control chart is another useful way to examine the variability in time-oriented data.

For example, the graph below represents a time series plots of deaths from a strange disease for the period, 1970-1990.

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Number	6	5	2	4	5	6	6	7	5	8	10
Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	-
Number	9	9	7	6	5	6	3	3	3	3	-



It is seen from the graph that the largest number of deaths during the 20-year period occurred in 1980 and the smallest number occurred in 1972. The number of deaths had initially decreased from 6 in 1970 to 2 in 1972 and increased steadily to 10 in 1980. It then decreased slowly from 1980 to 3 in 1987 and began to stabilize to 1990.

1-2.3 Trial Questions 1-2:

- 2.1(a)** Hospital records of 40 peptic ulcer patients provided the following information about their blood types: *A*, *B*, *O* and *AB*.

O	A	B	O	A	A	A	O	O	AB
O	AB	O	A	B	O	O	O	AB	A
B	A	A	O	O	A	A	O	AB	O
O	A	A	B	A	O	A	O	O	B

Prepare a frequency distribution table based on the blood types of peptic ulcer patients.

- 2.1(b)** The following are weights (in kg) of 80 persons measured to the nearest 0.01 kilogramme.

37.67	36.42	43.57	55.60	57.76	74.60	63.40	73.70
109.46	40.27	97.23	30.63	47.93	63.72	28.30	70.23
42.63	22.27	65.60	57.40	80.93	45.67	42.78	65.23
27.20	52.36	50.72	53.36	28.60	72.20	87.20	48.33
52.30	64.90	19.67	27.32	38.60	77.40	50.72	33.20
50.63	58.73	103.62	63.76	50.36	37.40	35.20	89.40
47.60	83.23	24.23	74.60	84.72	55.50	60.37	67.20
42.74	95.43	105.36	43.60	54.87	57.60	48.98	70.58
52.89	91.30	51.30	33.40	58.60	63.83	93.60	59.89
58.36	64.83	58.60	25.15	100.55	75.22	38.85	90.66

- (i) Construct a frequency distribution for the above data.
(ii) Use the above distribution to draw a histogram and describe the shape of the distribution.

- 2.1(c)** The following are the number of babies born during a year in sixty community hospitals.

56	57	30	55	27	45	56	48	45	49	32
47	57	46	37	58	52	34	54	42	32	59
35	24	59	54	32	26	40	28	53	54	29
42	42	53	50	34	39	26	59	58	49	53
30	53	21	28	29	24	52	57	43	46	54
31	22	31	24	57						

- (i) Construct a relative frequency distribution table,
(ii) A histogram and a relative frequency polygon,

- 2.2(a)** The table shown below gives the total monthly rainfall in Kumai in a certain year.

Month	June	July	August	Sept.	Oct.	Nov.
Rainfall (mm)	250	130	230	300	50	40

Draw a bar chart for the above data.

- 2.2(b)** The volume of raw materials and processed goods (in metric tones) produced by a certain factory in the first 5 years of its operations are shown in the table below.

Year	1	2	3	4	5
Raw Materials	20	30	35	40	80
Processed Goods	15	45	18	20	70

Draw a multiple bar chart to represent the data.

- 2.3(a)** The following are scores obtained by the students in a Statistics examination paper.

69	47	82	73	99	97	55	18	100	85	77
62	58	43	21	85	68	50	43	91	85	60
80	54	94	88	79	95	66	46	81	51	81
75	88	80	94	74	70	71	70	20	50	48

- (i) If you have to transform the data into a frequency distribution, what number of classes and class width would you use?
- (ii) Present the data in a frequency distribution and draw a histogram.
- (iii) Comment on the shape of the distribution.

- 3.3(b)** The following are the number of births per year per 1,000 population for 60 countries.

34	17	25	37	19	19	27	45	24	19	15
31	24	22	32	12	13	16	18	14	12	16
18	27	10	10	15	15	20	22	16	10	17
18	35	35	15	17	20	18	19	13	13	13
18	30	20	32	22	15	31	28	40	43	31
44	34	24	38	32	50	33	11	28	55	42

- (i) Organize the data into a frequency distribution. Start with a lower limit of 10 and use an interval of 5.
- (ii) Draw a histogram and frequency polygon.

- (iii) Draw a cumulative frequency polygon.
- (iv) If a country has a birth rate of 15 per 1,000 population, what percentage of the countries has a birth rate that is equal or greater than the birth rate in that country?

SESSION 2-2: NUMERICAL MEASURES

Graphical methods are very useful for presenting and conveying a rapid general description of data visually. However, in the absence of these visual representations, it becomes extremely difficult to give a verbal description or analysis of the data. Graphical methods are also not effective for purposes of performing statistical inference. These limitations of the graphical methods can be overcome by the use of numerical descriptive measures. Numerical measures convey a good mental picture of graphical representation of the frequency distribution of data collected and are also useful in making inferences concerning the sampled population. A numerical descriptive measure is a single value that provides information about the data collected. Most descriptive measures used to summarize a set of observations or data fall into *Measures of Central Tendency, Dispersion, Position and Shape*.

2-2.1 Measures of Central Tendency

Measures of central tendency are the averages which determine the *central location* or *middle* of the data. An average, in Statistical, is a numerical value that is typical of (and effectively represents) a given set of data. There are several types of such averages, each possessing particular properties. The most commonly used measures of central tendency are the *arithmetic mean, weighted mean, trimmed mean, mode, median, geometric mean and harmonic mean*.

2-2.1.1 The Arithmetic Mean

The *Arithmetic mean* or simply, the *mean* is the best known and most commonly used average. It is defined for both ungrouped and grouped data as follows:

Let $x_1, x_2, x_3, \dots, x_n$ be the observations forming the data set. Then the mean of the n observations is defined by which sum of observations is divided by the total number of observations. However, if the observations, $x_1, x_2, x_3, \dots, x_k$ occur in a frequency

distribution with corresponding frequencies, $f_1, f_2, f_3, \dots, f_k$, then the mean is given by

$$\begin{aligned}\bar{x} &= \frac{1}{n} \left(\sum_{i=1}^n x_i \right) = \frac{1}{n} (x_1 + x_2 + \dots + x_n), \text{ or} \\ &= \frac{1}{n} \left(\sum_{i=1}^k f_i x_i \right) = \frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{f_1 + f_2 + \dots + f_k},\end{aligned}$$

where $n = \sum_{i=1}^k f_i$ the total frequency.

When the observations are grouped in a frequency distribution, then x_i becomes the class mark or midpoint of the i^{th} class boundary with frequency, f_i . The mean is then

defined by $\bar{x} = \frac{1}{n} \left(\sum_{i=1}^k f_i x_i \right)$.

The mean may also be computed using the formula, $\bar{x} = A + \frac{1}{n} \left(\sum_{i=1}^k f_i d_i \right)$, where $A =$

assumed mean, usually chosen to be the middle class mark $d_i = x_i - A$, called deviation of i^{th} class mark.

The arithmetic mean has the following properties:

- The algebra sum of the deviations of from the arithmetic mean is zero. That is the centre of gravity of observations - a point of balance and serves as the most typical central value of the data since $\sum_{i=1}^n x_i = n \bar{x}$.
- The sum of squares of the deviation of the observations from the mean is less than the squared deviations from any other point in the data. That is, $\sum_{i=1}^n (x_i - \bar{x})^2$ is minimum.
- If fixed value a is added or subtracted from each of the observations, the mean changes by the same amount. However, if each observation in the data is multiplied or divided by a fixed constant b the means is also multiplied or divided by b .

The arithmetic mean, however, has some advantages and disadvantages which are presented as follows:

- It is unique. This means that for a given set of data there is one and only one arithmetic mean.

- The concept of the mean is familiar to most people and intuitively clear. It is widely understood and well suited for further statistics analysis.
- Its computation uses all the values of the observations. Hence it is very sensitive by the extreme values of the data. That is, the arithmetic mean may be distorted by the extremely high or small values of the data.
- It may result in an impossible value where the data are discrete, for example, having 3.53 children!!

2-2.1.2 The Weighted and Trimmed Means:

We may sometimes associate with each observation certain weighting factor or weight depending on the significance attached to the observation. The computed mean is called the *weighted mean*. It is used when a simple average fails to give an accurate reflection of the relative importance of the items or observations being averaged. The weighted mean is defined as follows:

Let the observed values, $x_1, x_2, x_3, \dots, x_n$ form the set of data with weights $w_1, w_2, w_3, \dots, w_n$ respectively. Then the weighted mean,

$$\bar{x} = \frac{1}{n_w} \sum_{i=1}^n w_i x_i = \frac{w_1 x_1 + w_2 x_2 + \dots + x_n}{w_1 + w_2 + \dots + w_n},$$

where $n_w = \sum_{i=1}^n w_i = \text{total weight}$.

Occasionally, a given set of data will have one or more unusually small and/or unusually large observations that significantly influence the value of the mean. In this situation, the mean may provide a poor description of the central location of the data. To remove the effect of the unusually small or large observations we *eliminate* or *trim* a percentage of the small and large observations from the data. The arithmetic mean of the remaining data is called the *trimmed mean*. For example, 5% trimmed of the mean removes the smallest 5% of the data values and the largest 5% of the data values. The 5% trimmed mean is then computed as the mean of the middle 90% of the data. In general, an α percent trimmed mean is the mean obtained after α percent of the smallest and α percent of the largest items in the data have been removed.

2-2.1.3 The Geometric and Harmonic Means

The geometric mean (g_m) is defined as the n th root of the product of the n observations, x_1, x_2, \dots, x_n forming the data. That is,

$$g_m = \sqrt[n]{x_1, x_2, \dots, x_n} = \sqrt[n]{\prod_{i=1}^n x_i} = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}$$

The computation of the geometric mean is made quite easier by taking its logarithm.

That is, $\log(g_m) = \frac{1}{n} \sum_{i=1}^n \log x_i$, from which we can take antilogarithm of $\log(\bar{x}_g)$ and obtain \bar{x}_g (i.e., $g_m = \text{anti log}(g_m)$). For k -grouped data, the geometric mean is defined as

$$g_m = \sqrt[n]{x_1^{f_1} \cdot x_2^{f_2} \cdot \dots \cdot x_k^{f_k}} = \left(\prod_{i=1}^k x_i^{f_i} \right)^{\frac{1}{k}}, \text{ where } x_i \text{'s are the class marks.}$$

The geometric mean is used primarily to average data for which the ratio of consecutive terms remains approximately constant, rates of change, ratios, economic index numbers, population sizes over consecutive time periods, etc. It has the following properties:

- The geometric mean cannot be computed if any of the observations is zero or negative. It holds for only positive observations.
- The product of the observations remains unchanged if the geometric mean is substituted for each individual observation.
- The sum of deviations of the logarithms of the original observations above or below the logarithm of the geometric mean is equal to zero. That is,

$$\sum_{i=1}^n (\log x_i - g_m) = 0 .$$

- The geometric mean may be more a representative average than the arithmetic mean when the values are rising or falling at steady rate overtime. For example, if a population of a state is growing at a rate of 10% every 10 years and has a population of 1 million in 1980, then the population of 1990 will be 1.1 million while that of 2000 is 1.2 million. The geometric mean, $g_m = \sqrt[3]{(1.0)(1.1)(1.2)} = 1.09696131$, which is a bit lower than the arithmetic

$$\text{mean, } \bar{x} = \frac{1.0+1.1+1.2}{3} = 1.1.$$

- For a given set of data, x_1, x_2, \dots, x_n , the arithmetic mean, \bar{x} is greater than the geometric mean, g_m .

The harmonic mean (h_m) is the reciprocal of the arithmetic mean of the reciprocals observations. That is, given the observations, x_1, x_2, \dots, x_n ,

$$h_m = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \text{ or } \frac{1}{h_m} = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$$

For grouped data, the harmonic mean,

$$h_m = \frac{n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \dots + \frac{f_k}{x_k}} = \frac{n}{\sum_{i=1}^k f_i \frac{1}{x_i}}$$

The harmonic mean is useful in processing ratio data that have physical dimension, example, miles per gallon, cost per mile, etc.. It is used to work out average speeds, rates of production, etc.

2-2.1.4 The Median

The median is the middle-ranked value of an ordered array data. It divides the data set into two equal parts after the observations have been arranged in order of magnitude. It is computed as follows:

Let x_1, x_2, \dots, x_n be the observations arranged increasing order of magnitude. The median, denoted M is defined as the middle most measurement in the ordering, if the number of measurements, n is odd. If n is even, the median becomes the arithmetic mean of the two middle most measurements. That is,

$$M = \begin{cases} x_{\frac{1}{2}(n+1)^{th}}, & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{\frac{1}{2}n^{th}} + x_{(\frac{1}{2}+1)n^{th}}), & \text{if } n \text{ is even} \end{cases}$$

For grouped data, the median is defined as the point at or below/above which exactly 50% of the observations fall. The class interval in which the median is located is called *median class interval*. The median is estimated approximated by the following methods.

- *Use of Histogram:* The median is obtained by drawing a vertical line dividing the histogram into two equal parts.
- *Use of Cumulative Frequency Curve:* This is used to determine the 50th observation, which is the median.
- *The Interpolation Method:* This estimates the median from median class

boundary using the formula, $m = l_m + \left(\frac{\frac{n}{2} - f_{cm}}{f_m} \right) C_m$, where

l_m = lower class boundary of the median class, f_{cm} = cumulative frequency just before the median class, f_m = frequency of the median class, C_m = class width of median class boundary and n = total number of observations (total frequency)

The median has the following properties:

- The median divides the set of data in such a way that at least 50% of the observations are equal to or less than it and at least 50% of the observations are equal or greater than it.
- The median is influenced only by the number of observations and not by the observations in the data. It is therefore a useful or highly desirable measure for skewed distributions such as income and scores like grades and rates.
- The sum of the absolute deviations of the observations from the median, M is less than the sum of the absolute derivations from any other point in the distribution. That is, $\sum_{i=1}^n |x_i - M|$ is a minimum. The median is often considered for analysis because of this property.

The following gives some advantages and disadvantages of the median:

- It is not affected by the extremely high or low values and therefore becomes useful when these extreme values are unknown.
- It is easy to compute and always exists.
- It may fail to reflect the full range of value and is unsuitable for further statistical analysis.
- Its computation ignores completely the actual size of the observations except those in the middle of the data.
- It is not likely to be a representative measure when the observations are few.

2-2.1.5 The Mode

The mode is defined as the most frequent observed value of a given set of observations. For example, if more people die of malaria than any other disease, then we say that malaria is the modal cause of death. For ungrouped data, the mode is determined by a mere inspection where we note the most frequent observation as would be illustrated by the given examples.

The typical usage of the mode is as follows:

- A modal grade of students is the grade most students receive.
- The most typical wage usually refers to the modal wage.
- Modal size of shoe is the typical size in the sense that more people buy this size than any other.
- The mode is useful in business planning for identifying those products in greatest demand. For example, a shirt or dress manufacture is interested in the size which is of greatest demand. Similarly, in scheduling the production of a drug, a manufacturer is interested in the drug that is most commonly prescribed by physicians. These measurements are best described by the mode.

Some advantages and disadvantages of the mode are:

- It is more appropriate average to use than the mean in situations where it is useful to know the most common observation, where large proportion of the observations are equal to it. For example, type of product mostly demanded by customers.
- It is easy to obtain and not affected by the extreme values of the data.
- It is mostly used by manufacturers since it gives a better idea of what particular size of a product to manufacture in excess of the others. For example, a shoe-maker is more interested in the modal size of a shoe he manufactures than the mean or median.
- It does not take into account all the observations and may not be unique.
- It is unsuitable for further statistical analysis.

For a grouped data we shall have a set of observations occurring frequently in a particular class, called *model class*. The mode from a grouped data can only be

approximately estimated by the following methods, each of which may give different value.

- *The Crude Method:* This uses the model class mark as the mode.
- *Use of Interpolation by formula:* The mode denoted M_0 is estimated by the

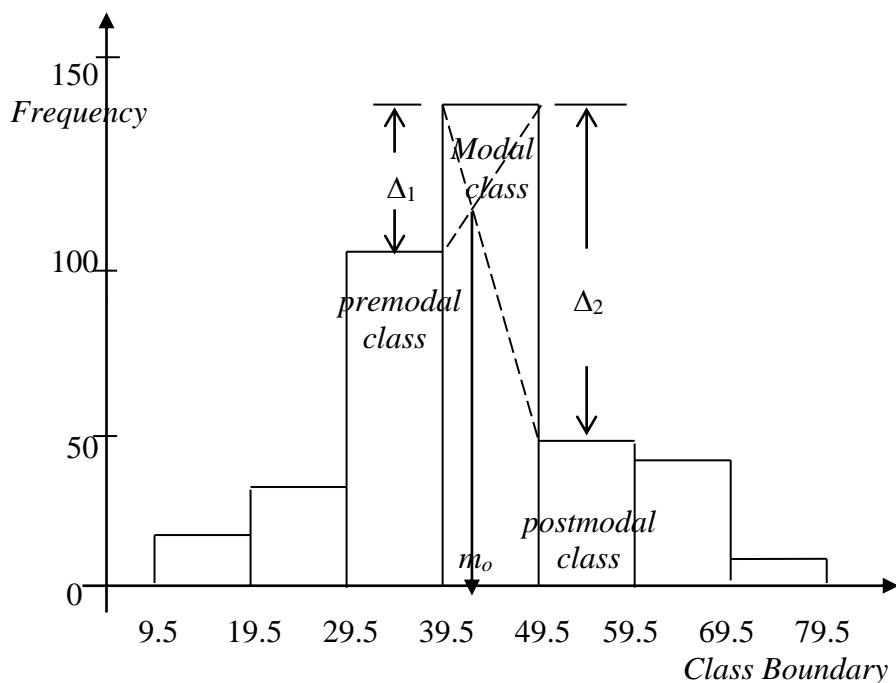
$$\text{formula, } m_0 = l_0 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) C, \text{ where}$$

l_0 = lower class boundary of modal class

Δ_1 = the absolute difference between the frequencies of pre-modal and modal classes (i.e. excess of modal frequency over frequency of next lower class)

Δ_2 = the absolute difference between the frequencies of the post-modal and modal classes (i.e. excess of modal frequency over frequency of next higher class)
 C = class width of model class

- *Use of histogram:* The mode is obtained from the modal class as illustrated in the diagram below.



Example 2.10

2.10(a) The following are measurements of ages (in years) of twelve school children.

Calculate the mean age of the school children.

10.3 11 13 8.3 5.7 10 11 14 7.5 8.2 7.8 9

(b) Consider the frequency distribution of the size of households for 65 workers.

Calculate the mean of the household size.

Size, x_i	5	6	7	8	9	10	11
No. of workers, f_i	8	10	15	13	8	6	5

Solution:

(a) The mean age of school children,

$$\begin{aligned}\bar{x} &= \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \\ &= \frac{10.3 + 11 + 13 + 8.3 + 5.7 + 10 + 11 + 14 + 7.5 + 8.2 + 7.8 + 9}{12} \\ &= \frac{115.8}{12} = 9.65 \text{ years}\end{aligned}$$

(b) From the given distribution we obtain the following table:

Size, x_i	5	6	7	8	9	10	11
No. of workers, f_i	8	10	15	13	8	6	5
$f_i x_i$	40	60	105	104	72	60	55

The mean size of a household is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_i = \frac{496}{65} = 7.63 \approx 8 \text{ members per worker}$$

2.11(a) The distribution below gives measurements on 40 different subjects.

Class interval	Class mark (x_i)	No. of subjects, frequency (f_i)	Deviation $d_i = (x_i - A)$	$f_i d_i$
110 – 119	114.5	1	-30	-30
120 – 129	124.5	3	-20	-60
130 – 139	134.5	7	-10	-70
140 – 149	144.5	14	0	0
150 – 159	154.5	8	10	80
160 – 169	164.5	5	20	100
170 – 179	174.5	2	30	60
<i>Total</i>	-	40	0	80

Using an assumed mean of $A = 144.5$, we compute the mean of the distribution as follows:

$$\bar{x} = A + \frac{1}{n} \sum_{i=1}^k f_i d_i = 144.5 + \frac{80}{40} = 146.5$$

2.11(b) A student's final end of semester examination marks in six courses are: 56, 68, 65, 70, 78 and 80. If the credits for the courses are 4, 3, 3, 4, 3 and 2 respectively, determine the approximate average mark.

- (b) Suppose a student went out and spent ₦150,000 as follows:

Item	Shirt	Book	Belt	Shoe	Bulb
Cost per item (GH¢)	18	8.5	6	25	2.5

Solution:

- (a) The approximate average mark is given by the weighted mean,

$$\begin{aligned}\bar{x}_w &= \frac{1}{n_w} \sum_{i=1}^6 w_i \\ &= \frac{4(56) + 3(68) + 3(65) + 4(70) + 3(78) + 2(80)}{4+3+3+4+3+2} \\ &= \frac{1297}{19} = 68.26\end{aligned}$$

- (b) On the average, the cost of an item is given by the arithmetic mean,

$$\begin{aligned}\bar{x} &= \frac{\text{total cost}}{\text{total number of items}} \\ &= \frac{18+8.5+6+25+2.5}{5} \\ &= \frac{60.0}{5} = \text{GH¢}12.00\end{aligned}$$

If instead of GH¢60.00, the student wishes to spend a total amount of GH¢200.00 as follows:

Item	Quantity	Cost/item (GH¢)	Total cost (GH¢)
Shirt	3	18.0	54.00
Book	9	8.5	76.50
Belt	2	6.0	12.00
Shoe	2	25	50.00
Bulb	3	2.5	7.50
<i>Total</i>	19	-	200.00

The average cost of one of the items purchased is given by the weighted mean,

$$\bar{x}_w = \frac{3(18.0) + 9(8.5) + 2(6.0) + 2(25.0) + 3(2.5)}{3+9+2+2+3}$$

$$= \frac{200.00}{19} = \text{GH¢}10.53$$

- (c) Consider the data below which give the number of hours of television viewing per week for a sample of 17 persons. Find the 10% trimmed mean of the given data.

14	9	12	4	20	26	17	15	18	10	6
16	15	8	5	23	11					

Solution

Arranging the data in order of magnitude we have,

4	5	6	8	9	10	11	12	14	15	15
16	17	18	20		23	26				

We find 10% of the total observations (17) which is approximately 2. We then remove the first two values from the extreme ends of the array data and compute the 10% trimmed mean as follows:

$$\begin{aligned} \text{10\% trimmed mean} &= \frac{6 + 8 + 9 + \dots + 18 + 20}{13} \\ &= \frac{161}{13} = 12.385 \end{aligned}$$

- (d) (i) Find the geometric and harmonic means of the following data: 25, 18, 15, 27, and 30.
- (ii) The distribution below gives the daily wages (in dollars) of 100 workers of a firm in a certain state. Compute the geometric and harmonic means of the distribution.

Daily wages	47-49	50-52	53-55	56-58	59-61	62-64
No. of workers	5	10	40	30	12	3

Solution

(i) The geometric mean,

$$g_m = \sqrt[5]{(25).(18).(15).(27).(30)} \\ = \sqrt[5]{5,467,500} \approx 22.26167, \text{ or}$$

$$\log(g_m) = \frac{1}{5} \left(\sum_{i=1}^5 \log x_i \right) \\ = \frac{1}{5} (15.51433203) = 3.102866974$$

Hence, $g_m = \text{anti log}(3.102866974) \approx 22.2617$

The harmonic mean,

$$h_m = \frac{5}{\frac{1}{25} + \frac{1}{18} + \frac{1}{15} + \frac{1}{27} + \frac{1}{30}} \\ = \frac{5}{0.232592593} \\ \approx 22.65978$$

(ii) For computations of geometric and harmonic means we obtain the following

table:

Daily wages, x_i	No. of workers, f_i	$\log x_i$	$f_i \log x_i$	f_i/x_i
48	5	3.87129	19.35645	0.10417
51	10	3.93183	39.31830	0.19608
54	40	3.98898	159.5592	0.74074
57	30	4.04305	121.2915	0.52632
60	12	4.09434	49.13208	0.20000
63	3	4.14313	12.42939	0.04762
Total	100	-	401.08692	1.81493

- $\log(g_m) = \frac{1}{100} \left(\sum_{i=1}^6 f_i \log x_i \right) = \frac{1}{100} (401.08692) = 4.0108692$

$$g_m = \text{anti log}(4.0108692) = 55.19482506$$

- $h_m = \frac{n}{\sum_{i=1}^k f_i/x_i} = \frac{100}{1.81493} = 55.09854375$

Example 2.11

- (a) Find the median of the following data:

(i) 5 4 2 9 7 6 21 11 13 18 10

(ii) 18 6 3 6 11 7 21 5 9 8 8 10

- (iii) The rating of job performance of workers in an establishment yielded the following results:

Job Performance	Excellent	Unsatisfactory	Very good	Average	Below Average	Good
No. of workers	15	15	66	20	45	40

Solution:

- (i) The number of observations in the given data, n is 11, which is odd. Arranging the data in increasing order of magnitude, the median is the $\frac{1}{2}(n+1)^{th} = 6^{th}$

observation: 2 4 5 6 7 $\boxed{9}$ 10 11 13 18 21


$$\text{Thus, } M = \frac{1}{2}(11+1)^{th} = 6^{th} \text{ observation} = 9$$

- (ii) The number of observations for a given set of data, $n = 12$. Arranging in order of magnitude:

3 6 6 8 8 $\boxed{9 \quad 11}$ 12 15 18 20 25


Hence the median is given by

$$M = \frac{1}{2} \left[\frac{n}{2}^{th} + \left(\frac{n}{2} + 1 \right)^{th} \right] \text{ observation}$$

$$= \frac{1}{2} (6^{th} \text{ observation} + 7^{th} \text{ observation}) = \frac{1}{2} (9 + 11) = 10$$

- (iii) We first arrange the performance in increasing order of rating

Job Performance	Unsatisfactory	Below Average	Average	Good	Very Good	Excellent
No. of workers	15	45	20	40	65	15
Cum.frequency	15	60	80	120	185	200

The median is the $\frac{1}{2}(200)^{th} = 100^{th}$ rating which falls in the *Good* category rating. Hence median of the rating is *Good*.

- (b) Compute the median for distribution below.

<i>Length (mm)</i>	<i>Frequency</i>	<i>Cum. Frequency</i>
118 – 126	3	3
127 – 135	5	8
136 – 144	9	17
145 – 153 ← median class	12 ← F_m	29
154 – 162	5	34
163 – 171	4	38
172 – 180	2	40
<i>Total</i>	40	-

Solution:

The middle observation is the 20th measurement which is located in the class interval 145 – 153 with class boundary 144.5 – 153.5. The median is obtained by interpolation as follows:

$$m = L_m + \left(\frac{\frac{n}{2} - F_m}{f_m} \right) C$$

$$= 144.5 + \left(\frac{40 - 17}{12} \right)(9) = 144.5 + \left(\frac{3}{12} \right)(9) = 144.5 + 2.25 = 164.75 \text{ mm}$$

- (c) Find the mode of the following set of data:

11, 11, 11, 12, 12, 13, 13, 12, 13, 17, 13,
 18, 13, 14, 14, 15, 14, 13, 16, 13, 21, 21,
 23, 13, 14, 13

Solution:

The mode of this set of data of 25 values is 13 because it is the most frequent occurring value. It occurs 9 times.

The mode of a distribution of data may not exist and even if it exists, it may not be unique. Consider the two data sets given below.

- (i) 10, 21, 33, 54, 40, 18, 53, 29, 8
 (ii) 3, 6, 9, 3, 10, 4, 6, 3, 1, 6, 2, 5, 6

There is no mode in the first set of data in (i) since all the observations are different. However, in the second set of data in (ii) there are two modes namely 3 and 6. They both occur four times and no other value occurs as often as that. The data is thus said to be *bimodal data*. A set of data which has a single mode is known as *unimodel data*.

- (d) Consider the distribution of 200 measurements of weights of an item observed at various locations.

<i>Class Boundary</i>	<i>Frequency</i>
3.67 – 3.79	3
3.79 – 3.91	9
3.91 – 4.03	28
4.03 – 4.15 (modal class)	54
4.15 – 4.27	51
4.27 – 4.39	31
4.39 – 4.51	17
4.51 – 4.75	7
<i>Total</i>	200

The most frequent class is 4.03 – 4.15 because it has the highest number of observations, 54. Thus 4.03 – 4.15 becomes the modal class. The mode is estimated as follows:

(i) By the *Crude method*, $m_0 = \frac{4.03 + 4.15}{2} = 4.09$

(ii) Using the formula, $m_0 = l_0 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) C$, where

$$l_0 = 4.03, \text{ lower limit of modal class}$$

$$\Delta_1 = 54 - 28 = 26, \Delta_2 = 54 - 57 = 3, \text{ and}$$

$$C = 4.15 - 4.03 = 0.12, \text{ class width of modal class,}$$

$$m_0 = 4.03 + \left(\frac{26}{26+3} \right) (0.12) = 4.03 + 0.11 = 4.15$$

2-2.2 Measures of Dispersion

When an average is used to describe a given set of data it tends to give a very misleading result unless it is identified and accompanied by supplementary information which indicates the amount of deviations of the various observations from the average. The degree to which the numerical data tend to spread about an average is the *dispersion or variation* of the data. Variation or dispersion is a very important characteristic of data. A measure of dispersion of a given set of data is important in two ways: It is used to show the degree of variation among the values in the given data. For example, a low dispersion of wages of workers indicates that workers are

approximately paid equal wages while a high dispersion gives an impression what workers are paid wages which are significantly different. It is also used to supplement an average description of data or to compare one group of data with another. When the dispersion is high, the average is of little or no significance but when it is low the value of the average becomes significant or highly representative. If the mean pulse rates of two patients in a hospital are the same but different in variability, the one with smaller variation may have a stable condition than the other one whose pulse rate fluctuates widely.

Numerous measures of dispersion exist, the most commonly being the *range*, *mean deviation*, *variance* (or *standard deviation*), *quartile deviation* and *coefficient of variation*.

2-2.2.1 The Range

The *range* is the simplest measure of dispersion. The range of set of measurements $x_1, x_2, x_3, \dots, x_n$ is defined as *the difference between the largest and smallest measurements*. In the case of grouped data, the range is defined as *the difference between the last and the first class marks*.

The range has the following properties:

- The range is easy and quicker is to compute and easily understood, as naturally, there is curiosity about the minimum and maximum values. It is very useful in stock market reports, which frequency give prices in terms of their ranges, quoting high and low prices over a time period. It is also often used in engineering and medical reports.
- It is affected by the one or two extreme values of the data and not very sensitive to the number of observations of the data.
- It is a very crude and generally, not a useful measure of variation. It does not tell anything about the dispersion of the values which fall between the two extreme values. It is used as “quick” and “easy” indication of variability. The range is widely used in Statistical Process Control (SPC) applications. It is used, for instance, in industrial quality control to keep a close check on raw materials or products by observing, and charting, the range of small samples taken regular intervals of time.

- The range is a rough estimate of dispersion and unsuitable for further statistical analysis. It supplements the mean description of data and not very sensitive to the number of observations of the data.

2-2.2.2 The Mean Deviation

The *mean deviation (MD)* is a measure of the average amount by which the observations, $x_1, x_2, x_3, \dots, x_n$, forming the data differ from the arithmetic mean, \bar{x} .

It is defined as follows:

- $MD = \frac{1}{n} \left(\sum_{i=1}^n |x_i - \bar{x}| \right)$, for ungrouped data, and
- $MD = \frac{1}{n} \left(\sum_{i=1}^n f_i |x_i - \bar{x}| \right)$, for grouped data.

The mean deviation has the following properties:

- The mean deviation is easily understood. It is a measure of dispersion which shows by how much, on average, each observation differs/deviates from the arithmetic mean.
- It is not greatly affected by extreme the observations. Its computation takes into account all the observed values.
- It is very useful in dealing with simple samples and situations where no elaborate analysis is required. It is unsuitable for further statistical analysis.

2-2.2.3 The Variance and Standard Deviation

The variance (or standard deviation) is the most preferred used measure of dispersion. The variance of a set of observations $x_1, x_2, x_3, \dots, x_n$ is the average of the squared deviations from the arithmetic mean. It is denoted by σ^2 and s^2 population and sample data respectively. That is,

- $\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2$ and $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$,

where μ and \bar{x} are the population and sample means respectively. The computation of sample variance (s^2) divides by $(n-1)$ instead of n to provide good estimator for the

population variance (σ^2), which will underestimate it. It is noted that for large sample size ($n > 30$) s^2 and σ^2 are approximately the same.

The *standard deviation* is defined as the positive root of the variance,

- $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$ or $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

The variances may be expanded, as shown below, for easier usage.

- $\sigma^2 = \frac{1}{N} \left(\sum_{i=1}^N (x_i - \mu)^2 \right) = \frac{1}{N} \left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 \right]$ Similarly,
- $s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]$

For grouped data,

- $\sigma^2 = \frac{1}{N} \sum_{i=1}^N f_i (x_i - \mu)^2 = \frac{1}{N} \left[\sum_{i=1}^N f_i x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N f_i x_i \right)^2 \right]$
- $s^2 = \frac{1}{n-1} \sum_{i=1}^n f_i (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n f_i x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n f_i x_i \right)^2 \right]$

where x_i is the class mark for the i^{th} class. If $d = (x_i - A)$ is the deviation of x_i from the assumed mean, A , then

- $s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n f_i d_i^2 - \frac{1}{n} \left(\sum_{i=1}^n f_i d_i \right)^2 \right]$

The following properties of the variance or standard deviation must be well-noted:

- It considers all the observations in the distribution. It has desirable properties which make it suitable for further statistical analysis.
- When each observation of the data is increased (or decreased) by affixed number, the standard deviation remains unchanged. However, when each observation is multiplied (or divided) by a fixed number the standard deviation is also multiplied (or divided) by that fixed number.

2-2.2.4 The Coefficient of Variation:

The standard deviation is useful as a measure of dispersion within a given set of data. Sometimes, we may be interested in comparing variations between two or more sets of data. The standard deviation or the variance can be used for this purpose when the variables are given in the same units and are such that their means are approximately equal. For instance, comparing the distributions of annual incomes and absenteeism for a group of employees, the number of defective articles produced in a batch and ages of workers involved in the production, weights of adults and babies or cholesterol levels (measured in milligrams per $100ml$) of persons and their weights (measured in pounds) is meaningless or difficult since they would distinctively be different. Obviously, it is impossible to compare directly the standard deviation of £500,000 for the annual incomes distribution with the standard deviation of 3.5 days for the distribution of absenteeism.

In order to make a meaningful comparison of the dispersion in incomes and absenteeism, we need to convert each of these standard deviations to a relative value. This relative measure of dispersion is called the *coefficient of variation* (CV). The coefficient of variation is defined as the ratio of the standard deviation to the arithmetic mean, usually expressed in percentage. That is, $CV = \frac{\text{standard deviation}(s)}{\text{mean}(\bar{x})} \times 100\%$

The coefficient of variation becomes a very useful measure of dispersion for comparing distributions of data when the data are in different units (such as dollars and days absent) or are in the same units but the means are far apart (such incomes of the top executives and incomes of the unskilled employees).

Example 2.11:

- (a) Consider the pulse rates of two patients in a hospital for ten different days:

Patient A: 77 76 70 69 70 69 75 78 70 71
Patient B: 59 92 60 80 71 65 50 88 95 70

Find the mean pulse rate for each patient and the corresponding range.

Solution:

The mean pulse rates for patients A and B are:

$$\bar{x}_A = \frac{77 + 76 + 70 + 69 + 70 + 69 + 75 + 78 + 70 + 71}{10} = \frac{725}{10} = 72.5$$

$$\bar{x}_B = \frac{59 + 92 + 60 + 80 + 71 + 65 + 50 + 88 + 95 + 70}{10} = \frac{730}{10} = 73$$

The ranges of data A and B are:

$$R_A = 78 - 69 = 9; R_B = 95 - 50 = 45$$

The two patients seem approximately have equal mean pulse rates over the ten days.

However, the pulse rates of patient B are more widely dispersed than patient A.

- (b) The ages of six HIV/AIDS patients in a hospital are 38, 26, 14, 41, 22 and 30 years. Find the mean deviation.

Solution:

The arithmetic mean and mean deviation are:

$$\bar{x} = \frac{38 + 26 + 14 + 41 + 22 + 30}{6} = \frac{171}{6} = 28.5$$

$$MD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{6} \left(\begin{aligned} & |38 - 28.5| + |26 - 28.5| + |14 - 28.5| + |41 - 28.5| \\ & + |22 - 28.5| + |30 - 28.5| \end{aligned} \right) \\ = \frac{1}{6} (9.5 + 2.5 + 14.5 + 12.5 + 6.5 + 1.5) = \frac{1}{6} (47) = 7.83,$$

which means that age distribution of the HIV/AIDS patients in the hospital deviates, on the average, by 7.83 years.

- (c) Consider the distribution of weight of 20 goats after feeding experiment.

Weight	Class Mark, x_i	Frequency, f_i	$f_i x_i$	$d_i = (x_i - \bar{x})$	$f_i x_i - \bar{x} $
42 – 47	44.5	3	133.5	-10.2	30.6
48 – 53	50..5	7	353.5	-4.2	29.4
54 – 59	56.5	5	282.5	1.8	9.0
60 – 65	62.5	3	187.5	7.8	23.4
66 – 71	68.5	2	137.0	13.8	27.6
Total	-	20	1094.0	-	120.0

The arithmetic mean and mean deviation from the table are as follows:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n f_i x_i = \frac{1094.0}{20} = 54.70 \text{ kg}$$

$$MD = \frac{\sum_{i=1}^k f_i |x_i - \bar{x}|}{n} = \frac{120}{20} = 6 \text{ kg}$$

Example 2.12:

- (a) (i) The prices (in dollars) of a certain commodity on eight different sessions were 38, 11, 8, 60, 52, 68, 32 and 19. Find the variance and the standard deviation of the data, where $\sum_{i=1}^n x_i = 288$, and $\sum_{i=1}^n x_i^2 = 13,942$.
- (ii) Compute the variance and standard deviation for the grouped sample data:

Mark	Class Mark, x_i	Frequency, f_i	$f_i x_i$	$f_i (x_i - \bar{x})^2$	$f_i x_i^2$
10 – 19	14.5	5	72.5	2,493.14	1,051.25
20 – 29	24..5	20	490	3,040.58	12,005.0
30 – 39	34.5	10	345	54.29	11,902.5
40 – 49	44.5	14	623	823.60	27,723.5
50 – 59	54.5	5	270	1,561.14	14,851.25
60 – 69	64.5	4	258	3,062.52	16,641.0
70 – 79	74.5	2	149	2,838.06	11,100.5
<i>Total</i>	-	60	2,210	13,873.33	95,275.0

Solution:

- (i) The mean price of the commodity and its variance are as computed below:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^x x_i = \frac{288}{8} = \$36.0$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^x (x_i - \bar{x})^2 = \frac{1}{7} (3574) = 510.57143, \text{ or}$$

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^x x_i^2 - \frac{1}{n} \left(\sum_{i=1}^x x_i \right)^2 \right] = \frac{1}{7} \left[13,942 - \frac{1}{8} (288)^2 \right] = 510.57143$$

Hence the standard deviation, $s = \sqrt{510.57} = 22.59583$

- (ii) The sample mean and variance the of the distribution:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_i = \frac{2210}{60} = 36.83$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (x_i - \bar{x})^2 = \frac{13,873.33}{59} = 235.14, \text{ or}$$

$$s^2 = \frac{1}{n-1} \left[\sum f_i x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] = \frac{1}{59} \left[95,275 - \frac{1}{60} (2210)^2 \right] = 235.14$$

Hence the standard deviation, $s = \sqrt{235.14} = 15.33$

- (b) Consider the distribution of weights of 20 goats after feeding experiment in **Example 2.11(c)** and assume a mean of 56.5 ($A = 56.5$), compute the mean and the standard deviation, where $d_i = x_i - A$, and the sum of $f_i d_i$ and $f_i d_i^2$ are -36 and 1,080 respectively.

The arithmetic mean and standard and deviation:

$$\begin{aligned} \bar{x} &= A + \sum_{i=1}^k f_i d_i \\ &= 56.5 + \frac{-36}{20} = 54.7 \text{ kg} \end{aligned}$$

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^k f_i d_i^2 - \frac{1}{n} \left(\sum_{i=1}^k f_i d_i \right)^2 \right]} \\ &= \sqrt{\frac{1}{19} \left[1080 - \frac{1}{20} (-36)^2 \right]} = 7.53 \text{ kg} \end{aligned}$$

- (c) A student was asked to compute the mean and standard deviation of a random sample of ten numbers and found the results to be 20 and 15 respectively. He however, realized that two of the numbers, 21 and 13, were wrongly recorded as 12 and 31. Use the correct numbers to obtain the correct values of the mean and standard deviation.

Solution:

The wrong sum of the numbers, $\sum_{i=1}^{10} x_i = 20(10) = 200$, and the correct sum of numbers,

$$\sum_{i=1}^{10} x_i = 200 - (12 + 31) + (21 + 13) = 191$$

The wrong sum of squares of the numbers, $\sum_{i=1}^n x_i^2 = 15^2(9) + \frac{1}{10}(200)^2 = 6025$. The

correct sum of squares of numbers, $\sum_{i=1}^{10} x_i^2 = 6025 - (12^2 + 31^2) + (21^2 + 13^2) = 5530$

Now the correct mean and standard deviation are respectively

$$\bar{x} = \frac{\text{correct sum}}{10} = \frac{191}{10} = 19.1$$

$$s = \sqrt{\frac{1}{n-1} \left[\text{correct sum of squares} - \frac{(\text{correct sum})^2}{n} \right]} = \sqrt{\frac{1}{9} \left[5,530 - \frac{(191)^2}{10} \right]} = 14.46$$

- (d) A set of examination marks has a mean of 35 and a standard deviation of 3. The marks are to be scaled so that the mean becomes 45 and standard deviation, 6. If the equation of the transformation is $y_i = ax_i + b$, find the values of the constants, a and b . Find also the scaled mark which corresponds to the mark of 40 in the original set of data.

Solution:

Let marks be x_1, x_2, \dots, x_n , and the scaled marks, y_1, y_2, \dots, y_n , where $y_i = ax_i + b$

$$\sum_{i=1}^x y_i = a \sum_{i=1}^x x_i + nb \Leftrightarrow \bar{y} = a\bar{x} + b \quad \dots \dots \dots \quad (2.1)$$

The variance of y_i , $i = 1, 2, \dots, n$.

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (ax_i + b - (a\bar{x} + b))^2, \text{ from (2.1)}$$

$$\sigma_y^2 = \frac{a^2}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = a^2 \sigma_x^2, \text{ from which we have}$$

$$\sigma_y = a \sigma_x, \quad \dots \dots \dots \quad (2.2)$$

Hence given $\bar{x} = 35$, $\sigma_x = 3$, $\bar{y} = 45$, and $\sigma_y = 6$ we substitute into equations (2.1) and (2.2) and obtain,

$$45 = 5a + b \quad \dots \dots \dots \quad (2.3)$$

$$6 = 3b \quad \dots \dots \dots \quad (2.4)$$

Solving equations (2.3) and (2.4) simultaneously, we have $a = 2$ and $b = -25$.

The transformation equation then becomes, $y_i = 2x_i - 25$, and a mark of 40 is therefore scaled to $y = 2(40) - 25 = 55$.

Example 2.13:

- (a) During the past few months, one runner averaged 12 miles per week with a standard deviation of 2 miles, while another runner averaged 24 miles per week with a standard deviation of 3 miles. Which of the two runners is relatively more consistent in his weekly running habits?
- (b) The variation in the annual incomes of executives is to be compared with the variation in incomes of unskilled employees. For a sample of executives, the mean income is \$500,000 with standard deviation of \$50,000 while that of the unskilled employees have a mean of \$22,000 with standard deviation, \$2,200. Compute the coefficients of variation for a meaningful comparison of variation in annual incomes.
- (b) In order to choose two measuring instruments A and B, each was used to measure the diameter of 20 coins. Instrument A give the following measurements in centimetres.

3.35	3.54	3.55	3.53	3.52	3.54	3.57	3.54	3.56	3.54	3.53
3.55	3.52	3.53	3.55	3.54	3.53	3.55	3.56	3.55		

- (i) Using an assumed mean of 3.54 cm, calculate the mean, standard deviation and coefficient of variation of the measurements.
- (ii) If instrument B gave the same mean as A but its standard deviation was 0.010745cm which of the two instruments is better? Give reasons.
- (iii) What would be the mean and the standard deviation if each of the above measurements is decreased by 0.25cm?

Solution:

- (a) Computing the coefficients of variation for the two runners,

$$\text{Runner I : } CV_1 = \frac{s_1}{\bar{x}_1} = \frac{2}{12} = 0.167(16.7\%)$$

$$\text{Runner II : } CV_2 = \frac{s_2}{\bar{x}_2} = \frac{3}{24} = 0.125(12.5\%)$$

The second runner is relatively more consistent in his weekly running habits since its CV is quite less than the first.

- (b) We are tempted to say that there is more dispersion in the annual incomes of the executives because \$500,000 is greater than \$22,000. The coefficients of variation are computed to make a meaningful comparison of variation in annual incomes.

$$\text{The } CV \text{ for incomes of executives} = \frac{50,000}{100,000} = 0.10(10\%)$$

$$\text{The } CV \text{ for incomes of unskilled employees} = \frac{2,200}{22,000} = 0.10(10\%),$$

from which we conclude that there is no difference in the relative dispersion of the two groups.

- (c) Given the assumed mean, $A = 3.54$, we have the table:

Class Mark, x_i	Frequency, f_i	$A = 3.54$ $d_i = x_i - A$	$f_i d_i$	$f_i d_i^2$
3.51	1	-0.03	2,493.14	1,051.25
3.52	2	-0.02	3,040.58	12,005.0
3.53	5	-0.01	54.29	11,902.5
3.54	5	0.00	823.60	27,723.5
3.55	5	0.01	1,561.14	14,851.25
3.56	2	0.02	3,062.52	16,641.0
-	60	-	-0.03	0.0035

- (i) The arithmetic mean, standard deviation and coefficient of variation of instrument A are, respectively

$$\bar{x}_A = A + \frac{1}{n} \sum f d = 3.54 + \frac{-0.03}{20} = 3.5385$$

$$s_A = \sqrt{\frac{1}{n-1} \left[\sum f d^2 - \frac{1}{n} (\sum f d)^2 \right]}$$

$$= \sqrt{\frac{1}{19} \left[0.0035 - \frac{1}{20} (-0.03)^2 \right]} = 0.013485$$

$$CV_A = \frac{s \text{ standard deviation}}{\text{mean}}$$

$$= \frac{0.013485}{3.5385} = 0.0037615$$

- (ii) If instrument B has same mean as A with standard deviation, 0.010745, then instrument B gives better diameter measurements for coins since its is less variable than instrument A .
- (iii) If all the measurements are decreased by 0.25, the mean also decreases by 0.25 cm but the standard deviation remains unchanged.

2-2.3 Measures of Position and Shape

2-2.3.1 Measures of Position

Suppose there are agitations among workers in an establishment that they are drastically underpaid compared with other people with similar experience and performance. One way to tackle the problem is to obtain the salaries of these other workers and demonstrate that, comparatively, the salaries of these aggrieved workers are indeed on a low side.

To evaluate the salaries of the workers compared with the entire workers, we would use a measure of position. Measures of position are indicators of how a particular value fits in with all the other data values. They actually determine the relative position of a particular observation, indicating what fraction of the whole set of data is below/above this particular observation. The most commonly used measures of position are the quartiles, deciles, and percentiles.

- *Quartiles:* The quartiles divide a set of data into quartiles such that the first or lower quartile (Q_1) has 25% of the observations falling below it, the second or middle quartile (Q_2), popularly known as the median, has 50% of the observations falling below/above it and the third or upper *quartile* (Q_3) has

75% of the observations falling below it when the data is arranged in order of magnitude. The difference between Q_3 and Q_1 is called *inter-quartile range (IQR)* (that is, $IQR = Q_3 - Q_1$, which is a measure of dispersion. The *semi inter-quartile range or quartile deviation (QD)* is defined by, $OD = \frac{1}{2}(Q_3 - Q_1)$.

$$Q_1 = l_1 + \left(\frac{\frac{n}{4} - F_1}{f_1} \right) c_1,$$

l_1 = lower class boundary of the first quartile class, F_1 = cumulative frequency just before the first quartile class, f_1 = frequency of the first quartile class, c_1 = class width of median class boundary and n =total number of observations(total frequency)

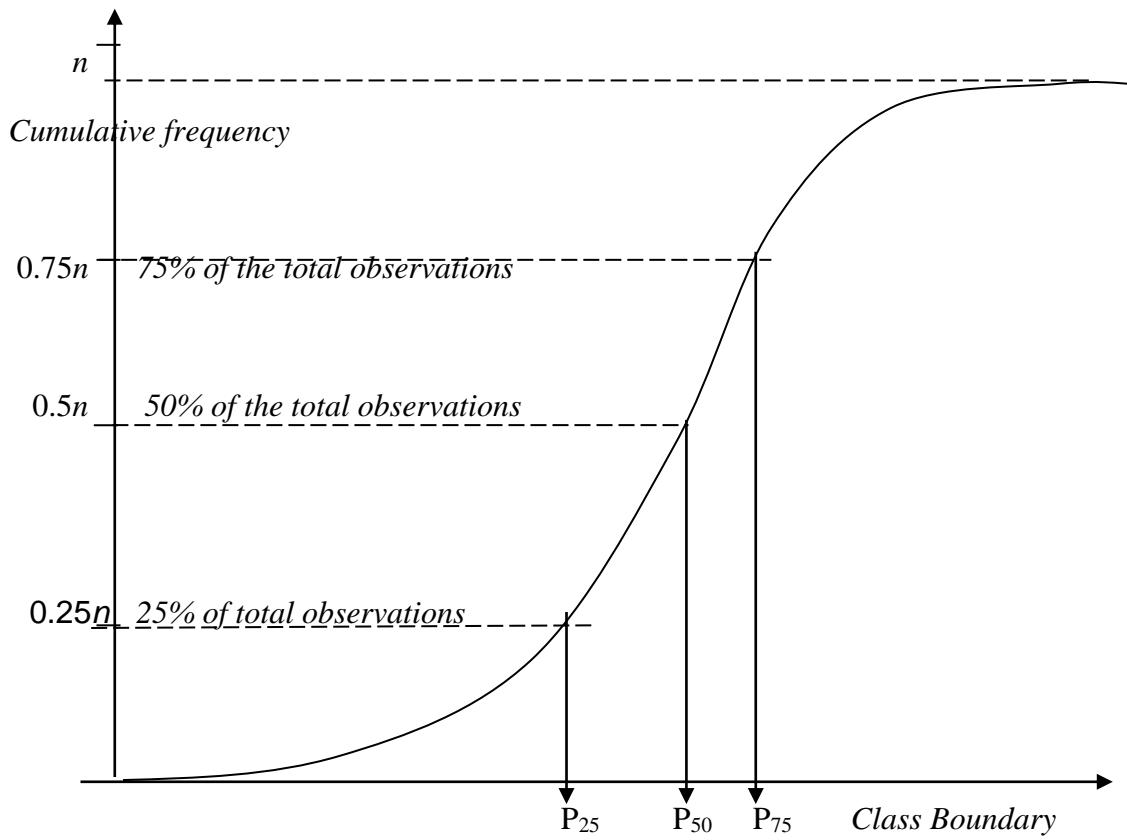
$$Q_3 = l_3 + \left(\frac{\frac{3n}{4} - F_3}{f_3} \right) c_3$$

l_3 = lower class boundary of the third quartile class, F_3 = cumulative frequency just before the third quartile class, f_3 = frequency of the first quartile class, c_3 = class width of median class boundary and n =total number of observations(total frequency)

- *Deciles:* The deciles are the values that divide the set of data into ten equal parts. They are denoted by $D_1, D_2, D_3, \dots, D_9$ and are such that 10%, 20%, 30% . . . 90% of the observations fall below $D_1, D_2, D_3, \dots, D_9$ respectively.
- *Percentiles:* The percentiles divide the data into 100 equal parts. They are denoted by $P_1, P_2, P_3, \dots, P_{99}$ and are such that 1%, 2% . . . 99% of the observations fall below $P_1, P_2, P_3, \dots, P_{99}$ respectively. They are used when dealing with large amount of data. The 25th percentile (P_{25}) is equal to Q_1 , 50th percentile (P_{50}) is the median(Q_2 or M) and 75th percentile (P_{75}) is equal to Q_3 .

Measures of position for grouped data can be determined by the *interpolation method* and *use of the cumulative frequency curve*.

- *The Interpolation Method:* The k^{th} percentile for a grouped data is determined by the formula, $P_k = l_k + \frac{c_k}{f_k} (nk - F_k)$, where l_k = the lower limit of the class in which the k^{th} percentile lies, F_k = the cumulative frequency just before the k^{th} percentile class boundary, f_k = the frequency of the k^{th} percentile class boundary, and c_k = the class width of the k^{th} percentile class boundary.
- *The Use Cumulative Frequency Curve:* The quartiles, deciles or percentiles may also be determined from cumulative frequency curve as illustrated below.
- *The Use Cumulative Frequency Curve:* The quartiles, deciles or percentiles may also be determined from cumulative frequency curve as illustrated below.



2-2.3.2 Measures of Shape

The shape of a frequency distribution of n data observations, $x_1, x_2, x_3, \dots, x_n$, represented graphically by a histogram/frequency polygon can be described using various measures of shape. Measures of shape determine whether the distribution of data exhibits a symmetric pattern or stretch out in a particular direction. Two of such measures of shape are the *skewness* and *kurtosis*.

- *Skewness:* The skewness of a distribution indicates its degree of symmetry or non-symmetry. It is measured by the *Pearson Coefficient of Skewness* (S_k), defined by $S_k = \frac{3(\text{mean} - \text{median})}{s \tan \text{dard deviation}} = \frac{3(\bar{x} - m)}{s}$, which ranges from -3 to 3 .
 3. If $S_k = 0$, $(\bar{x} = m)$ and the distribution is said to be *symmetric*.

If $S_k > 0$, $(\bar{x} > m)$ and the distribution is said to be skewed to *the right* or *positively skewed*.

If $S_k < 0$, $(\bar{x} < m)$ and the distribution is said to be *skewed to the left* or *negatively skewed*.

The graph below gives the shapes of the symmetrical and two skewed distributions for various values of S . Also, it shows the relationship among the arithmetic mean (\bar{x}), median (M) and mode (M_o).

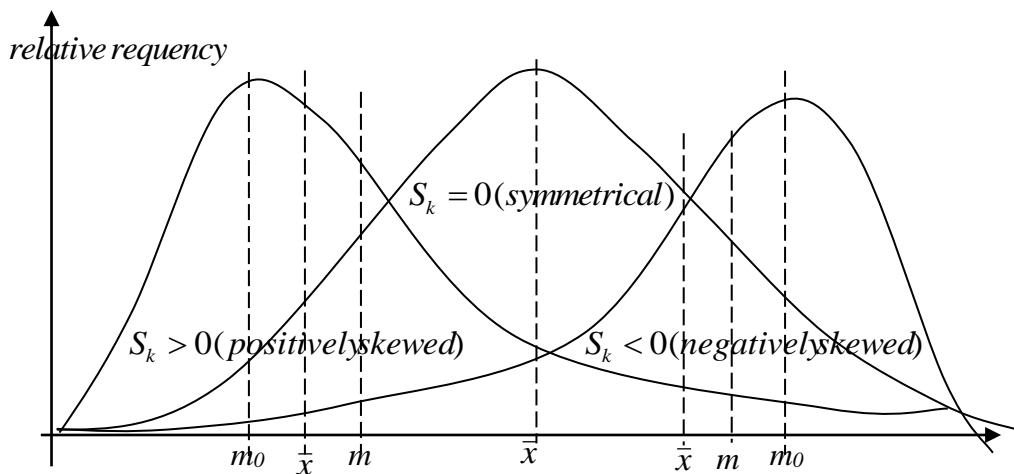


Figure 2.1: Graph of Symmetrical and Non-Symmetrical Distributions

From the relationship it is established that

$$\bar{x} - m_0 = 3(\bar{x} - m_0) \text{ or } m_0 = \bar{x} - 3(\bar{x} - m_0)$$

- *Peakness:* The degree of *peakness* or *kurtosis* of a distribution is described by the coefficient of kurtosis, k defined by $k = \frac{1/2(Q_3 - Q_1)}{P_{90} - P_{10}}$, and compared with 3.

If the value of $k = 3$, the distribution is said to be *symmetrical* or *normal*. If $k < 3$, The distribution flattens at the centre than the normal distribution (the individual observations scatter widely about the mean). If $k > 3$, the distribution is more peaked than the normal distribution (the observations are closer to the mean). These are illustrated as below.

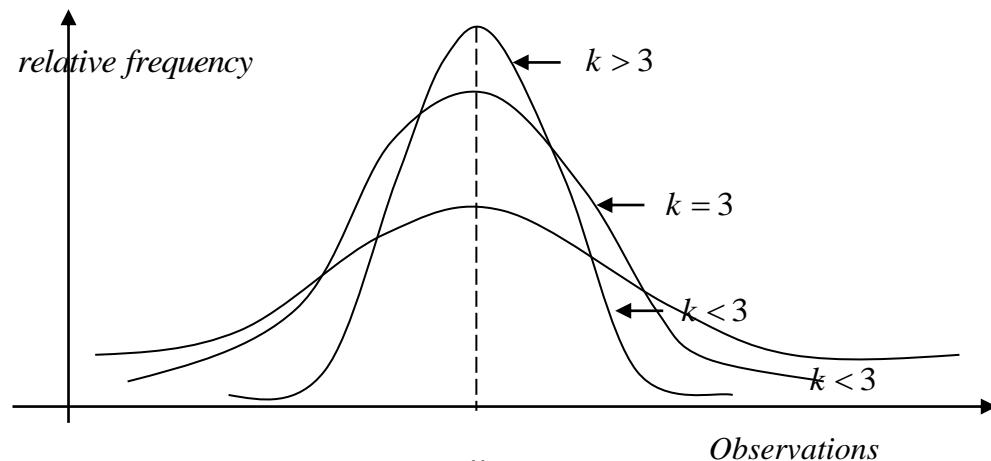


Figure 2.2: Graphs of Distributions indicating their Peakness

Example 2.14

- (a) The lengths of stay on the cancer floor of a hospital were organised into a frequency distribution. The mean length of stay was 28 days, the median, 25 days and mode, 23 days. The standard deviation was computed to be 4.2 days.
- (i) The distribution is positively skewed because the mean is the largest of the three measures of central tendency.
- (ii) The coefficient of skewness which generally lies between is computed

$$\text{as } S_k = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3(28 - 25)}{4.2} = 2.14,$$

which indicates that a substantial amount of positive skewness. Apparently, a few cancer patients are staying in the hospital for a long time, causing the mean to be larger than the mean or mode.

- (b) The data below show the age distribution of cases of malaria reported during year at a hospital.

34	17	25	37	19	19	27	19	44	24	24
22	32	12	13	16	18	14	12	16	14	17
10	16	22	20	15	15	10	10	14	17	20
18	13	32	13	13	18	30	24	34	44	31
43	40	28	31	15	22	15	31	18	27	35
35	20	32	38	32						

- (i) Organize the data into a frequency distribution table.
(ii) Calculate and interpret the coefficient of skewness, and kurtosis.

Solution:

- (i) The required frequency distribution is

Age	Class Mark, x_i	Frequency, f_i	$f_i x_i$	$f_i x_i^2$
9.5-14.5	12	11	132	1,051.25
14.5-19.5	17	19	223	12,005.0
19.5-24.5	22	9	198	11,902.5
24.5-29.5	27	4	108	27,723.5
29.5-34.5	32	9	288	14,851.25
34.5-39.5	37	4	148	16,641.0
39.5-44.5	42	4	168	
Total	-	60	1,365	36,095

- (ii) The arithmetic mean and variance are:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_i = \frac{1365}{60} = 22.75 \text{ years}$$

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^k f_i x_i^2 - \frac{1}{n} \left(\sum_{i=1}^k f_i x_i \right)^2 \right] = \frac{1}{59} \left[36,095 - \frac{1}{60} (1,365)^2 \right] = (9.244)^2$$

Hence the standard deviation, $s = 9.244$ years

- (iii) The median, lower and upper quartiles of the distribution are, respectively of the distribution,

$$m = l_m + \frac{c_m}{f_m} \left(\frac{n}{2} - F_m \right) = 19.5 + \frac{5}{9} (30 - 30) = 19.5 \text{ years} = 14.5 + \frac{5}{19} (30 - 11)$$

$$Q_1 = l_1 + \frac{c_1}{f_1} \left(\frac{n}{4} - F_1 \right)_1 = 14.5 + \frac{5}{19} (15 - 11) = 15.55 \text{ years}$$

$$Q_3 = l_3 + \frac{c_3}{f_3} \left(\frac{3n}{4} - F_3 \right) = 29.5 + \frac{5}{9} (45 - 43) = 30.61 \text{ years}$$

- (iv) The 10th and 90th percentiles are, respectively

$$P_{10} = l_{10} + \frac{c_{10}}{f_{10}} (0.1n - F_{10}) = 9.5 + \frac{5}{12} (6 - 0) = 12 \text{ years}$$

$$P_{90} = l_{90} + \frac{c_{90}}{f_{90}} (0.9n - F_{90}) = 34.5 + \frac{5}{4} (54 - 52) = 37 \text{ years}$$

- (v) The coefficients of skewness and of kurtosis:

$$S_k = \frac{3(\bar{x} - m)}{s} = \frac{3(22.75 - 19.5)}{9.244} = 1.05474 > 0$$

indicating that the age distribution of cases of the disease is positively skewed.

$$k = \frac{\frac{1}{2}(Q_3 - Q_1)}{P_{90} - P_{10}} = \frac{\frac{1}{2}(39.61 - 15.55)}{37 - 10.05} = 0.44638 < 3,$$

which means that the distribution slightly flattens at the centre than the normal distribution (i.e. the data is scattered about the mean).

2-2.4 Trial Questions 2-2:

- 2.2(a)** (i) Classify methods of statistical analysis and briefly explain the situations under which each may be employed.
- (ii) State the need of studying Statistics in your programme of study. What significant role does a statistician play in a scientific investigation?
- (iii) State the stages involved in a statistical investigation. Discuss the methods of data collection.
- (iv) Distinguish between standard deviation and coefficient of variation. State two desirable properties of each of these measures of dispersion.

- 2.2(b)** The width of electronic components from a production process was measured by an instrument. The results obtained (in mm) are as shown by the data below.

8.1	4.9	3.5	5.2	6.0	2.8	7.1	4.3	7.2	7.9	7.5	7.7
6.4	6.7	8.5	6.6	3.9	8.2	6.3	5.8	7.3	9.4	5.1	8.5
8.3	5.2	4.9	7.0	7.4	5.5	4.6	7.3	6.6	6.2	7.9	7.3
8.7	6.6	6.0	3.7	6.3	9.0	7.2	7.1	4.4	8.0	8.6	6.1
6.0	7.3	2.0	6.3	7.5	6.9	5.0	4.8	5.3	4.1	7.2	5.2

- (i) Estimate the number of classes and class width using the Surges' Rule and corrected to one decimal place
- (ii) Construct a grouped frequency distribution for the given data.
- (iii) Compute the mean, median, mode and the standard deviation.
- (iv) Compute the coefficients of skewness and kurtosis

- 2.2(c)** The following table gives the frequency distribution of average weekly expenditure (in thousands of Cedis) of a random sample 100 students at KNUST.

<i>Expenditure (x)</i>	58 – 62	63 - 67	68 - 72	73 – 77	78 – 82
<i>Frequency (f)</i>	15	<i>m</i>	10	<i>n</i>	10
<i>fx</i>	900	<i>65m</i>	700	<i>75n</i>	800

- (i) Assuming that the mean is ₦70,750.00, find the values of *m* and *n*.
- (ii) Compute the standard deviation, given that the sum of the data is 54,625.
- (iii) Compute the coefficient of variation of the data.

- 2.2(d)** The frequency distributions below show the weekly sales of cars for two companies during a current year.

Company A		Company B	
No. of sales	frequency	No. of sales	Frequency
0	6	0	10
1	25	1	11
2	15	2	12
3	5	3	9
4	1	4	8
5	0	5	2

- (i) What are the most frequent weekly sales made by companies A and B?
 - (ii) What are the mean number of cars sold by companies A and B?
 - (iii) What are the standard deviation for the number of sales made by companies A and B?
 - (iv) Compute the coefficient of variations for the number of sales made by companies A and B.
 - (v) Use the above results to describe the two distributions.
- 2.2(e)** A study of the test scores for a course in Principles of Management and years of service of the employees enrolled in a Business programme resulted in a mean score of 200 with standard deviation, 40 and mean number of years of service of 20 with standard deviation of 2. Compare the relative dispersion in the two distributions using the coefficient of variation.

- 2.2(f)** The annual income for randomly selected secretaries is recorded in thousands of dollars as follows:

10.3	9.8	10.1	13.2	15.4	10.0	13.6	12.2	9.7	12.6	10.2
11.0	16.4	8.9	9.4	8.9	10.6	10.4	10.9	10.5	12.1	1.8
8.1	12.0	13.2	12.2	12.4	14.5	10.5	9.7			

- (i) Obtain a grouped frequency distribution for the data using the Sturges Rule
- (ii) Compute the coefficient of variation of annual income of the secretaries.
- (iii) Compute the coefficients of skewness and kurtosis.

INTRODUCTION TO PROBABILITY THEORY

Probability Theory is an important area in Mathematics that is concerned with random (or chance) phenomenon. The study of probability theory has attracted much attention because of its intrinsic interest and successful applications to many areas within the physical, biological, social sciences, engineering and in the business world. It is the underlying foundation on which the important methods of inferential statistics are built. In this unit we introduce the concepts of probability of an event including that of several events and then show how these probabilities can be computed under such situations.

Learning Objectives

This unit aims to develop a basic understanding of Probability Theory and skills necessary to compute probabilities in a variety of important circumstances. This means that after studying the unit, students will be able to:

- Understand the probability measure and how it is determined.
- Explain the basic terms and also state the basic rules or theorems of probability.
- Apply the rules/theorems to compute probabilities of events.
- Apply the counting techniques to compute the probability of events.

SESSION 1-3: GENERAL CONCEPTS

1-3.1 Introduction

Life, they say, is full of uncertainties. This may be observed in several situations some of which are in a forecast of the weather, measurement of the blood pressure of a patient, a football coach's assessment of the chances of his/her team winning a match, a driver pondering over being caught for parking illegally and a civil engineer pondering over the exact load a bridge can endure before collapsing, where in each case we see an element of uncertainty. This is because the weather is often wrongly forecasted, we cannot tell the blood pressure level of the patient at that particular point in time, the coach knows that there is no such a thing as sure win, the driver knows that drivers are often caught when they parked illegally, and bridges often collapse for exerting too

much load upon them. Most often, in our daily lives, we would like to measure the likelihood of an outcome of an event or activity. This can be determined by performing an experiment. However, the outcomes of some experiments are *random* (or *not predictable*). A toss of a *fair coin* or *die* and the experiments quoted above, are all examples of random experiments.

The *probability* of an event is measure of belief that (or how likely) the event will occur. The concept of Probability Theory becomes necessary when dealing with physical, biological and social mechanisms that generate observations which cannot be predicted with certainty. The theory of probability can be thought of as that branch of Mathematics that is concerned with calculating the probabilities of outcomes of random experiments or uncertain events. It originated from games of chance (gambling) and has now become important tool due to wide variety of practical problems it solves and the role it plays in Science. It is also the basis of statistical analysis of data, which is widely used in industry and in experimentation.

1-3.2 Applications of Probability Theory

The study of uncertainties has been found to have wide applications in the following situations:

- It is used in industries to determine the reliability of certain equipment.
- The government uses it to determine fiscal and economic policies. Economists use it in predicting the rise or decline of an inflation rate.
- It is used by biologists in the study of genetics.
- It is used by the insurance companies in the calculations of insurance premiums and the probable life expectancies of their policy holders.
- It is used by business managers in determining which products to manufacture, which products to advertise and through which medium: TV, radio, magazine, newspaper, and etc. advertisements.
- The quality control in manufacturing product development decisions are based on probability theory.
- In survival analysis, a patient can be assessed his/her chance of survival after he/she has been diagnosed of a disease.
- It is used by investors to decide which particular stock has greater chance for future growth than any other stock.

1-3.3 Terminologies and Notations

- *Experiment:* An experiment is any process that generates well-defined outcomes. There are two types of experiments, namely *deterministic* and *random* (or *chance*) *experiment*. In the deterministic experiments the observed results are not subject to chance while the outcomes of random experiments cannot be predicted with certainty. A random experiment could be simple as tossing a coin or die and observing an outcome or as complex as selecting 50 students from KNUST campus testing them for the HIV/AIDS disease.
- *Trial:* A trial is a single performance of an experiment (that is, a repetition of an experiment).
- *Outcome:* The possible result of each trial of an experiment is called outcome. When an outcome of an experiment has equal chance of occurring as the others the outcomes are said to be *equally likely*. For example, the toss of a coin and a die yield the possible outcomes in the sets, $\{H, T\}$ and $\{1, 2, 3, 4, 5, 6\}$ and a play of a football match yields $\{\text{win}(W), \text{loss}(L), \text{draw}(D)\}$
- *Sample Space:* It is the set of all possible outcomes of an experiment. The letter, S , denotes it. Each element or outcome of the experiment is called *sample point* or *outcome*. For example,

- (i) The results of two and three tosses of a coin give the sample spaces:

$$S = \{HH, HT, TH, TT\}$$

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

- (ii) The outcomes of two tosses of a die are as shown in the *Table 3.1*.

<i>First Toss</i>	<i>Second Toss:</i>					
	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Table 3.1: Outcomes of two Dice

- (iii) A toss of a die and coin simultaneously give the results,
- $$S = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$$
- (iv) Drawing a card from a packet of playing cards has a sample space with 52 cards, made up *13 Heart*, *13 Spade*, *13 Diamond* and *13 Club* cards.

- *Event:* An event is a collection of one or more outcomes from an experiment which is a subset of a sample space. It is denoted by a capital letter. For example we may have:
 - The event of observing exactly two heads (H 's) in three tosses of a coin,
 $A = \{HHT, HTH, THH\}$
 - The event of obtaining a total score of 8 on two tosses of a die,
 $B = \{(2,6), (3,5), (4,4), (5,3), (6,2)\}$
 - A newly married couple planning to have three children. The event of the family having two girls is $D = \{BGG, GBG, GGB\}$
- *Tree Diagram:* The tree diagram represents pictorially the outcomes of random experiment. The probability of an outcome which is a sequence of trials is represented by any path of the tree. For example,
 - A couple planning to have three children, assuming each child born is equally likely to be a boy (B) or girl (G). Figure 3.1 gives the tree diagram of the three births.

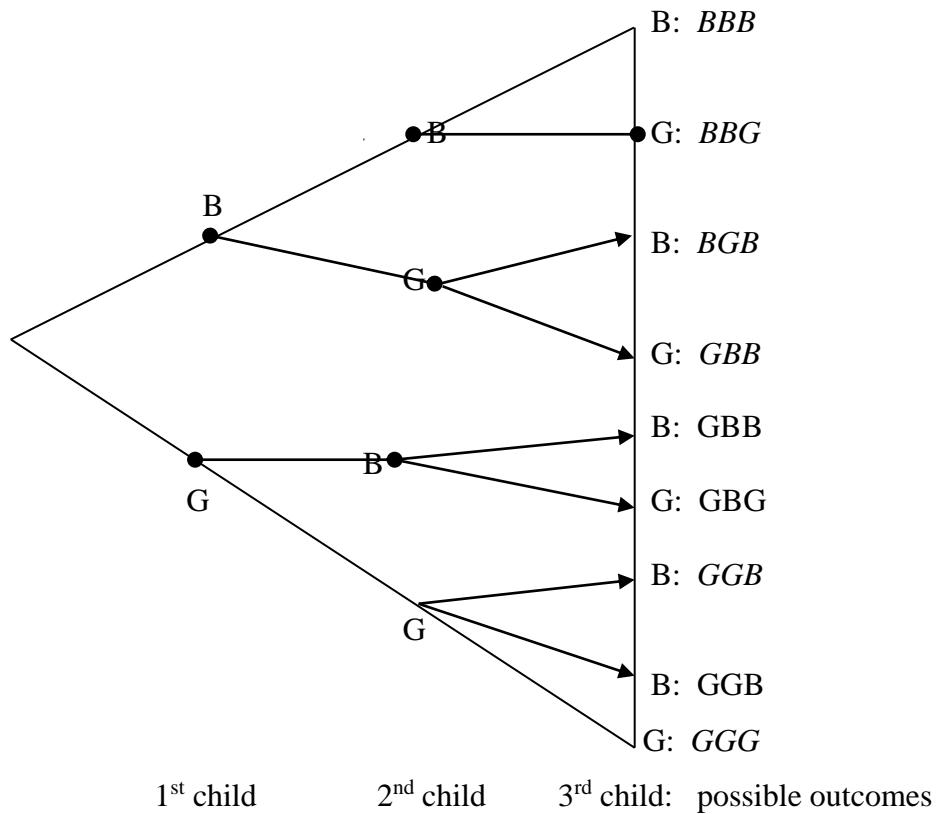


Figure 3.1: Tree Diagram of 3 Births

- (ii) A soccer team on winning (W_T) or losing (L_T) a toss can defend either post A or B . It plays the match and either win (W), draw (D) or lose (L). We illustrate the experiment on a diagram in Figure 3.2 as follows:

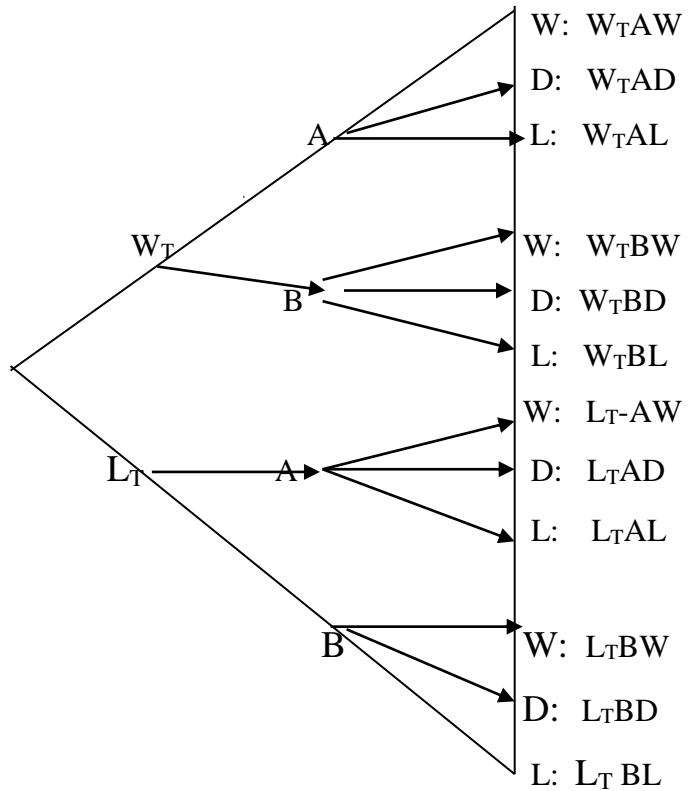


Figure 3.2: Outcomes of Soccer Match

SESSION 2-3: DETERMINATION OF PROBABILITY OF EVENTS

2-3.1 Introduction

The probability of an event A , denoted, $P(A)$, gives the numerical measure of the likelihood of the occurrence of event A which is such that, $0 \leq P(A) \leq 1$. If $P(A) = 0$, the event A is said to impossible to occur, if $P(A) = 1$, A is said to be certain and if $P(A) = 0.5$, the event is just as likely to occur as not. If \bar{A} is the complement of the event A , then $P(\bar{A}) = 1 - P(A)$, the probability that event A will not occur.

There are three main schools of thought in defining and interpreting the probability of an event. These are the *Classical Definition*, *Empirical Concept* and the *Subjective Approach*. The first two are referred to as the *Objective Approach*.

2-3.2 The Classical Definition

This is based on the assumption that all the possible outcomes of the experiment are equally likely. For example, if an experiment can lead to n mutually exclusive and equally likely outcomes, then the probability of the event A is defined by

$$P(A) = \frac{n(A)}{n(S)} = \frac{\text{number of ways } A \text{ can occur}}{\text{number of ways the expt. can proceed}} = \frac{\text{number of successful outcomes}}{\text{number of possible outcomes}}$$

The classical definition of probability of event A is referred to as *priori probability* because it is determined before any experiment is performed to observe the outcomes of event A .

2-3.3 The Empirical Concept

This concept uses the relative frequencies of past occurrences to develop probabilities for future. The probability of an event A happening in future is determined by observing what fraction of the time similar events happened in the past. That is,

$$P(A) = \frac{\text{number of times event } A \text{ occurred in the past}}{\text{total number of observations}}$$

The relative frequency of the occurrence of the event A used to estimate $P(A)$ becomes more accurate if the trials are largely repeated. The relative frequency approach of defining $P(A)$ is sometimes called *posterior probability* because $P(A)$ is determined only after event A is observed. For example, in studying the peak demand at a power plant, an electrical engineer observed that on 85 of 120 days randomly selected from past records, the peak demand occurred between 6.30 and 8.00pm. The probability of this occurring is $\frac{85}{120} = 0.708$, which is based on repeated experimentation and observation.

2-3.4 The Subjective Definition

The subjective concept of probability is based on the degree of belief (personal opinion) through the evidence available. The probability of an event A may therefore be assessed through experience, intuitive judgment or expertise. For example, determining the probability of containing an oil spill before it causes a widespread damage, getting a cure of a disease or raining today may consider some factors. An environmental scientist called upon to assess the situation of the oil spill will base his prediction on his

informed personal opinion on the type of spill, the amount of oil spilled, the weather condition during clean-up operation and nearness of beaches.

The main disadvantage of this personal approach is that it is always applicable as anyone can have a personal opinion about anything. Its main disadvantage is that its accuracy depends on the accuracy of the information available and ability of for the information to be assessed correctly.

Example 3.2:

3.2(a) Consider the problem of a couple planning to have three children, assuming each child born is equally likely to be a boy(B) or a girl(G).

- (i) List all the possible outcomes in this experiment.
- (ii) What is the probability of the couple having exactly two girls?

Solution:

- (i) The sample space for this experiment is

$$S = \{ BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG \}$$

- (ii) Let A be the event of the couple having exactly two girls. Then,

$$A = \{ BGG, GBG, GGB \}$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{8}$$

3.2(b) Suppose a card is randomly selected from a packet of 52 playing cards.

- (i) What is the probability that it is a “Heart”?
- (ii) What is the probability that the card bears the number 5 or a picture of a queen?

3.2(c) A box contains 4 red, 2 black and 3 white balls. What is the probability of drawing a red ball?

Solution:

- (b) Let the sample space be the set, $S = \{\text{playing cards}\}$, $A = \{\text{Heart cards}\}$, $B = \{\text{cards numbered 5}\}$, $Q = \{\text{cards with a picture of queen}\}$. Then $n(S) = 52$, $n(A) = 13$, $n(B) = 4$ and $n(Q) = 4$

$$(i) P(A) = \frac{n(A)}{n(S)} = \frac{13}{52} = \frac{1}{4}$$

$$(ii) P(B \text{ or } Q) = P(B) + P(Q) = \frac{n(B)}{n(S)} + \frac{n(Q)}{n(S)} = \frac{4}{52} + \frac{4}{52} = \frac{2}{13}$$

- (c) The sample space, $S = \{4R, 2B, 3W\text{-balls}\}$ and let $R = \{\text{red balls}\}$.

$$\text{Then, } P(R) = \frac{n(R)}{n(S)} = \frac{4}{9}$$

- 3.2(d)** A die is tossed twice. List all the outcomes in each of the following events.

Compute the probability of each event.

- (i) The sum of the scores is less than 4
- (ii) Each toss results in the same score
- (iii) The sum of scores on both tosses is a prime number.
- (iv) The product of the scores is at least 20

Solution:

The sample space for the experiment is the set of ordered paired (m, n) , where m and n each takes the values 1, 2, 3, 4, 5 and 6. Thus,

$$S = \{(1,1), (1,2), (1,3), \dots, (6,5), (6,6)\}, \text{ where } n(S) = 36$$

- (i) $A = \{\text{sum of scores less than 4}\} = \{(1,1), (1,2), (2,1)\}$

$$P(A) = \frac{3}{36} = \frac{1}{12}$$

- (ii) $B = \{\text{each toss results in the same score}\}$

$$= \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$$

$$P(B) = \frac{6}{36} = \frac{1}{6}$$

- (iii) $D = \{\text{sum of scores on both tosses is prime}\}$

$$= \{(1,1), (1,2), (1,4), (1,6), (2,1), (2,3), (2,5), (3,2), (3,4), (4,3), (5,2), (5,6), (6,1), (6,5)\}$$

$$P(D) = \frac{14}{36} = \frac{7}{18}$$

- (iv) $E = \{\text{product of the scores is at least 20}\}$

$$= \{(4,5), (4,6), (5,4), (5,5), (5,6), (6,4), (6,5), (6,6)\}$$

$$P(E) = \frac{8}{36} = \frac{2}{9}$$

SESSION 3-3: PROBABILITY OF COMPOUND EVENTS 3-3.1 Definitions

Two or more events are combined to form a single event using the two set operations, \cup and \cap . The event

- $(A \cup B)$ occurs, if either A or B or both occur(s).
- $(A \cap B)$ occurs, if both A and B occur.

We define the following terms associated with compound events:

- *Mutually Exclusive Events*: Two or more events which have no common outcome(s) (never occur at the same time) are said to be *mutually exclusive*. If A and B are mutually exclusive events of an experiment, then

$$A \cap B = \emptyset \text{ and } P(A \cap B) = 0$$

- *Exhaustive Events*: Let A , B and D be mutually exclusive events of the sample space (S) such that they partition the sample space ($S = A \cup B \cup D$). Then A , B and D are said to be mutually exclusive events.
- *Independent Event*: Two or more events are said to be independent if the probability of occurrence of one is not influenced by the occurrence or non-occurrence of the other(s). Mathematically, the two events, A and B are said to be independent, if and only if $P(A \cap B) = P(A) \cdot P(B)$.

However, if A and B are such that, $P(A \cap B) = P(A) \cdot P(B|A)$, they are said to be *conditionally independent*.

- *Conditional Probability*: Let A and B be two events in the sample space, S with $P(B) > 0$. The probability that an event A occurs given that event B has already occurred, denoted $P(A|B)$, is called the conditional probability of A given B . The conditional probability of A given B is defined as.

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{n(A \cap B)}{n(B)}, P(B) > 0 \end{aligned}$$

Example 3.3:

3.3(a) (i) Let A and B be events such that $P(A) = 0.6$, $P(B) = 0.5$ and $P(A \cup B) = 0.8$. Find $P(A|B)$. Are A and B independent?

(ii) In a certain population of women, 40% have had breast cancer, 20% are smokers and 13% are smokers and have had breast cancer.

If a woman is selected at random from the population, what is the probability that she had breast cancer, smokes or both?

Solution:

(i) Given $P(A) = 0.6$, $P(B) = 0.5$ and $P(A \cup B) = 0.8$

- $P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.6 + 0.5 - 0.8 = 0.3$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.3}{0.5} = \frac{3}{5} = 0.6$$

- $P(A) \cdot P(B) = (0.6)(0.5) = 0.3 = P(A \cap B)$, which means that A and B are independent.

(ii) Let B be the event of women with breast cancer and W the event of women who smoke. Then, $P(B) = 0.40$, $P(W) = 0.20$, $P(A \cap B) = 0.13$ and

$$P(B \cup W) = P(B) + P(W) - P(B \cap W) = 0.4 + 0.20 - 0.13 = 0.47$$

3.3(b) The probability that a man hits a target is $\frac{1}{2}$ and that of his son and daughter are $\frac{2}{5}$ and $\frac{1}{5}$ respectively. If they all fire together find the probability that

- (i) they all miss the target,
- (ii) exactly one shot hits the target.,
- (iii) at least one shot hits the target, and
- (iv) the man hits the target given that exactly one hit is registered.

3.3(c) Suppose a batch contains 10 items of which 4 are defective. Two items are drawn at random from the batch one after the other, without replacement. What is the probability that:

- (i) both are defective?
- (ii) the second item is defective?

Solution:

- (b) Let M , S and D be the events of the man, son and daughter hitting target respectively which are all independent. Then $P(M) = \frac{1}{2}$, $P(S) = \frac{2}{5}$ and

$$P(D) = \frac{1}{5}$$

- (i) The probability they all miss target,

$$P(\bar{M} \cap \bar{S} \cap \bar{D}) = [1 - P(M)].[1 - P(S)].[1 - P(D)]$$

$$\begin{aligned} &= \frac{1}{2} \cdot \frac{3}{5} \cdot \frac{4}{5} \\ &= \frac{6}{25} = 0.24 \end{aligned}$$

- (ii) The probability that exactly one shot hits target,

$$\begin{aligned} P(M \cap \bar{S} \cap \bar{D}) \text{ or } \bar{M} \cap S \cap \bar{D} \text{ or } \bar{M} \cap \bar{S} \cap D) \\ &= P(M \cap \bar{S} \cap \bar{D}) + P(\bar{M} \cap S \cap \bar{D}) + P(\bar{M} \cap \bar{S} \cap D) \\ &= \frac{1}{2} \cdot \frac{3}{5} \cdot \frac{4}{5} + \frac{1}{2} \cdot \frac{2}{5} \cdot \frac{4}{5} + \frac{1}{2} \cdot \frac{3}{5} \cdot \frac{1}{5} \\ &= \frac{12}{50} + \frac{8}{50} + \frac{3}{50} \\ &= \frac{23}{50} = 0.46 \end{aligned}$$

- (iii) The probability at least one shot hits target.

$$\begin{aligned} &= 1 - P(\text{all miss target}) \\ &= 1 - P(\bar{M} \cap \bar{S} \cap \bar{D}) \\ &= 1 - \frac{6}{25} = \frac{19}{25} = 0.76, \text{ [from (i)]} \end{aligned}$$

- (iv) Let E be the event that exactly one hit is registered. Then $P(E) = \frac{23}{50}$,

from (ii) and the required probability is given as follows:

$$\begin{aligned} P(M / E) &= \frac{P(M \cap E)}{P(E)} = \frac{P(M \cap \bar{S} \cap \bar{D})}{P(E)} \\ &= \frac{\frac{1}{2} \cdot \frac{3}{5} \cdot \frac{4}{5}}{\frac{23}{50}} \\ &= \frac{12}{23} = 0.522 \end{aligned}$$

(c) Let D_1 and D_2 be the events, first and second item drawn is defective respectively. Then, $P(D_1) = \frac{4}{10}$, $P(D_2|D_1) = \frac{3}{9}$, $P(D_2|D_1^c) = \frac{4}{9}$

(i) The probability that both are defective,

$$\begin{aligned} P(D_1 \cap D_2) &= P(D_1)P(D_2|D_1) \\ &= \frac{4}{10} \cdot \frac{3}{9} = \frac{2}{15} \end{aligned}$$

(ii) The probability that second item is defective,

$$\begin{aligned} P(D_1 \cap D_2 \text{ or } \bar{D}_1 \cap D_2) &= P(D_1 \cap D_2) + P(\bar{D}_1 \cap D_2) \\ &= P(D_1)P(D_2|D_1) + P(\bar{D}_1)P(D_2|\bar{D}_1) \\ &= \frac{2}{15} + \frac{6}{10} \cdot \frac{4}{9} = \frac{18}{45} = \frac{2}{5} = 0.4 \end{aligned}$$

3-3.2 Some Basic Rules/Theorems of Probability

3-3.2.1 Axioms of Probability

Let S be a sample space, E , class of events and P , a real-valued function defined on E . Then P is called probability function or measure and $P(A)$, the probability of the event A , if the following axioms hold:

- A.1: For every event A , $0 \leq P(A) \leq 1$
- A.1: $P(S) = 1$
- A.3: If A and B are mutually exclusive events, then,

$$P(A \cup B) = P(A) + P(B)$$
- A.4: If $A_1, A_2, A_3, \dots, A_n$ is a sequence of n mutually exclusive events, then,

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) \quad \text{or} \quad P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

The following theorems arise directly from the above axioms:

- Theorem 1: If \emptyset is the empty set, the $P(\emptyset) = 0$.
- Theorem 2: If \bar{A} is the complement of an event A , then $P(\bar{A}) = 1 - P(A)$.
- Theorem 3: If $A \subseteq B$, then $P(A) \leq P(B)$.

- *Theorem 4:* If A and B are two events, then $P(\bar{B}) = P(A) - P(A \cap B)$
- *Theorem 5:* If A and B are any two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$
- *Corollary:* For any events A_1, A_2 and A_3 ,

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) \\ &\quad - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3) \end{aligned}$$

3-3.2.2 The Addition Rule:

Let $A_1, A_2, A_3, \dots, A_n$ be events of the samples space, S . Then

- $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
- $$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) \\ &\quad - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3) \end{aligned}$$

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

- $P(A_1 \cup A_2) = P(A_1) + P(A_2)$
- $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$
- $P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$

3-3.2.3 The Multiplication Theorem:

If $A_1, A_2, A_3, \dots, A_n$ are events of the sample space, S , then

- $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 | A_1)$
- $P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2)$
- $$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \times \dots \times \\ &\quad P(A_n | A_1 \cap A_2 \cap \dots \cap A_{n-1}) \end{aligned}$$

3-3.2.4 The Total Probability Rule:

Suppose the sample space, S is partitioned into mutually exclusive events, $A_1, A_2, A_3, \dots, A_n$. Let B any other event in S intersecting with all the n events.

Then we have,

$$\begin{aligned} B &= S \cap B \\ &= (A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n) \cap B \\ &= (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B) \cup \dots \cup (A_n \cap B), \end{aligned}$$

where $(A_i \cap B)$, for $i = 1, 2, 3, \dots, n$ are also mutually exclusive events, then

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) + \dots + P(A_n \cap B) \\ &= P(A_1).P(B|A_1) + P(A_2).P(B|A_2) + P(A_3).P(B|A_3) + \dots + P(A_n).P(B|A_n) \\ &= \sum_{i=1}^n P(A_i).P(B|A_i), \end{aligned}$$

which is the total probability rule for finding $P(B)$. The general form of this rule is called *Bayes' Theorem* which states as follows. The conditional probability of A_i given B is defined as

$$P(A_i|B) = \frac{P(A_i).P(B|A_i)}{\sum_{i=1}^n P(A_i).P(B|A_i)}$$

Example 3.3:

3.3(d) A die is tossed once. If one (1) appears a ball is drawn from *Box 1* and if two (2) or three (3) appears a ball is drawn from *Box 2*, otherwise a ball is drawn from *Box 3*. The contents of the boxes are as follows:

Box 1: 5 white, 3 Green and 2 Red balls; *Box 2*: 1 white, 6 Green and 3 Red balls; *Box 3*: 3 white, 1 Green and 6 Red balls.

Find the probability that.

- (i) The ball chosen is red.
- (ii) Box 2 is selected given that the ball chosen is red.

Solution:

(d) Let B_1 , B_2 and B_3 be the events of selecting a ball from the boxes and R , the event of choosing a red ball. Then, we have: $B_1 = \{5W, 3G, 2R\}$,

$B_2 = \{1W, 6G, 3R\}$, $B_3 = \{3W, 1G, 6R\}$, where

$$\begin{aligned} P(B_1) &= \frac{1}{6}, \quad P(R|B_1) = \frac{2}{10} \\ P(B_2) &= \frac{2}{6}, \quad P(R|B_2) = \frac{3}{10} \\ P(B_3) &= \frac{3}{6}, \quad P(R|B_3) = \frac{6}{10} \end{aligned}$$

- (i) The probability that the ball chosen is red,

$$\begin{aligned} P(R) &= P(B_1).P(R|B_1) + P(B_2).P(R|B_2) + P(B_3).P(R|B_3) \\ &= \frac{1}{6} \cdot \frac{2}{10} + \frac{2}{6} \cdot \frac{3}{10} + \frac{3}{6} \cdot \frac{6}{10} = \frac{13}{30} \end{aligned}$$

(ii) The required probability is

$$P(B_2|R) = \frac{P(B_2 \cap R)}{P(R)} = \frac{P(B_2).P(R|B_2)}{P(R)}, \text{ by multiplication rule.}$$

$$= \frac{\cancel{2}/6 \cdot \cancel{3}/10}{\cancel{13}/30} = \cancel{3}/13$$

- 3.3(f)** A population is composed of 60% men and 40% women. It is known that 75% of the men and 45% of the women smoke cigarettes. What is the probability that a person in the population observed smoking is a man?
- 3.3(g)** Three candidates running for the office of the SRC presidency of KNUST all promise not allow an increase of AFUF next academic year. The probabilities of the candidates, C_1 , C_2 and C_3 , winning the election are 0.45, 0.25 and 0.30 respectively. The probabilities for an increase in AFUF should C_1 , C_2 and C_3 win the election are 0.60, 0.15 and 0.45 respectively.
- (i) What is the probability that there will be increase in AFUF?
 - (ii) If coming next academic year it is found in the newspaper that there has been an increase in AFUF, what is the probability that candidate C_1 was elected? Interpret your result.

Solution:

- (f) Let M and W be the events of observing a man and woman respectively and B be the event, smoking of cigarettes. Then $P(M) = 0.60$, $P(B|M) = 0.75$, $P(W) = 0.40$, $P(B|W) = 0.45$, and
- $$\begin{aligned} P(B) &= P(M).P(B|M) + P(W).P(B|W) \\ &= (0.60)(0.75) + (0.40)(0.45) \\ &= 0.63 \end{aligned}$$

Hence the probability that a person observed smoking is a man is

$$\begin{aligned} P(B|M) &= \frac{P(M).P(B|M)}{P(B)} \\ &= \frac{(0.60)(0.75)}{0.63} = \frac{5}{7} = 0.71 \end{aligned}$$

- (g) Let A be the event of an increase in AFUF. Then given the probabilities,

$$P(C_1) = 0.45, \quad P(A|C_1) = 0.60, \quad P(C_2) = 0.25, \quad P(A|C_2) = 0.15, \text{ and}$$

$P(C_3) = 0.30, \quad P(A|C_3) = 0.45$, we compute the following:

$$\begin{aligned} \text{(i)} \quad P(A) &= P(C_1).P(A|C_1) + P(C_2).P(A|C_2) + P(C_3).P(A|C_3) \\ &= (0.45)(0.60) + (0.25)(0.15) + (0.30)(0.45) \\ &= 0.27 + 0.0375 + 0.135 = 0.4425 \end{aligned}$$

$$\begin{aligned} \text{(ii)} \quad P(C_1|A) &= \frac{P(C_1).P(A|C_1)}{P(A)} = \frac{P(C_1 \cap A)}{P(A)} \\ &= \frac{(0.45).(0.60)}{0.4425} = 0.61017, \end{aligned}$$

which gives a high likelihood of candidate C_1 being elected as SRC president.

- 3.3(h)** A lot of 50 spacing washers contain 30 washers that are thicker than the target dimension. Suppose that three washers are selected at random without replacement from the lot. What is the probability that

- (i) all the three washers are thicker than the target?
- (ii) the third washer selected is thicker than the target if the first two washers selected are thinner than target?
- (iii) the third washer selected is thicker than the target?

- 3.3(j)** Consider a sample space made up of adults in a town who have completed the requirements for a university degree. The table below indicates their employment status categorized according to sex.

Sex	Employed (E)	Unemployed (U)	Total
Male(M)	400	100	500
Female(F)	200	200	400
Total	600	300	900

If one of these adults is to be chosen at random what is the probability that the person is:

- (i) man and employed? (ii) employed?
- (iii) man given that he is employed?

Solution:

- (h) (i) The probability that all the 3 washers are thicker,

$$P(T_1 T_2 T_3) = P(T_1).P(T_2).P(T_3)$$

$$= \frac{30}{50} \cdot \frac{29}{49} \cdot \frac{28}{48} = \frac{29}{140} = 0.2071 \text{ or } \frac{^{30}C_3}{^{50}C_3}$$

- (ii) The probability that the third washer is thicker than the first two,

$$P(\bar{T}_1 \bar{T}_2 T_3) = \frac{20}{50} \cdot \frac{19}{49} \cdot \frac{30}{48} = \frac{19}{196} = 0.097$$

- (iii) The probability that the third washer is thicker,

$$\begin{aligned} & P(T_1 T_2 T_3 \text{ or } T_1 \bar{T}_2 T_3 \text{ or } \bar{T}_1 \bar{T}_2 T_3) \\ &= P(T_1 T_2 T_3) + P(T_1 \bar{T}_2 T_3) + P(\bar{T}_1 \bar{T}_2 T_3) \\ &= \frac{30}{50} \cdot \frac{29}{49} \cdot \frac{28}{48} + \frac{30}{50} \cdot \frac{20}{49} \cdot \frac{29}{48} + \frac{20}{50} \cdot \frac{19}{49} \cdot \frac{30}{48} \\ &= 0.207 + 0.148 + 0.097 = 0.452 \end{aligned}$$

- (j) Given that M and E are events of choosing a man and employed person

respectively. Then, $P(M) = \frac{500}{900} = \frac{5}{9}$, $P(F) = \frac{400}{900} = \frac{4}{9}$

$$(i) P(\text{man and employed}) = P(M \cap E) = \frac{400}{900} = \frac{4}{9} = 0.4444$$

$$(ii) P(\text{employed}) = P(E) = \frac{600}{900} = \frac{2}{3} = 0.6667, \text{ or}$$

$$P(E) = P(M).P(E|M) + P(F).P(E|F)$$

$$P(E) = \frac{5}{9} \cdot \frac{4}{5} + \frac{4}{9} \cdot \frac{1}{2} = \frac{20}{45} + \frac{4}{18} = \frac{2}{3}$$

- (iii) The probability of choosing a man given that he is employed,

$$\begin{aligned} P(M|E) &= \frac{P(M \cap E)}{P(E)} \\ &= \frac{P(M).P(E|M)}{P(E)} \\ &= \frac{\frac{5}{9} \cdot \frac{4}{5}}{\frac{2}{3}} = \frac{2}{3} \end{aligned}$$

SECTION 4-3: APPLICATION OF COUNTING TECHNIQUES

The classical definition of probability of an event A , $P(A)$ requires the knowledge of the number of outcomes of A and the total possible outcomes of the experiment. To find these outcomes we list such outcomes explicitly, which may be impossible if they are too many. Counting Techniques may be very useful to determine the number of outcomes and compute $P(A)$. We shall examine three basic counting techniques, namely the *Multiplication Principle*, *Permutation* and *Combination*.

4-3.1 The Multiplication Principle

The *Multiplication principle*, also known as the *Basic counting principle* states as follows. *If an operation can be performed in n_1 ways, a second operation can be performed in n_2 ways and so on for k^{th} operation which can performed in n_k ways, then the combined experiment or operations can be performed in $n_1 \times n_2 \times n_3 \times \dots \times n_k$ ways.*

For example,

- Tossing a coin has two possible outcomes and tossing a die has six possible outcomes. Then the combined experiment, tossing the coin and die together results in $2 \times 6 = 12$ possible outcomes: $H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6$
- The number of different ways for a man to get dressed if he has 8 different shirts and 6 different pairs of trousers is $8 \times 6 = 48$
- The number of ways a three-figure integer be formed from the numbers, 4, 3, 5, 6 and 7 if no number is used twice or more is $5 \times 4 \times 3 = 60$.

Applying the multiplication principle, results in the other two counting techniques, namely *Permutation* and *Combination*, used to find the number of possible ways when a fixed number of items are to be picked from a lot without replacement.

4-3.2 Permutation of Objects

An ordered arrangement of objects is called a *permutation*. The number of permutations of

- (i) n distinct objects, taken all together is $n! = n(n-1)(n-2) \times \dots \times 3 \times 2 \times 1$

- (ii) n distinct objects taken k at a time is ${}^n P_k$ or $P(n, k) = \frac{n!}{(n-k)!}$, where $k < n$.
- (iii) n objects consisting of groups of which n_1 of the first group are alike, n_2 of the second group are alike and so on for the k^{th} group with n_k objects which are alike is $\frac{n!}{n_1!.n_2!.n_3! \dots n_k!}$, where $n = n_1 + n_2 + \dots + n_k$
- (iv) n distinct objects arranged in a circle, called *circular permutations* is given by

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

For example,

- The number of possible permutations of the letters, A , B and C is $3! = 6$. The required permutations are ABC , BAC , ACB , BCA , CAB and CBA .
- The number of permutations of 10 distinct digits taken two at a time

$$= {}^{10}P_2 = \frac{10!}{(10-2)!} = 10 \times 9 = 90..$$

- The number of permutations of the letters forming the following 14-letter word, $S C I E N T I F I C A L L Y$, which contains 2C's, 3I's, 2L's, and 1's of the rest of letters $= \frac{14!}{2!.3!.2!} = 3,632,428,800$
- The number of circular permutations of 6 persons sitting around a circular table
 $= 5! = 120$

4-3.3 Combinations of Objects

A *combination* is a selection of objects in which the order of selection does not matter.

The number of ways in which k objects can be selected from n distinct objects, irrespective of their order is defined by

$${}^n C_k \text{ or } \binom{n}{k} = \frac{n!}{(n-k)! k!} = \frac{{}^n P_k}{k!}$$

For example,

- the number of ways choosing a committee of 5 from 9 persons is

$${}^9 C_5 = \frac{9!}{4!5!} = 126.$$

- The number of combinations of the letter a , b , c , d and e , taken 3 at time is $\binom{5}{3} = 10$ which are listed follows: abc , abd , abe , acd , ace , ade , bcd , bce , bde , cde

Example 3.4

- 3.4(a)** (i) In how many ways can a three-figure integer is formed from the numbers: 4, 3, 5, 6 and 7 if any number can be used more than once?
- (ii) In a certain examination paper, students are required to answer 5 out of 10 questions from *Section A* another 3 out of 5 questions from *Section B* and 2 out of 5 questions from *Section C*. In how many ways can the students answer the examination paper?

Solution:

- (i) The first, second and third numbers, each can be chosen in 5 ways. The total number of ways $= 5 \times 5 \times 5 = 125$
- (ii) The number of ways of answering the questions in *Section A*
 $= 10 \times 9 \times 8 \times 7 \times 6 = 30,240$
- The number of ways of answering the questions in section *B*
 $= 5 \times 4 \times 3 = 60$
- The number of ways of answering the questions in section *C*
 $= 5 \times 4 = 20$
- Hence the students can answer the questions in the three sections in
 $= 30,240 \times 60 \times 20 = 36,288,000$

- 3.4(b)** A company codes its customers by giving each customer an eight character code. The first 3 characters are the letter A , B and C in any order and the remaining 5 are the digits 1, 2, 3, 4 and 5 also in any order. If each letter and digit can appear only once then number of customers the company can code is obtained as follows:

The first 3 letters can be filled in $3!$

The next 5 digits can be filled in $5!$

Then the required number $= 3! \times 5! = 720$

3.4(c) In many ways can 4 boys and 2 girls seat themselves in a row if

- (i) The 2 girls are to sit next to each other?
- (ii) The 2 girls are not to sit next to each other?

Solution:

(i) If we regard the 2 girls as a separate persons ($B_1 B_2 B_3 B_4 G_1 G_2$), then the number of arrangements of 5 different persons, taken all at a time = $5!$

The 2 girls can exchange places and so the required number of ways they can seat themselves = $5! \times 2! = 240$

(ii) The number of ways the boys can arrange themselves = $4!$

The number of ways the 2 girls can occupy the arrowed places:

$$\uparrow B_1 \uparrow B_2 \uparrow B_3 \uparrow B_4 \uparrow = {}^5P_2 = 5 \times 4$$

The required number of permutations (with the 2 girls not sitting next to each other) = $4! \times 5 \times 4 = 480$

3.4(d) Find the number of ways in which a committee of 4 can be chosen from 6 boys and 5 girls if it must

- (i) Consist of 2 boys and 2 girls.
- (ii) Consist of at least 1 boy and 1 girl.

Solution:

(i) The number of ways of choosing 2 boys from 6 and 2 girls from 5

$$= \binom{6}{2} \cdot \binom{5}{2} = 15 \times 10 = 150$$

(ii) For the committee to contain at least 1 boy and 1 girl we have

$$1B3G, 2B2G \text{ or } 3B1G$$

The required number of ways

$$\begin{aligned} &= \binom{6}{1} \cdot \binom{5}{3} + \binom{6}{2} \cdot \binom{5}{2} + \binom{6}{3} \cdot \binom{5}{1} \\ &= 6(10) + 15(10) + 20(5) = 130 \end{aligned}$$

3.4(e) (i) A school Parent-Teacher committee of 5 members is to be formed from 6 parents, 2 teachers and the principal. In how many ways can the committee be formed in order to include

- (α) The principal? (β) Exactly four parents?
- (γ) Not more than four parents?

- (ii) Four balls are drawn from a bag of 12 balls of which 7 are blue and 5 are red. In how many of the possible combinations of 4 balls is at least a red?

Solution:

- (i) (α) If the principal is to be included then we select 4 people from the remaining 8. Hence required number of ways the committee is formed

$$= \binom{1}{1} \cdot \binom{8}{4} = 70$$

- (β) The number of ways of selecting 4 parents out of 6 = $\binom{6}{4}$. The number of ways of selecting the remaining number from the 3 (2 teachers and the principal) = $\binom{3}{1}$

Therefore the number of ways of selecting exactly 4 parents

$$= \binom{6}{4} \cdot \binom{3}{1} = 15 \times 3 = 45$$

- (γ) The number of ways of forming a 5-member committee = $\binom{9}{5}$
The number of ways of selecting 5 parents from 6 = $\binom{6}{5}$

Therefore the required number of ways of selecting a committee with not more

$$\text{than 4 parents} = \binom{9}{5} - \binom{6}{5} = 126 - 6 = 120$$

- (ii) If at least one red is to be included then the combinations include

$$1R\ 3B, \text{ with number of combinations} = \binom{5}{1} \binom{7}{3} = 175$$

$$2R\ 2B, \text{ with number of combinations} = \binom{5}{2} \binom{7}{2} = 210$$

$$3R\ 1B, \text{ with number of combinations} = \binom{5}{3} \binom{7}{1} = 70$$

$$4R, \text{ with number of combinations} = \binom{5}{4} \binom{7}{0} = 5$$

$$\text{The total number combinations} = 175 + 210 + 70 + 5 = 460$$

3.4(f) A board consist of 12 men and 8 women. If a committee of 3 members is to be formed what is the probability that

- (i) It includes at least one woman?
- (ii) It includes more women than men?
- (iii) A particular woman is included?

Solution:

The number of ways of forming the committee of 3 from $(12M + 8W) = \binom{20}{3} = 1140$

(i) The probability that at least one women

$$= P(1W2W) + P(2W1M) + P(3W)$$

$$= \frac{\binom{8}{1} \cdot \binom{12}{2} + \binom{8}{2} \cdot \binom{12}{1} + \binom{8}{3}}{1140}$$

$$= \frac{528 + 336 + 56}{1140} = \frac{920}{1140} = \frac{46}{57}$$

(ii) The probability that more women than men

$$= (2W1M) + (3W)$$

$$= \frac{336}{1140} + \frac{56}{1140} = \frac{196}{575} = 0.34087,$$

(iii) The probability that a particular women is included

$$= P(\text{that woman and any other } 2)$$

$$= \frac{^{19}C_2}{1140} = \frac{171}{1140} = 0.15$$

3.4(g) A box contains 6 red, 3 white and 5 blue balls. If three balls are drawn at random, one after the other without replacement, find the probability that

- | | |
|------------------------|-------------------------------|
| (i) All are red | (ii) 2 are red and 1 is white |
| (ii) at least 1 is red | (iv) one of each colour |

Solution:

$$(i) \quad \text{Pr}(3 \text{ red balls}) = \frac{\text{no. of selection of 3 from 3}}{\text{no. of selections 3 from 14}}$$

$$= \frac{^6C_3}{^{14}C_3} = \frac{6 \times 5 \times 4}{14 \times 13 \times 12} = \frac{5}{91}$$

$$(ii) \quad \Pr(2 \text{ red and } 1 \text{ white ball}) = \frac{{}^6C_2 \cdot {}^3C_1}{{}^{14}C_3} = \frac{45}{364}$$

$$(iii) \quad P(\text{at least 1 red}) = 1 - P(\text{none is red}) = 1 - \frac{{}^8C_3}{{}^{14}C_3} = 1 - \frac{2}{13} = \frac{11}{13}$$

$$(iv) \quad P(\text{one of each colour})$$

$$= \frac{{}^6C_1 \cdot {}^3C_1 \cdot {}^5C_1}{{}^{14}C_3} = \frac{6 \times 3 \times 5}{364} = \frac{45}{182}$$

4-3.4 Trial Questions 1-3:

(a) (i) Prove that for any two events A_1 and A_2 , $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$

(ii) Show that if A and B are independent events, then

$$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \cdot P(\bar{B})$$

(b) (i) A smoke detector system uses two devices, A and B . If smoke is present the probability that it will be detected by device A is 0.95, B is 0.90 and both devices is 0.88. If smoke is present, find the probability that it will be detected by either device A , B or both. Also, find the probability that it will be undetected.

(ii). A box contains 10 balls: 4 red and 6 blue. A second box contains 16 red and unknown numbers of blue balls. A single ball is drawn from each box. The probability that both balls are the same colour is 0.44. Calculate the number of blue balls in the second box.

(c) Two methods, A and B are available for teaching a certain industrial skill. The failure rate is 20% for A and 10% for B . However, B is more expensive and hence used only 30% of the time while A is used for the other 70%. A worker is taught the skill by one of the methods but fails to learn it correctly. What is the probability that he/she was taught by method A .

(d) The probability that a man will pass a certain examination is $\frac{1}{4}$ and the corresponding probability for his wife is $\frac{1}{3}$. Find the probability that

(i) Both will pass the examination.

(ii) Neither will pass the examination.

(iii) At least one of them will pass the examination.

- 2.(a) Suppose that A and B are independent events such that the probability that neither occurs is a and the probability of B is b show that

$$P(A) = \frac{1-b-a}{1-b}.$$

- (b) A student is suffering from malaria and she is to see the doctor at the University Hospital. The probabilities that she will be given an injection, tablet and both are 0.46, 0.38 and 0.22 respectively. What is the probability that the student will be given:
- (i) either injection, tablet or both.
 - (ii) injection but not tablet.
 - (iii) injection or tablet but not both.
- (c) A large group of people is to be checked for common symptoms of a certain disease. It is thought that 20% of people that posses symptom A alone, 30% posses symptom B alone, 10% posses both symptoms and the rest have neither symptom. If a person is randomly chosen from the group, find the probability that he/she has
- (i) Neither symptom. (ii) At least one symptom.
 - (iii) Both symptoms given that he/she has symptom B .

- 3.(a) (i) A die is weighted so that 4 is twice as likely to appear as 1, 2, 3, or 5 and 6 is twice as likely to appear as 4, find the probability of obtaining an even number on a single roll of the die.
- (ii) A basket contains 10 oranges of which 4 are rotten. Find the probability that if the oranges are taken from the basket one at a time without replacement, the third orange taken is rotten.
- (b) (i) Given the two events A and B and the probabilities, $P(A) = 0.5$, $P(B) = 0.25$, $P(A|B) = 0.8$, use the appropriate laws of probability to compute $P(A \cap B)$ and $P(A \cup B)$.
- (ii) A bag contains 4 red and 6 green identical marbles. If two marbles are drawn together at random from the bag, find the probability that they are of the same colour.

- 4.(a) A certain television (TV) set is found to be defective. It could have been manufactured at any of the four factories A , B , C and D . Of all such TV sets, 10% are produced at A , 15% at B , 55% at C and 20% at D . It was determined that 3% of the sets produced at A , 1.5% of those produced at B , 2% of those produced at C and 5% of those produced at D are defective. For each factory, find the probability that the defective set came from that.
- (b) A medical diagnostic test for a certain disease will yield either a positive or a negative reaction. If you have the disease there is a 0.99 chance that the test result will be positive while if you do not have the disease, there is a 0.90 chance that the test will be negative. If it is estimated that 0.03 of the population have the disease, find the probability that
- (i) You have the disease, given that you have a positive reaction.
 - (ii) You do not have the disease, given that you have a negative reaction.
- (c) There is an incidence rate of 0.01 of a disease in a certain community. Of those having the disease, 95% test positive when a certain diagnostic test was applied while those not having the disease, 90% test positive when the test was applied. Suppose that an individual from this community is randomly selected and given the test.
- (i) Find the probability of the individual selected has the disease and test positive.
 - (ii) Compute the probability that the individual selected tests positive.
 - (iii) Find the probability that the individual selected has the disease given that he/she tests positive.
- (d) An insurance company classifies policyholders as either *good* (G), *bad* (B) and *questionable* (Q). It is on record that the risk of good, bad and questionable drivers having an accident is 0.02, 0.09 and 0.04 respectively. If 57 percent of the current policyholders are good drivers, 23 percent are bad drivers and 20 percent are questionable drivers, and an accident is reported, what is the probability that the person
- (i) Is a good driver?
 - (ii) Is a questionable driver?
 - (iii) Is bad driver?

- 6.(a) (i) Prove the theorems arising from the Axioms of Probability.
- (ii) Explain the terms:
- *mutually exclusive events,*
 - *exhaustive events,*
 - *independent events, and*
 - *conditional probability of an event.*
- (b) Of the patients reporting to a clinic with the symptoms of sore throat and fever, 25% have strep throat, 50% have an allergy and 10% have both.
- (i) What is the probability that a patient selected at random has either strep throat, an allergy or both?
- (ii) Are the occurrence of strep throat and allergy independent?

PROBABILITY DISTRIBUTIONS

This unit combines methods of descriptive statistics and that of concepts of probability discussed in the previous two units to develop the notion of a random variable, a variable whose numerical observed value is determined by chance, and probability distribution, the probability associated with each observed value. The main topics presented include the concepts of probability distributions and expectation of random a variable.

Learning Objectives

The unit aims to introduce students to important definitions, concepts and computational techniques associated with random variable. This means that after studying the unit, students will be able to:

- Define a random variable and distinguish between discrete and continuous random variable.
- Identify numerical data as either discrete or continuous random variables and give some examples.
- State the properties of probability distributions and compute the probabilities of the various numerical values of a random variable.
- Compute the expected value (mean), variance, median and the mode of a random variable.

SESSION 1-4: CONCEPTS OF PROBABILITY DISTRIBUTIONS

1-4.1 Random Variables

The generated outcomes of most experiments that are performed, as we noted in the introductory study of Probability, are random. The random outcomes of these experiments can be represented by simple real numbers. For example, in a study on operation of a supermarket, the random experiment here might involve the random selection of a customer leaving a store.

The potential variables could be the number of items purchased (x) and the time spent for the service (y). It should be noted that until a customer is selected there is uncertainty about the values of x and y . Other classic examples of numerical

outcomes of random experiments are the amount of gasoline that is lost to evaporation during the filling of a gas tank, the number of bits in error in a digital communication channel, the weights of newly born babies in a health centre, the number of bacterial per unit area in the study of drug control on bacterial growth, the number of molecules in a sample of gas, the number of accidents occurring at an intersection in a period of time, the sugar content of samples of food drinks, the number of heads (H 's), x observed in three tosses of a coin results, and so on.

The numerical outcomes of these experiments which can change from experiment to experiment are called *random variables*. Random variables are useful tool for describing random experiments and of great interest in Probability Theory. The probabilities of such random variables are always needed to make statistical inferences about a given situation (or population).

- Definition 4.1:

A random variable is a real-valued function defined on the outcomes of a chance experiment. (or a function that assigns a real number of each outcome in the sample space of a random experiment.

Random variables are denoted by uppercase letters, such as X , Y and Z and the corresponding lowercase letters such as x , y and z are used to denote the particular values of X , Y and Z . We refer to the set of possible numbers of a random variable as the *range*. They are classified as either being *discrete* or *continuous*, depending upon the range of values they assume.

- Definition 4.2:

A random variable, X is said to be discrete if it can take on only a finite number or a countably infinite possible values of X .

Discrete random variables represent counts associated with real phenomena. For example, the number of errors that a machine makes in an assembly operation, the number of customers awaiting to be served at a supermarket, the number of accidents at a particular plant for a given period, the number of bacterial per unit area in the study of drug control on bacteria growth, the number of defective television sets in a shipment., the number of errors detected in accounting records, etc.

- Definition 4.3:

A random variable, X is said to be continuous if it can assume infinitely many values within an interval of real numbers.

Examples are the length of time to complete an operation in a manufacturing plant, the heights or weights of a group of people, the amount of energy produced by utility a company for a given period, the length of life span of an electric bulb, the amount of sugar in a bottled drink, the length of time a customer waits for a service at a counter, etc. A complete description of a random variable is to specify its probability distribution which is discussed in the following sequel.

1-4.2 Probability Distributions

1-4.2.1 Introduction

Probability distributions provide a means of determining the probability of different values of the random variable occurring in an experiment. They are used in Statistics to provide a description of a population data. That is, the probability distribution of a random variable X , denoted $p(x)$ or $f(x)$, is a description of the set of possible values of X along with the probability, $p(x)$ or $f(x)$ associated with each of the possible values. Some probability distributions are used so extensively in statistical analysis that special formulae and/or tables have been developed for computing the probabilities associated with them. They are classified as either *discrete* or *continuous*, depending upon the numerical values their random variables can assume.

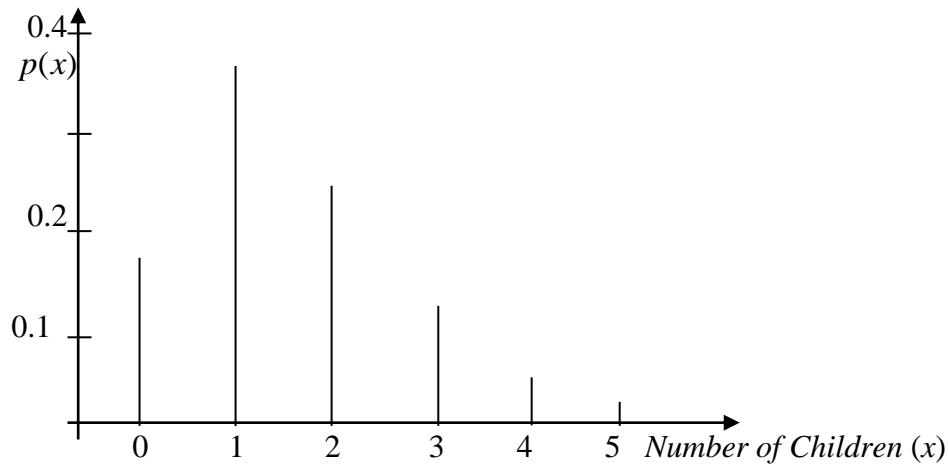
1-4.2.2 Types of Probability Distributions

A probability distribution is described as either discrete or continuous.

- *Discrete Distributions:* The probability distribution for a discrete random variable X is a *formula, table, graph* or *any device* that specifies the probability associated with each possible value of X . For example, a study on 300 families in a community was conducted, noting the number of children, X and its occurrence, f in a family results the following distribution:

X	0	1	2	3	4	5
f	54	114	72	42	12	6
$p(x)$	0.18	0.38	0.24	0.14	0.04	0.02

The distribution is further illustrated on a line histogram as follows:



- Definition 4.4:

The probability that X takes a discrete value, denoted, $P(X = x)$ or $p(x)$ is called probability mass function (pmf), if the following properties are satisfied:

- (i) $p(x) = P(X = x)$
- (ii) $0 \leq p(x) \leq 1$ or $p(x) \geq 0$, for all x , meaning, probability that the random variable X assumes a value x is always between 0 and 1, inclusive.
- (iii) $\sum_x p(x) = 1$, the sum of all probabilities is equal to 1.

- *Continuous Probability Distributions:* The relative frequency behaviour of continuous random variable, X is modelled by a function, $f(x)$ which is more often called *probability density function (pdf)*. The graph of $f(x)$ is a smooth curve defined over a range of interval the random variable, X assumes.

The area under the graph of $f(x)$ gives the probability that x falls in an interval. Thus, the probability that X assumes a value within the interval $[a, b]$ is the area bounded by $x = a$, $x = b$, $x = 0$ and $f(x)$.

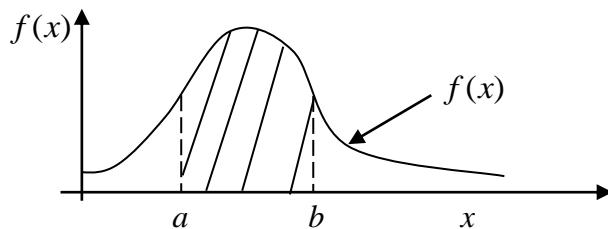


Figure 4.1: Graph of probability density function

That is, from the diagram above,

$$P(a \leq x \leq b) = \int_a^b f(x)dx = \text{shaded area in Figure 4.1.}$$

- Definition 4.4:

The probability distribution, $f(x)$ is said to be probability density function of the continuous random variable, x if for an interval of real numbers $[a, b]$ the following properties are satisfied:

(i) $f(x) \geq 0$, for any value of x .

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

(iii) $P(a \leq x \leq b) = \int_a^b f(x)dx$, where $-\infty \leq a \leq x \leq b \leq \infty$. If $a = b$, then

$$P(a \leq x \leq a) = P(x = a) = \int_a^a f(x)dx = 0,$$

probability that a continuous random assumes a particular value, a is zero.

- *Cumulative Distribution Functions:* The cumulative distribution function (cdf) for a random variable x , denoted, $F(x)$ is defined by $F(x) = P(X \leq x)$.

If x is a discrete random variable with probability mass function, $p(x)$

then, $F(x) = \sum_t^x p(t)$, which is a step function.

If X is, however, a continuous random variable with probability density

function, $f(x)$, then $F(x) = \int_{-\infty}^x f(t)dt$,

where $-\infty \leq x \leq \infty$, $f(x) = \frac{dF(x)}{dx}$ and $P(x_1 \leq x \leq x_2) = F(x_2) - F(x_1)$.

In each case, $F(x)$ is a monotonic increasing function with the following properties:

(i) $F(a) \leq F(b)$, wherever $a \leq b$, and

(ii) The limit of $F(x)$ to the left is 0 and to the right is 1. That is,

$$\lim_{x \rightarrow -\infty} F(x) = 0 \text{ and } \lim_{x \rightarrow \infty} F(x) = 1$$

(iii) $0 \leq F(x) \leq 1$

Example 4.2:

4.2(a) Construct probability distributions for the following random variables:

- (i) The number of heads when four fair coins are tossed.
- (ii) The difference between the results of two fair dice rolled together.

Solution:

- (i) The sample space for tossing four fair coins:

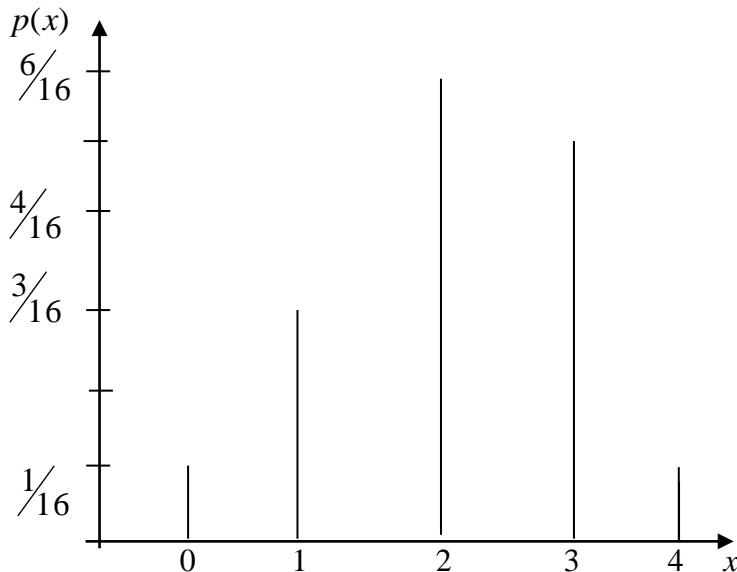
$$S = \{HHHH, HHHT, HHTT, HHTH, HTHH, HTHT, HTHH, HTTT, THHH, THHT, THTT, THTH, TTTH, TTHT, TTTH, TTTT\}$$

The random variable, X is the number of heads occurring in that experiment which assumes the values, $X = 0, 1, 2, 3, 4$. The required probability distributions is

X	0	1	2	3	4
p(x)	1/16	3/16	6/16	5/16	1/16

where $\sum_{x=0}^4 p(x) = 1/16 + 3/16 + 6/16 + 5/16 + 1/16 = 1$
and $p(x) > 0$, for each x .

Representing on a line histogram:



- (ii) The sample space for the experiment is given by the table below.

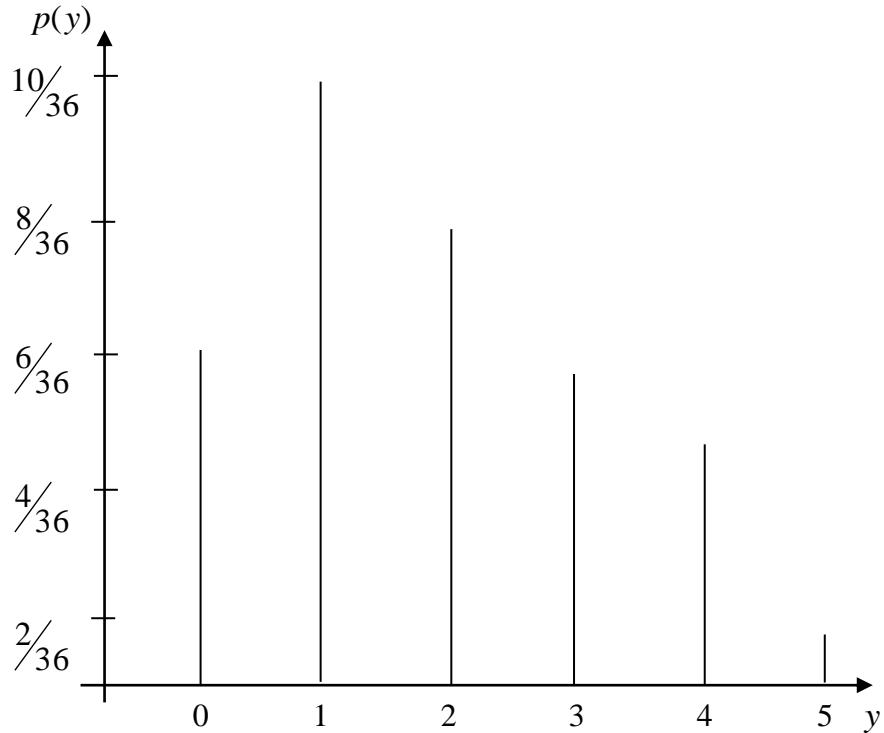
D_1	D_2					
	1	2	3	4	5	6
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

The possible values of this random variable (difference between the results of the two rolled dice), y are 0, 1, 2, 3, 4, and 5. The required probability distribution is

Y	0	1	2	3	4	5
$p(y)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

where $\sum_{y=1}^5 p(y) = \frac{1}{36}(6 + 10 + 8 + 6 + 4 + 2) = 1$ and $p(y) > 0$, for all y . The

line histogram is given by



- 4.2(b)** The number of telephone calls received in an office between 12.00 noon and 1.00 pm has the probability function given by

x	0	1	2	3	4	5	6
$p(x)$	0.05	0.20	0.25	0.20	0.10	0.15	0.05

- (i) Verify that it is probability mass function.
- (ii) Find the probability that there will be 3 or more calls.
- 4.2(c)** Verify that the following probability distribution functions are probability mass function.

$$(i) \quad p(x) = \begin{cases} \frac{1}{21}(2x+3), & x = 1, 2, 3 \\ 0 & , elsewhere \end{cases}$$

$$(ii) \quad p(x) = \begin{cases} k(x-1), & x = 3, 4, 5 \\ 0 & , elsewhere \end{cases}$$

Solution:

- (b) (i) To verify that it is probability mass function, we have $p(x) > 0$, for $x = 0, 1, 2, 3, 4, 5$ and 6 and

$$\sum_{i=1}^6 p(x) = 0.05 + 0.20 + 0.25 + 0.20 + 0.10 + 0.15 + 0.05 = 1$$

$$(ii) \quad P(x \geq 3) = \sum_{x=3}^6 p(x) \\ = P(3) + P(4) + P(4) + P(6) \\ = 0.20 + 0.10 + 0.15 + 0.05 = 0.50$$

- (c) (i) $p(x) > 0$, for all x , and

$$\sum_{x=1}^3 p(x) = \frac{1}{21} \sum_{x=1}^3 (2x+3) = \frac{1}{21} \{2(1) + 3 + 2(2) + 3 + 2(3) + 3\} \\ = \frac{1}{21} (5 + 7 + 9) = 1$$

- (ii) We determine k by assuming $p(x)$ is probability mass function,

$$\sum_{x=3}^5 p(x) = \sum_{x=3}^5 k(x-1) = 1$$

$$k(x-1) = k \{(3-1) + (4-1) + (5-1)\} = 1$$

$$9k = 1 \Leftrightarrow k = \frac{1}{9}$$

4.2(d) (i) Let x be a continuous random variable with probability density function,

$$f(x) = \begin{cases} \frac{1}{6}x + k, & 0 \leq x \leq 3 \\ 0, & \text{elsewhere} \end{cases}$$

Evaluate k and hence find $P(1 \leq x \leq 2)$

(ii) Determine the value of k and hence compute the probabilities, $P(1 \leq x \leq 2)$ and $P(x > 2)$.

$$f(x) = \begin{cases} kx, & 0 \leq x \leq 3, k > 0 \\ 3k(4-x), & 3 < x \leq 4 \\ 0, & \text{otherwise} \end{cases}$$

Solution:

(i) Given the probability density function,

$$f(x) = \begin{cases} \frac{1}{6}x + k, & 0 \leq x \leq 3 \\ 0, & \text{elsewhere} \end{cases}$$

then,

$$\begin{aligned} \text{(i)} \quad \int_0^3 f(x)dx &= 1 \\ \int_0^3 \left(\frac{1}{6}x + k\right)dx &= 1 \\ \frac{1}{12}x^2 + kx \Big|_0^3 &= 1 \\ \left[\frac{1}{12}(3)^2 + 3k\right]_0^3 - 0 &= 1 \\ \frac{3}{4} + 3k &= 1 \\ 3k &= \frac{1}{4} \Leftrightarrow k = \frac{1}{12} \end{aligned}$$

$$\text{Hence, } f(x) = \begin{cases} \frac{1}{12}(2x+1), & 0 \leq x \leq 3 \\ 0, & \text{elsewhere} \end{cases}$$

$$\begin{aligned} P(1 \leq x \leq 2) &= \int_1^2 \frac{1}{12}(2x+1) dx \\ &= \frac{1}{12} \left[x^2 + x \right]_1^2 \\ &= \frac{1}{12} [(2^2 + 2) - (1^2 + 1)] \\ &= \frac{1}{12} (6 - 2) = \frac{1}{3} \end{aligned}$$

- (ii) For $f(x)$ is probability density function, $f(x) \geq 0$ for all values of x and $k > 0$. We also show that,

$$\int_0^4 f(x)dx = 1$$

$$\int_0^3 kx dx + \int_3^4 3k(4-x) dx = 1$$

$$\left(\frac{kx^2}{2}\right)_0^3 + 3k\left(4x - \frac{x^2}{2}\right)_3^4 = 1$$

$$\frac{9k}{2} + 3k [(16-8) - (12-\frac{9}{2})] = 1$$

$$\frac{9k}{2} + \frac{3k}{2} = 1$$

$$6k = 1$$

$$k = \frac{1}{6}$$

$$\text{Hence, } f(x) = \begin{cases} \frac{1}{6}x & , 0 \leq x \leq 3 \\ \frac{1}{2}(4-x), & 3 < x \leq 4 \\ 0 & , \text{ elsewhere} \end{cases}$$

$$P(1 \leq x \leq 2) = \int_1^2 f(x)dx$$

$$= \int_1^2 \frac{1}{6}x dx = \left| \frac{x^2}{12} \right|_1^2$$

$$= \frac{1}{12}(2^2 - 1) = \frac{1}{4}$$

$$P(x > 2) = \int_2^4 f(x)dx$$

$$= \int_2^3 \frac{1}{6}x dx + \int_3^4 \frac{1}{2}(4-x) dx$$

$$= \left| \frac{x^2}{12} \right|_2^3 + \frac{1}{2} \left| 4x - \frac{x^2}{2} \right|_3^4$$

$$= \frac{1}{12}(9-4) + \frac{1}{2}(16-8) - \frac{1}{2}(12-\frac{9}{2})$$

$$= \frac{5}{12} + \frac{1}{4}$$

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

4.2(e) Given the probability mass function,

x	0	1	2	3
$p(x)$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{8}$

Find the cumulative distribution function.

Solution:

$$(a) \quad F(x) = P(X \leq x) = \sum_{x=0}^3 p(x)$$

$$F(0) = P(X \leq 0) = p(0) = \frac{1}{4}$$

$$F(1) = P(X \leq 1) = p(0) + p(1) = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}$$

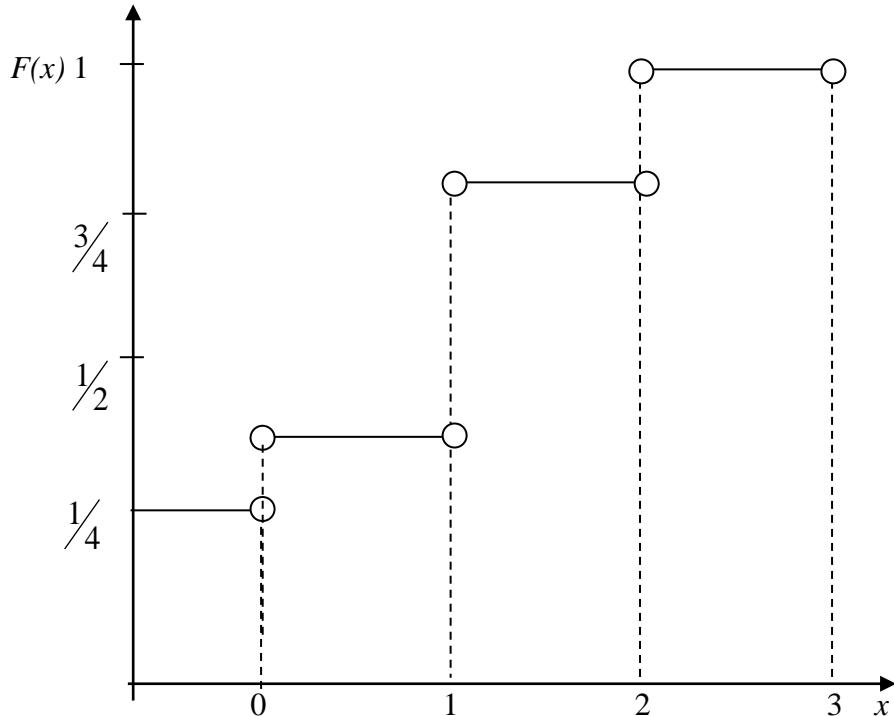
$$F(2) = P(X \leq 2) = p(0) + p(1) + p(2) = \frac{1}{4} + \frac{1}{8} + \frac{1}{2} = \frac{7}{8}$$

$$F(3) = P(X \leq 3) = p(0) + p(1) + p(2) + p(3) = \frac{1}{4} + \frac{1}{8} + \frac{1}{2} + \frac{1}{8} = 1$$

Hence the cumulative distribution is

X	0	1	2	3
$F(x)$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{7}{8}$	1

which is illustrated graphically below:



4.2(f) For a discrete random variable x with cumulative distribution function is given by $F(x) = kx$, $x = 1, 2, 3$,

- (i) Find the value of the constant k , and
- (ii) Evaluate the probability $P(x < 3)$. Hence determine the probability distribution of x .

Solution

- (i) From the properties of cumulative distribution function,

$$F(3) = P(x \leq 3) = 1$$

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

$$\text{Hence, } F(x) = \frac{1}{3}x \text{ for } 1, 2, 3$$

(ii) $P(x < 3) = F(2) = \frac{1}{3}(2) = \frac{2}{3}$

The probability distribution function of x , $p(x)$ is obtained as follows:

$$P(x = 1) = F(1) - F(0) = \frac{1}{3} - 0 = \frac{1}{3} = p(1)$$

$$P(x = 2) = F(2) - F(1) = \frac{2}{3} - \frac{1}{3} = \frac{1}{3} = p(2)$$

$$P(x = 3) = F(3) - F(2) = 1 - \frac{2}{3} = \frac{1}{3} = p(3)$$

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

or $p(x) = \frac{1}{3}x$, for $x = 1, 2, 3$.

4.2(g) Find the cumulative distribution functions of the following probability density functions:

(i) $f(x) = \begin{cases} \frac{1}{2}x, & \text{for } 0 \leq x \leq 2 \\ 0, & \text{elsewhere} \end{cases}$ (iii) $f(x) = \begin{cases} \frac{1}{2}e^{-\frac{x}{2}}, & x > 0 \\ 0, & \text{elsewhere} \end{cases}$

(ii) $f(x) = \begin{cases} x, & 0 \leq x \leq 1 \\ \frac{1}{2}, & 1 < x \leq 2 \\ 0, & \text{elsewhere} \end{cases}$

Sketch the graph of $F(x)$ in each case.

Solution:

$$(i) \quad f(x) = \begin{cases} \frac{1}{2}x, & 0 \leq x \leq 2 \\ 0, & \text{elsewhere} \end{cases}$$

$$\begin{aligned} F(x) &= \int_0^x f(t) dt = \int_0^x \frac{1}{2}t dt \\ &= \frac{1}{4}x^2, \text{ for } 0 \leq x \leq 2 \\ &= 1, \text{ if } x > 2 \text{ and } 0, \text{ elsewhere.} \end{aligned}$$

$$\text{Hence, } F(x) = \begin{cases} \frac{1}{4}x^2, & 0 \leq x \leq 2 \\ 1, & x > 2 \\ 0, & x < 0 \end{cases}$$

$$(ii) \quad f(x) = \begin{cases} x, & 0 \leq x \leq 1 \\ \frac{1}{2}, & 1 < x \leq 2 \\ 0, & \text{elsewhere} \end{cases}$$

$$\begin{aligned} F(x) &= \int_{-\infty}^x f(t) dt \\ F_1(x) &= \int_0^x t dt = \frac{1}{2}x^2, \quad 0 \leq x \leq 1 \end{aligned}$$

$$F_2(x) = F_1(1) + \int_1^x \frac{1}{2} dt = \frac{1}{2}x, \quad 1 < x \leq 2$$

$$F_3(x) = F_2(2) + 0 = 1, \quad x > 2$$

$$F(x) = \begin{cases} \frac{1}{2}x^2, & 0 \leq x \leq 1 \\ \frac{1}{2}x, & 1 < x \leq 2 \\ 1, & x > 2 \\ 0, & \text{elsewhere} \end{cases}$$

$$(iii) \quad f(x) = \begin{cases} \frac{1}{2}e^{-\frac{x}{2}}, & x > 0 \\ 0, & \text{elsewhere} \end{cases}$$

$$\begin{aligned} F(x) &= \int_0^x f(t) dt \\ &= \int_0^x \frac{1}{2}e^{-\frac{t}{2}} dt \\ &= \begin{cases} 1 - e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leq 0 \end{cases} \end{aligned}$$

SESSION 2-4: EXPECTATION OF RANDOM VARIABLES

2-4.1 Expected Value and Variance of Random Variables

- The *expectation* or *expected value* (or simply the *mean*) of the random variable, x is defined by

$$(i) \quad \mu = E(x) = \sum_x x p(x), \text{ if } x \text{ is discrete.}$$

$$(ii) \quad \mu = E(x) = \int_{-\infty}^{\infty} x f(x) dx, \text{ if } x \text{ is continuous and } -\infty \leq x \leq \infty.$$

- The *variance* of the random variable, x with probability distribution, $p(x)$ or $f(x)$ is defined by

$$\sigma^2 = Var(x) = E[(x - \mu)^2] = E(x^2) - \mu^2, \text{ where}$$

$$(i) \quad Var(x) = \sum_x (x - \mu)^2 p(x) \\ = \sum_x x^2 p(x) - \mu^2, \text{ if } x \text{ is discrete.}$$

$$(ii) \quad Var(x) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \\ = \int_a^b x^2 f(x) dx - \mu^2, \text{ if } x \text{ is continuous.}$$

The *standard deviation* of x is the square root $\sqrt{Var(x)}$. That is,

$$\sigma = \sqrt{Var(x)}$$

2-4.2 Median and Mode of Probability Distributions

- The median of a distribution of the random variable x is that value of $x = m$ such that $P(x \leq m)$ or $P(x \geq m) = 0.5$ or close to it. The *median* of the random variable, x with probability distribution, $p(x)$ or $f(x)$ is obtained by the equation:

$$(i) \quad \sum_x^m p(x) = \frac{1}{2} \text{ (or close to it), if } x \text{ is discrete.}$$

$$(ii) \quad \int_a^m f(x) dx = \frac{1}{2}, \text{ if } x \text{ is continuous and is such that } a \leq x \leq b.$$

- The *mode* of a distribution of random variable x is that value of $x = m_0$ that maximizes the probability distribution function, $p(x)$ or $f(x)$. If there is only one such x , it is called the *mode* of the distribution.

Example 4.3:

4.3(a) Compute the expected value (μ) and standard deviation (σ^2) of the random variable, x with the following probability distribution:

(i)

x	1	2	3	4	5
$p(x)$	0.1	0.3	0.2	0.3	0.1

(ii) $f(x) = \begin{cases} 6x(1-x), & 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$

Solution:

(i) Given the discrete probability distribution,

x	1	2	3	4	5
$p(x)$	0.1	0.3	0.2	0.3	0.1

The expected value of x or mean,

$$\mu = \sum_{x=1}^5 x p(x) = 1(0.1) + 2(0.3) + 3(0.2) + 4(0.3) + 5(0.1) = 3.0$$

The variance of x ,

$$\begin{aligned} Var(x) &= \sum_{x=1}^5 (x - \mu)^2 p(x) = \sigma^2 \\ &= (1 - 3)^2(0.1) + (2 - 3)^2(0.3) + (3 - 3)^2(0.2) + (4 - 3)^2(0.3) \\ &\quad + (5 - 3)^2(0.1) = 0.4 + 0.3 + 0 + 0.3 + 0.4 = 1.4, \text{ or} \end{aligned}$$

$$\begin{aligned} Var(x) &= \sum_{x=1}^5 x^2 p(x) - \mu^2 = \sigma^2 \\ &= 1^2(0.1) + 2^2(0.3) + 3^2(0.2) + 4^2(0.3) + 5^2(0.1) - (3)^2 \\ &= 0.1 + 1.2 + 1.8 + 4.8 + 2.5 - 9 = 1.4 \end{aligned}$$

Hence the standard deviation, $\sigma = \sqrt{1.4} = 1.1832$

(ii) Given the probability density function,

$$f(x) = \begin{cases} 6x(1-x), & 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$$

The mean of x is $\mu = E(x) = \int_0^1 x f(x) dx$, where

$$\begin{aligned}
\int_0^1 xf(x) dx &= \int_0^1 6x^2(1-x) dx \\
&= \int_0^1 (6x^2 - 6x^3) dx \\
&= \left[\frac{6}{3}x^3 - \frac{6}{4}x^4 \right]_0^1 \\
&= 2(1)^3 - \frac{3}{2}(1)^4 - 0 \\
&= 2 - \frac{3}{2} = \frac{1}{2} = 0.5
\end{aligned}$$

The variance of x ,

$$\begin{aligned}
\sigma^2 &= Var(x) = E(x^2) - \mu^2 \\
&= \int_0^1 x^2 f(x) dx - \mu^2 \\
&= \int_0^1 6x^3(1-x) dx - (0.5)^2 \\
&= \int_0^1 (6x^3 - 6x^4) dx - 0.25 \\
&= \left[\frac{6}{4}x^4 - \frac{6}{5}x^5 \right]_0^1 - 0.25 \\
&= \frac{3}{2} - \frac{6}{5} - 0.25 \\
&= \frac{3}{10} - \frac{1}{4} = \frac{1}{20} = 0.05
\end{aligned}$$

Hence the standard deviation, $\sigma = \sqrt{0.05} = 0.224$

- 4.3(b)** The weekly demand x for kerosene at a certain supply station has the probability distribution,

$$f(x) = \begin{cases} x & , 0 \leq x \leq 1 \\ \frac{1}{2} & , 1 < x \leq 2 \\ 0 & , \text{otherwise} \end{cases}$$

- (i) Determine the mean, median and the variance.
(ii) Find the mean and median of the distribution,

$$p(x) = \begin{cases} \frac{1}{9}(x-1), & x = 3, 4, 5 \\ 0 & , \text{elsewhere} \end{cases}$$

Solution:

$$(i) \quad \text{Given } f(x) = \begin{cases} x, & 0 \leq x \leq 1 \\ \frac{1}{2}, & 1 < x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

$$\begin{aligned}\mu = E(x) &= \int_0^2 x f(x) dx \\ &= \int_0^1 x^2 dx + \int_1^2 \frac{1}{2} x dx \\ &= \left[\frac{1}{3} x^3 \right]_0^1 + \left[\frac{1}{4} x^2 \right]_1^2 \\ &= \frac{1}{3}(1)^3 + \frac{1}{4}(2)^2 - \frac{1}{4}(1)^2 \\ &= \frac{1}{3} + 1 - \frac{1}{4} \\ &= \frac{13}{12} = 1.083\end{aligned}$$

Let m be the median. Then,

$$\begin{aligned}\int_0^1 x dx &= \frac{1}{2} x^2 \Big|_0^1 = \frac{1}{2} \\ \int_1^m \frac{1}{2} dx &= \frac{1}{2} - \int_0^1 x dx \\ \frac{1}{2} x \Big|_1^m &= \frac{1}{2} - \frac{1}{2} \\ \frac{1}{2} m - \frac{1}{2} &= 0 \Leftrightarrow m = 1\end{aligned}$$

The variance, $Var(x)$

$$\begin{aligned}Var(x) &= E(x^2) - \mu^2 \\ &= \int_0^2 x^2 f(x) dx - \mu^2 \\ &= \int_0^1 x^3 dx + \int_1^2 \frac{1}{2} dx - \left(\frac{13}{12} \right)^2 \\ &= \frac{1}{4} x^4 \Big|_0^1 + \frac{1}{2} x \Big|_1^2 - \frac{169}{144} \\ &= \frac{1}{4} + \frac{8}{6} - \frac{1}{6} - \frac{169}{144} \\ &= \frac{35}{144} = 0.243\end{aligned}$$

$$(ii) \quad p(x) = \begin{cases} \frac{1}{9}(x-1), & x = 3, 4, 5 \\ 0, & \text{otherwise} \end{cases}$$

$$\begin{aligned}\mu = E(x) &= \sum_{x=3}^5 x p(x) \Sigma x p(x) \\ &= \frac{2}{9}(3-1) + \frac{4}{9}(4-1) + \frac{5}{9}(5-1) \\ &= \frac{1}{9}\{6+12+20\} = \frac{38}{9} = 4.22\end{aligned}$$

For the median, m we have $P(x \leq m) \leq 0.5$ or ≥ 0.5

$$\begin{aligned}P(x \leq 4) &= p(3) + p(4) \quad \text{or} \quad P(x \geq 5) = p(5) \\ &= \frac{1}{9}(2) + \frac{1}{9}(3) \quad \text{or} \quad = \frac{1}{9}(4) \\ &= \frac{5}{9} = 0.56 \quad \text{or} \quad = \frac{4}{9} = 0.44\end{aligned}$$

Hence $m = 4$ since $P(x \leq 4)$ closer to 0.5 than $P(x \geq 4)$.

4.3(c) Let y have the probability distribution,

$$f(y) = \begin{cases} y, & 0 \leq y < \frac{1}{2} \\ \lambda(4-y), & \frac{1}{2} \leq y \leq 4 \\ 0, & \text{elsewhere} \end{cases}$$

- (i) Find the value of λ , and
- (ii) Use it to determine the mean, median and the standard deviation.

Solution:

- (i) To find λ we have,

$$\int_0^4 f(y) dy = 1$$

$$\int_0^{\frac{1}{2}} y dy + \lambda \int_{\frac{1}{2}}^4 (4-y) dy = 1$$

$$\frac{1}{2}y^2 \Big|_0^{\frac{1}{2}} + \lambda \left[4y - \frac{1}{2}y^2 \right]_{\frac{1}{2}}^4 = 1$$

$$\frac{1}{8} + \lambda \left\{ \left[4(4) - \frac{1}{2}(4)^2 \right] - \left[4\left(\frac{1}{2}\right) - \frac{1}{2}\left(\frac{1}{2}\right)^2 \right] \right\} = \frac{1}{8} + \lambda \left[(16-8) - \left(2 - \frac{1}{8}\right) \right] = 1$$

$$\frac{49}{8}\lambda = \frac{7}{8} \Leftrightarrow \lambda = \frac{1}{7}$$

$$\text{Hence, } f(y) = \begin{cases} y & , 0 \leq y < \frac{1}{2} \\ \frac{1}{7}(y - y) & , \frac{1}{2} \leq y \leq 4 \\ 0 & , \text{ elsewhere} \end{cases}$$

$$\begin{aligned} \text{(ii)} \quad \mu &= E(y) = \int_0^{\frac{1}{2}} y^2 dy + \int_{\frac{1}{2}}^4 \frac{1}{7} y (4-y) dy \\ &= \frac{1}{3} y^3 \Big|_0^{\frac{1}{2}} + \int_{\frac{1}{2}}^4 \frac{1}{7} (4y - y^2) dy \\ &= \frac{1}{24} + \frac{1}{7} \left[\left(2y^2 - \frac{1}{3} y^3 \right) \Big|_{\frac{1}{2}}^4 \right] \\ &= \frac{1}{24} + \frac{1}{7} \left[\left(32 - \frac{64}{3} \right) - \left(\frac{1}{2} - \frac{1}{24} \right) \right] \\ &= \frac{256}{168} = \frac{3}{2} = 1.5 \end{aligned}$$

Let m be the median. Then

$$\begin{aligned} \int_0^{\frac{1}{2}} y dy &= \left[\frac{1}{2} y^2 \right]_0^{\frac{1}{2}} = \frac{1}{8} \\ \int_0^{\frac{1}{2}} y dy + \int_{\frac{1}{2}}^m \frac{1}{7} (4-y) dy &= \frac{1}{2} \\ \int_{\frac{1}{2}}^m \frac{1}{7} (4-y) dy &= \frac{1}{2} - \frac{1}{8} \\ \frac{1}{7} \left[4y - \frac{1}{2} y^2 \right]_{\frac{1}{2}}^m &= \frac{3}{8} \\ \frac{1}{7} \left(4m - \frac{1}{2} m^2 \right) - \frac{1}{7} \left(2 - \frac{1}{8} \right) &= \frac{3}{8} \\ 4m - \frac{1}{2} m^2 &= 9 \Leftrightarrow m^2 - 8m + 9 = 0 \end{aligned}$$

Solving we have $m = 1.35$.

For the standard deviation, σ , we have

$$\begin{aligned} \sigma^2 &= \text{Var}(x) = E(x^2) - \mu^2 \\ &= \int_0^{y_2} y^3 dy + \int_{\frac{1}{2}}^4 \frac{1}{7} y^2 (4-y) dy - (1.5)^2 \\ &= \left[\frac{1}{4} y^4 \right]_0^{\frac{1}{2}} + \frac{1}{7} \left[\left(\frac{4}{3} y^3 - \frac{1}{4} y^4 \right) \right]_{\frac{1}{2}}^4 - 2.25 \\ &= \frac{73}{24} - 2.25 = \frac{19}{24} = 0.79167 \end{aligned}$$

Hence, the standard deviation, $\sigma = \sqrt{0.79167} = 0.88976$

2-4.3 Trial Questions 1-4:

- 1.(a) (i) What is a random variable? State and distinguish between the two types of random variables.
(ii) Give five examples of each type of random variable in (i).
(b) The table below shows that probability distribution of the number of children per household (x).

x	0	1	2	3	4	5
$p(x)$	0.18	0.30	0.24	0.14	0.10	0.04

- (i) Find the expected value and the standard deviation of x .
(ii) If it costs ₦5,500 to feed a child per household, find the expected cost.
(c) Let y have the probability density function given by

$$f(y) = \begin{cases} ky^2 + y & , 0 \leq y \leq 2 \\ 0 & , \text{ elsewhere} \end{cases}$$

- (i) Find the value of k . Compute the mean value of y and its standard deviation.
(ii) Find the probability, $P(1 \leq y \leq 1.5)$.

- 2.(a) Find the mean and the standard deviation of the following distributions,

$$(i) p(x) = \begin{cases} \frac{k}{2}(x-2) & , x = 1, 2, 3, 4 \\ 0 & , \text{ elsewhere} \end{cases}$$

y	0	1	2	3	4
$p(y)$	0.20	0.30	0.25	0.15	0.10

where y is the number of sales per week.

- (b) The probability density of a random, y is given by

$$f(y) = \begin{cases} \lambda y^2(1-y) & , 0 \leq y \leq 1 \\ 0 & , \text{ elsewhere} \end{cases}$$

- (i) Find the value of λ and the standard deviation of y
(ii) Given that the median of y is m , show that $6m^4 - 8m^3 + 1 = 0$

3.(a) A random variable x has probability density function,

$$f(x) = \begin{cases} kx & , 0 \leq x \leq 2 \\ 2k(3-x) & , 2 < x \leq 3 \\ 0 & , elsewhere \end{cases}$$

where k is constant.

- (i) Determine the value of k and sketch the graph of $f(x)$.
 - (ii) What are the mean and median of the distribution?
 - (iii) Find the value of a such that $P(x > a) = 0.25$.
- (b) Given the random variable x with probability density function,

$$f(x) = \begin{cases} ke^{-0.001x}, & x > 0 \\ 0 & , elsewhere \end{cases}$$

Find the value of k , the mean of x and the probability, $P(x > 1,050)$.

(c) The continuous random variable x has the probability density function,

$$f(x) = \begin{cases} \frac{1}{2}(x-2), & 2 \leq x \leq 3 \\ a & , 3 < x \leq 5 \text{ and } a > 0 \\ 2-bx & , 5 < x \leq 6 \text{ and } b > 0 \\ 0 & , elsewhere \end{cases}$$

- (i) Find the values of a and b , and sketch the graph of $f(x)$.
- (ii) Find the cumulative distribution function, $F(x)$ and sketch it.

SPECIAL PROBABILITY DISTRIBUTIONS

We now consider some probability distribution functions which arise very often in practice, giving the description of the situations which call for their uses. The ones discussed in this unit serve good models for large number of applications in engineering, physical and social sciences, as well as in business and economics. They are classified into discrete and continuous. The detailed discussions of these special distributions which include their properties and applications are presented in the following sections.

Learning Objectives

The unit aims at studying several important types of discrete and continuous random variables and their probability models (or distributions). This will eventually enable students to:

- Identify situations and model them using the appropriate probability models such as the Binomial, Negative Binomial, Poisson, and Normal distributions.
- Compute probabilities of random variables of these special probability distributions.

SESSION 1-5: THE DISCRETE DISTRIBUTIONS

1-5.1 The Binomial Distribution

The *Binomial distribution* is a discrete probability distribution used to model experiments consisting of sequence of observations of identical and independent trials, each of which results in one of the two outcomes. Such experiments which are actually generalization of Bernoulli trials are called *Binomial Experiment*. A Binomial Experiment exhibits the following properties:

- The experiment consists of n independent and identical trials.
- Each trial results in one of the two outcomes called *success* and *failure*.
- The probability of success in a single trial is p and remains the same from trial to trial. The probability of a failure, also in a single trial is $q = 1 - p$.
- The random variable of interest, x is the number of successes observed during the n trials.

There are several situations which may result in a random variable that may be satisfied by the conditions of the Binomial Experiment.

Some examples of these situations are a random selection of items from a manufacturing process for inspection is either *defective* or *non-defective*, an opinion poll during electioneering campaign where each of n persons interviewed will either *vote* or *not vote* for a particular candidate, interviewing a random sample of n students to determine whether a policy being introduced by the university authorities is *favoured* or *not favoured*, a number of patients undergoing through a medical treatment may either come out *successfully* or *not successfully*, a sequence of n shots at a target may result in a number of *hits* or *misses*, tossing a coin n times and observing the number of successes, *heads* or *tails*.

- *Definition 5.1:*

A random variable, x is said to have a Binomial distribution based on n trials with probability of a success, p if and only if

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

n where $0 < p < 1$ and the mean and variance are respectively,

$$\mu = E(x) = np \text{ and } \sigma^2 = \text{Var}(x) = np(1-p)$$

The random variable, x with Binomial distribution is simply denoted as $x \sim B(n, p)$.

The following diagrams (Figure 5.1 and Figure 5, 2) illustrate graphically the probability histograms of Binomial distribution.

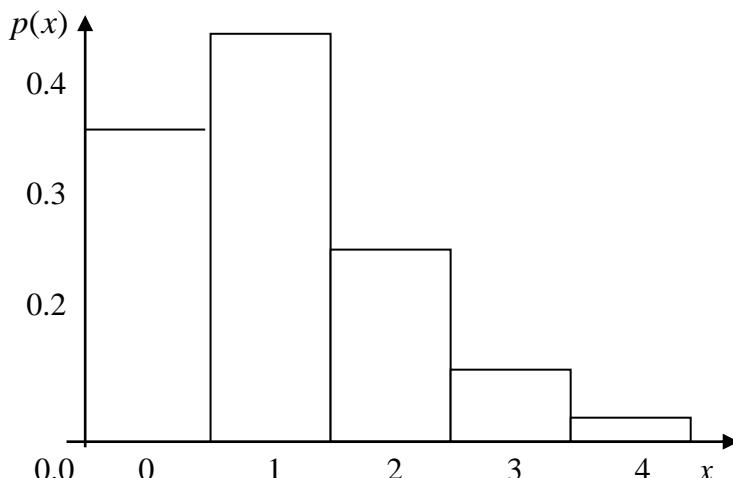
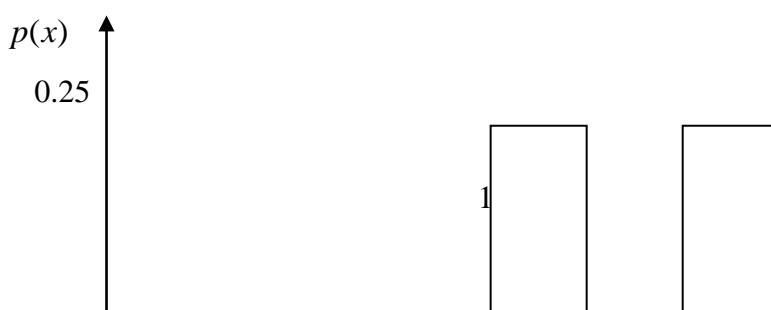


Figure 5.1: Binomial distribution for $n=10$ and $p=0.10$



0.20

0.15

0.10

0.05

0.0

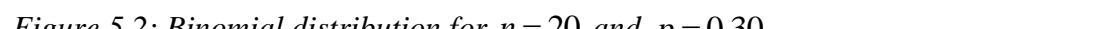


Figure 5.2: Binomial distribution for $n=20$ and $p=0.30$

$p(x)$

0.20

0.15

0.10

0.05

0.0

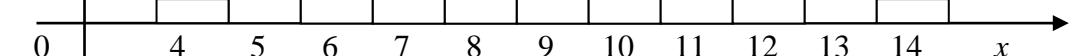


Figure 5.3: Binomial distribution for $n=50$, $p=0.50$

The term *binomial experiment* is derived from the fact that the probabilities, $p(x)$ at $x = 0, 1, 2, \dots, n$ are terms of the binomial expansion,

$$(p+q)^n = \sum_{x=0}^n \binom{n}{x} p^x q^{n-x}$$

$$\text{where } \sum_{x=0}^n p(x) = \sum_{x=0}^n \binom{n}{x} p^x q^{n-x} = (p+q)^n = 1$$

The generalization of the Binomial distribution is Multinomial distribution which arises when each trial of the experiment has more than two possible outcomes of the probabilities of the respective outcomes are the same for each trial. These two probability distributions have many applications because the binomial and multinomial

experiments occurs in sampling for defectives in industrial quality control, sampling of consumer preference for products, voting intentions in an opinion polls and in many other physical situations.

Example 5.1:

5.1(a) It is known that 25% of inhabitants of a community favour a political party A .

A random sample of 20 inhabitants was selected from the community and each person was asked he/she will vote for party A in an impending election. What is the probability that:

- (i) exactly two persons will vote for party A ?
- (ii) at least three persons will vote for party A ?
- (iii) fewer than two persons will vote for party A ?

5.1(b) A multiple-choice test consists of 15 questions each with five possible answers of which only one is correct. Suppose one of the students taking the test answers the questions by guessing. What is the probability that he answers at most 3 questions correctly?

Solution:

(a) The experiment here is testing the 20 inhabitants of the community to find whether they have blood group A . This satisfies the binomial experiment requirements, where $n = 20$, $p = 0.25$, $1 - p = 0.75$, and $x = \text{number of persons}$

with blood group A . Thus $p(x) = \binom{20}{x} (0.25)^x (0.75)^{20-x}$, $x = 0, 1, 2, \dots, 20$.

$$(i) \quad p(2) = \binom{20}{2} (0.25)^2 (0.75)^{18} = 190(0.0625)(0.00563771) = 0.066947808$$

$$(ii) \quad P(x \geq 3) = 1 - P(x \leq 2)$$

$$\begin{aligned} &= 1 - \sum_{x=0}^2 \binom{20}{x} (0.25)^x (0.75)^{20-x} \\ &= 1 - \left\{ \binom{20}{0} (0.25)^0 (0.75)^{20} + \binom{20}{1} (0.25)^1 (0.75)^{19} + \binom{20}{2} (0.25)^2 (0.75)^{18} \right\} \\ &= 1 - 0.09126043 = 0.90873957 \end{aligned}$$

$$\begin{aligned} (iii) \quad P(x < 2) &= \sum_{x=0}^1 \binom{20}{x} (0.25)^x (0.75)^{20-x} \\ &= \binom{20}{0} (0.25)^0 (0.75)^{20} + \binom{20}{1} (0.25)^1 (0.75)^{19} = 0.024312625 \end{aligned}$$

- (b) The number of correct answers, $x \square B\left(15, \frac{1}{5}\right)$. Then required probability is

$$\begin{aligned}
 P(x \leq 3) &= \sum_{x=0}^3 \binom{15}{x} \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{15-x} \\
 &= \binom{15}{0} \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{15} + \binom{15}{1} \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^{14} + \binom{15}{2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^{13} \binom{15}{3} \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^{12} \\
 &= \left(\frac{4}{5}\right)^{15} + 15 \left(\frac{1}{5}\right) \left(\frac{4}{5}\right)^{14} + 105 \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^{13} 455 \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^{12} \\
 &= 0.03518 + 0.13194 + 0.23090 + 0.25014 \\
 &= 0.64816
 \end{aligned}$$

- 5.1(c)** A large retailer purchases a certain kind of product from a manufacturer. The manufacturer indicates that the defective rate of the product is 3% in a shipment. The inspector of the retailer randomly picks 20 items of the product from a shipment.
- (i) What is the probability that there will be (3) defective items?
 - (ii) What is the probability that there will not be more than two (2) defective items?

- 5.1(d)** An oil exploration firm is formed with enough capital to finance 5 explorations. The equipment of particular exploration being successful is 0.15. Assume that the explorations are independent and conform to the properties of Binomial experiment.
- (i) Find the expected number of unsuccessful explorations and its variance.
 - (ii) Suppose the firm has a fixed cost of \$10,000 and that it costs \$25,000 to make a successful exploration. Find the expected total cost in the 5 explorations.
 - (iii) What is the probability that no or one successful exploration is made?

Solution:

- (c) Let the number of defective devices among the 20 randomly selected from a shipment be y and so $y \square B(n = 20, p = 0.03)$.
- (i) The required probability becomes

$$P(y=3) = \binom{20}{3} (0.03)^3 (0.97)^{20-3}$$

$$= 1,140 (0.03)^3 (0.97)^{17} = 0.018339526$$

- (ii) The required probability becomes

$$P(y \leq 2) = \sum_{y=0}^2 \binom{20}{y} (0.03)^y (0.97)^{20-y}$$

$$= \binom{20}{0} (0.03)^0 (0.97)^{20} + \binom{20}{1} (0.03)^1 (0.97)^{19} + \binom{20}{2} (0.03)^2 (0.97)^{18}$$

$$= (0.97)^{20} + 20(0.03)(0.97)^{19} + 190(0.0009)(0.97)^{18}$$

$$= 978991643$$

- (d) Let y be the number of successful explorations. Then given $n=5$ and probability of a successful exploration, $p=0.15$, we have

- (i) The expected number of unsuccessful explorations, x and its variance are:

$$E(x) = n(1-p) = 5(0.85) = 4.25 \approx 4$$

$$Var(x) = n(1-p)p = 5(0.85)(0.15) = 0.6375$$

- (ii) The total cost for making y explorations is given by

$$C = 10,000 + 25,000y, \text{ (in dollars), where the expected cost is}$$

$$E(C) = E(10,000 + 25,000y)$$

$$= 10,000 + 25,000E(y)$$

$$= 10,000 + 25,000(5)(0.15) = \$28,750$$

- (iii) The probability that 0 or 1 successful exploration is made,

$$P(x=0 \text{ or } x=1) = P(x=0) + P(x=1)$$

$$= \binom{5}{0} (0.15)^0 (0.85)^5 + \binom{5}{1} (0.15)^1 (0.85)^4$$

$$= (0.85)^5 + 5(0.15)(0.85)^4 = 0.8352$$

$$= (0.85)^5 + 5(0.15)(0.85)^4 = 0.4437 + 0.3915 = 0.8352$$

1-5.2 The Negative Binomial and Geometric Distributions

The binomial random variable is a count of number of successes in a series of n Bernoulli trials. The number of trials is fixed (predetermined) and the number of successes varies randomly from experiment to experiment. The negative binomial

random variable is a count of the number of trials required to obtain k successes. The number of successes here is fixed and the number of trials varies randomly. In this sense, the negative binomial random variable is considered as the opposite or negative of the binomial random variable. In particular, the negative binomial random variable arises in situations characterized by these properties:

- The experiment consists of a series of independent and identical Bernoulli trials, each with probability of successes, p .
- The trials are observed until the k^{th} success is obtained (k is fixed by the experimenter)
- The random variable y is defined as the number of identical and independent trials required to obtain k successes.

Typical of such situations are the number of job applicants interviewed until the k^{th} suitable applicant is found, the number of oil will needed to be drilled until the k^{th} successful oil well is hit, the number of shots fired before the first, second or third target was hit, the number of pregnancies required before the fifth girl-child is born.

- *Definition 5.2:*

(i) *The random variable y (the number of identical and independent trials required to obtain k successes) has probability distribution called the negative binomial distribution is defined by:*

$$p(y) = \binom{y-1}{k-1} p^k (1-p)^{y-k}, \quad y = k, k+1, k+2, \dots ; 0 < p < 1$$

where mean and variance given as follows: $E(y) = \frac{k}{p}$ and $Var(y) = \frac{k(1-p)}{p^2}$

(ii) *A special type of Negative Binomial distribution is Geometric distribution where the random variable y is defined as the number of identical and independent trials the experiment is performed until the first success occurs ($k=1$). The probability distribution is defined by*

$$p(y) = p(1-p)^{y-1}, \quad y = 1, 2, 3, \dots ; 0 < p < 1$$

where mean and variance become, $E(y) = \frac{1}{p}$ and $Var(y) = \frac{1-p}{p^2}$

Example 5.2

5.2. (a) A geological study indicates that an exploratory oil well drilled in a certain part of a state strikes oil with probability of 0.20.

- (i) Find the probability that the first strike of oil comes on the third well drilled.
- (ii) Find the probability that the third strike of oil comes on the sixth well drilled.
- (iii) Find the mean and variance of the number of wells that must be drilled if the company wants to set up three producing wells.

5.2(b) Ten percent of engines manufactured on an assembly line are defective.

If engines are randomly and independently selected one at a time and tested, what is the probability that

- (i) The first non-defective engine is found on the third trial;
- (ii) The third non-defective engine is found on or before the fifth trial.

Solution

(a) The probability of striking an oil well, $p = 0.2$

- (i) The probability of striking first oil on third well drilled,

$$p(y=3) = p(1-p)^{y-1} \\ = 0.2(0.8)^{3-1} = 0.2(0.8)^2 = 0.128$$

- (ii) The probability of striking third oil on the sixth well drilled

$$P(y=6) = \binom{y-1}{k-1} p^k (1-p)^{y-k}, \text{ where } k = 3 \\ = \binom{6-1}{3-1} (0.2)^3 (0.8)^3 = 10(0.2)^3 (0.8)^3 = 0.04096$$

(b) (i) The use of Geometric distribution,

$$p(x=3) = 0.9(0.1)^{3-1} = 0.009$$

- (ii) The use of Negative Binomial distribution,

$$P(x=6) = \binom{6-1}{3-1} \cdot (0.9)^3 (0.1)^3 = 0.00729$$

1-5.3 The Poisson Distribution

1-5.3.1 Introduction

The Poisson probability distribution provides a good model for a discrete random variable which results from an experiment called *Poisson Process*. The Poisson Process is characterised by the following assumptions or properties.

- The experiment consists of counting the number of times a particular event occurs during a given unit of time, area or volume.
- The occurrence or non-occurrence of the event in any interval of time space or volume is random and independent of the occurrence or non-occurrence of the event in any other interval.
- An infinite number of occurrences of event must be possible in the interval. Also in any infinitesimally small portion of the interval the probability of more than one occurrence of the event is negligible.
- The probability of the occurrence of an event in a given interval is proportional to the length of the interval.
- The mean number of events in an interval, denoted μ , is equal to its variance.

Typical examples of some experiments which may result a random variable that can be modelled by the Poisson process are the number of industrial accidents during a given period of time, number of flaws or defects on a square metre piece of material, number of radioactive or particles that decayed or emitted in a particular period of time, number of errors a typist makes in typing a page of a text, number of insurance claims received by an insurance company during a period of time, number of repairs on a kilometre of a road and number of telephone calls received by a switchboard in a unit time interval.

In each of the above examples the random variable represents the number of events occurring in a unit period of time or space during which the mean number of events, μ is expected to occur.

- *Definition 5.6:*

The Poisson distribution for the random variable, x , representing the number of occurrence of an event in a given interval of time, space or volume is defined by

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, 2, \dots, \text{and } \mu > 0,$$

where the mean and variance are the same. That is, $E(x) = \mu = \text{Var}(x)$.

The distribution of x may simply be denoted as $x \square P(\mu)$.

1-5.3.2 Poisson Approximation to Binomial

When the number of trials, n in a Binomial process is large, the computations of the binomial probabilities may be too tedious. The Poisson distribution can be used, as an alternative, to approximate the Binomial distribution. This is based on the convergence of the Binomial distribution as n becomes large ($n \rightarrow \infty$) as illustrated by the following theorem.

- Theorem 5.1:

Let μ be a fixed number and n an arbitrary positive integer. For each non-negative

$$\text{integer } x, \lim_{n \rightarrow \infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{\mu^x e^{-\mu}}{x!}, \text{ where } p = \frac{\mu}{n},$$

The tables below give an indication of the rate at which the Binomial distribution, converges to the Poisson distribution, where $\mu = np = 1$ in both cases.

(i) *Table 4.1: $n = 5, p = 0.2$ and so $\mu = np = 5(0.2) = 1$*

x	$\text{Binomial, } p(x) = \binom{5}{x} (0.2)^x (0.8)^{5-x}$	$\text{Poisson, } p(x) = \frac{1^x e^{-1}}{x!}$
0	0.328	0.368
1	0.410	0.368
2	0.205	0.184
3	0.051	0.061
4	0.006	0.015
5	0.000	0.003
6+	0.000	0.001

(ii) *Table 4.2: $n = 100, p = 0.01$ and so $\mu = np = 100(0.01) = 1$*

x	$\text{Binomial, } p(x) = \binom{100}{x} (0.01)^x (0.99)^{100-x}$	$\text{Poisson, } p(x) = \frac{1^x e^{-1}}{x!}$
0	0.366032	0.367879
1	0.369730	0.367879
2	0.184865	0.183940
3	0.060999	0.061313
4	0.014942	0.015328
5	0.002898	0.003066
6	0.000463	0.000511
7	0.000063	0.000073
8	0.000007	0.000009
9	0.000001	0.000001

We note from Table 4.1 (where $n = 5$) that for some x the agreement between the Binomial probability and the Poisson approximation is not very good. If n is large as 100, (as indicated in Table 4.2), the agreement is remarkable very good for all x .

Example 5.3:

5.3(a) The number of telephone calls received by an office averages 4 per minute.

Find the probability that:

- (i) No call will arrive in a given one-minute period.
- (ii) At least two calls will arrive in a given one-minute period.
- (iii) At least three calls will arrive in a given two-minute period.

5.3(b) The number of serious accidents, y in a manufacturing plant has approximately a Poisson distribution with a mean of 1.5 accidents per month. What is the probability that:

- (i) More than three accidents will occur within a period of one month?
- (ii) Fewer than three accidents will occur within a period six weeks?

Solution:

(a) The random variable, x is the number of telephone calls received with an average of $\mu = 4$ per minute. Hence x will have the Poisson distribution,

$$p(x) = \frac{4^x e^{-4}}{x!}, x = 0, 1, 2, \dots$$

$$(i) P(x = 0) = p(0) = \frac{4^0 e^{-4}}{0!} = e^{-4} = 0.018316$$

$$\begin{aligned} (ii) P(x \geq 2) &= 1 - P(x \leq 1) \\ &= 1 - \{P(x = 0) + P(x = 1)\} \\ &= 1 - \left(\frac{4^0 e^{-4}}{0!} + \frac{4^1 e^{-4}}{1!} \right) \\ &= 1 - (e^{-4} + 4e^{-4}) = 1 - 5e^{-4} \approx 0.908422 \end{aligned}$$

$$(iii) \mu = 4 \times 2 = 8 \text{ Calls received in two minutes.}$$

$$\begin{aligned} P(x \geq 3) &= 1 - P(x \leq 2) \\ &= 1 - \{P(x = 0) + P(x = 1) + P(x = 2)\} \\ &= 1 - \left(\frac{8^0 e^{-8}}{0!} + \frac{8^1 e^{-8}}{1!} + \frac{8^2 e^{-8}}{2!} \right) \\ &= 1 - (e^{-8} + 8e^{-8} + 32e^{-8}) = 1 - 41e^{-8} \approx 0.986246 \end{aligned}$$

- (b)** Given that y approximately have the Poisson distribution,

$$p(y) = \frac{1.5^y e^{-1.5}}{y!}, y = 0, 1, 2, \dots, \text{then}$$

$$\begin{aligned} \text{(i)} \quad P(y > 3) &= 1 - P(y \leq 2) = 1 - \sum_{y=0}^2 \frac{1.5^y e^{-1.5}}{y!} \\ &= 1 - \left(\frac{1.5^0 e^{-1.5}}{0!} + \frac{1.5^1 e^{-1.5}}{1!} + \frac{1.5^2 e^{-1.5}}{2!} \right) \\ &= 1 - (e^{-1.5} + 1.5 e^{-1.5} + 1.125 e^{-1.5}) \\ &= 1 - 3.25 e^{-1.5} \approx 0.191153 \end{aligned}$$

$$\text{(ii)} \quad \mu = 1.5 \times \frac{3}{2} = 2.25 \text{ accidents in six weeks,}$$

$$\begin{aligned} P(y < 3) &= \sum_{y=0}^2 \frac{2.25^y e^{-2.25}}{y!} \\ &= \frac{2.25^0 e^{-2.25}}{0!} + \frac{2.25^1 e^{-2.25}}{1!} + \frac{2.25^2 e^{-2.25}}{2!} \\ &= e^{-2.25} + 2.25 e^{-2.25} + 2.53125 e^{-2.25} \\ &= 5.78125 e^{-2.25} \approx 0.609339 \end{aligned}$$

- 5.3.(c)** The probability that a car will breakdown after falling into a pot-hole on a road is 0.00015. If 20,000 cars travel along the road, find the expected number of break-downs and probability that at least one car will break down.

- 5.3(d)** A book has 300 pages and the probability of finding misprints, x , in a page is 0.015. Find the probability of detecting misprints in at most one page of the book using the *Binomial distribution* and *Poisson Approximation to the Binomial*.

Solution:

- (c)** (i) The expected number of break-downs,

$$\mu = np = 20,000 (0.00015) = 3$$

- (ii) The probability that at least one car will break down, using the Poisson Approximation to Binomial,

$$\begin{aligned} P(x \geq 1) &= 1 - P(x = 0) \\ &= 1 - \frac{3^0 e^{-3}}{0!} = 1 - e^{-3} = 0.950213 \end{aligned}$$

- (d) (i) By the Binomial distribution, $n = 300$ and $p = 0.015$. The required probability,

$$\begin{aligned}
 P(x \leq 1) &= P(x = 0) + P(x = 1) \\
 &= \binom{300}{0} (0.015)^0 (0.985)^{300} + \binom{300}{1} (0.015)^1 (0.985)^{299} \\
 &= (0.985)^{300} + 300 (0.015) (0.985)^{299} \\
 &= 0.010737 + 0.049051 \\
 &\approx 0.059788
 \end{aligned}$$

- (ii) Using the Poisson Approximation, $\mu = np = 300(0.015) = 4.5$

$$\begin{aligned}
 P(x \leq 1) &= P(x = 0) + P(x = 1) \\
 &= \frac{4.5^0 e^{-4.5}}{0!} + \frac{4.5 e^{-4.5}}{1!} \\
 &= e^{-4.5} + 4.5 e^{-4.5} \\
 &= 5.5 e^{-4.5} \approx 0.061099
 \end{aligned}$$

- 5.3(e)** A missile protection system consists of 10 radar units operating independently. Suppose each has a probability of 0.70 of detecting a missile entering a zone that is covered by all units.

- (i) Describe how the operation of the systems fits into the binomial process.
(ii) How many radar units would be required if the probability of detecting a missile by at least one unit is 0.998?

Solution

- (i) To decide whether operation of the missile protection system meets the binomial experiment, we must have the following requirements:
- The experiment consists of ten trials where each trial determines whether a particular radar unit detects the aircraft.
 - Each trial in one of two outcomes, the success being “detecting missile” and the failure, “not detecting missile”
 - The probability of success (a radar unit detecting missile) is 0.70.
 - The trials are independent because the radar units operate independently.
 - The random variable of interest, y is the number of successes in the operation of ten radar units. Hence $x \square B(n = 10, p = 0.70)$.

(ii) $P(x \geq 1) = 1 - P(x = 0) = 0.998$

$$1 - P(x = 0) = 1 - \binom{n}{0} (0.70)^0 (0.30)^n = 0.0998$$

$$(0.30)^n = 0.002$$

$$n \log_e (0.30) = \log_e (0.002)$$

$$n = \frac{\log_e (0.002)}{\log_e (0.30)} = 5.162 \approx 5$$

1-5.4 Trial Questions 1-5:

- 1.(a) A sales person has found that the probability of a sale on a single contact is approximately 0.03. If the salesperson contacts 100 prospects, what is the approximate probability of making at least one sale?
- (b) Many utility companies have begun to promote energy conservation by offering discount rates to customers who keep their energy usage below certain established subsidy standards. A recent EPA report notes that 70% of residents in a town have reduced their electricity consumption sufficiently to qualify for discount rates. Suppose 20 residential subscribers are randomly selected from the town.
 - (i) Explain how this situation can be modelled by the binomial distribution.
 - (ii) Find the expected number of residents who qualify for the subsidy.
 - (iii) What is the probability that at least four qualify for the favourable rates?
- 2.(a) From a large lot of new tyres n are randomly sampled by a potential customer and the number of defectives x is observed. If at least one defective tyre is observed in the sample of n , the entire lot is rejected by the customer.
 - (i) If 10% of the tyres are defective, find n such that the probability of rejecting the lot is approximately 0.9.
 - (ii) Use your result in (i) to find the probability that at least three defective tyres would be observed
- (b) It is conjectured that an impurity exists in 30% of all drinking wells in a certain rural community. It is too expensive to test all the many wells in the area and so a random sample of 10 wells was selected and subjected to testing. Is it likely to find 7 or more impure wells for drinking?

- 3.(a) A random variable y is described by the Geometric distribution. State the characteristics of y and derive its probability distribution. Also show that the distribution of y is probability mass function. Hence derive the mean and variance of the distribution of y .
- (b) The employees of a firm that manufactures insulation are being tested for indications of asbestos in their lungs. The firm is requested to send three employees who have positive indication of asbestos to a medical centre for further testing. If 40% of the employees have positive indication of asbestos in their lungs
- (i) Find the probability that ten employees must be tested in order to find three positives.
 - (ii) Find the expected value and variance of the total cost of conducting the tests necessary to locate the three positives if each test cost \$20.
- 4.(a) The customers arriving at a check-out counter in a departmental store follows the Poisson distribution at an average of five customers per hour.
- (i) During a particular one-hour period, determine the probability that at least three customers will arrive.
 - (ii) What is the probability that in a given two-hour period, exactly two customers will arrive?
 - (iii) In a thirty-minute period, determine the probability that at most two customers will arrive.
- (b) In a certain communication system, there is an average of 1 transmission error per 10 seconds. What is the probability of observing at least 2 errors within a duration of one-half minute?
- (c) The production of fuses are packaged in boxes after manufacturing. The probability of a defective fuse in a box of 200 fuses from the manufacturing process is 2 percent which is independent of the other fuses.
- (i) Estimate the expected number of fuses and its standard deviation?
 - (ii) What is the approximate probability that there will be fewer than two defective fuses in the box?

SESSION 2-5: THE NORMAL DISTRIBUTION

2-5.1 Introduction

The *Normal distribution* is one of the most widely used probability distribution for modelling random experiments. It provides a good model for continuous random variables involving measurements such as time, heights/weights of persons, marks scored in an examination, amount of rainfall, growth rate and many other scientific measurements.

- *Definition 5.9:*

The probability density function for the normal random variable, x which is simply called normal distribution is defined by

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, & -\infty < x < \infty \\ 0, & \text{elsewhere} \end{cases}$$

where $\sigma > 0$, $-\infty < \mu < \infty$ and the mean and variance of the measurements, x are

$$E(x) = \mu \text{ and } \text{Var}(x) = \sigma^2$$

If a random variable is modelled by the Normal distribution with mean, μ and variance, σ^2 , then it is simply denoted as $x \sim N(\mu, \sigma^2)$. The graph of the Normal probability distribution is *bell-shaped smooth curve*, as illustrated in the diagram below.

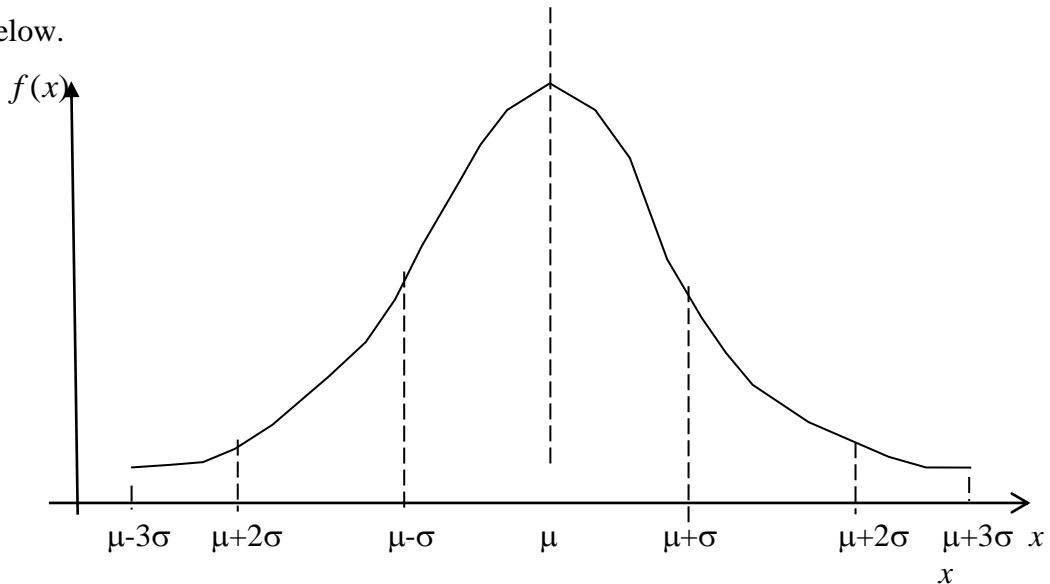


Figure 5.4: The Normal Curve

2-5.2 Properties of Normal Curve:

The normal curve has the following desirable properties, accounting for the widespread applications of the Normal distribution.

- The normal curve is symmetrical about its mean, μ .
- The mean, median and mode of x are the same.
- The total area under the normal curve is equal to 1.
- The probability distribution of the normal random variable, x is completely determined by its two parameters μ and σ .
- The curve is asymptotic to its horizontal axis.
- The probabilities of the normal random variable, x are given by the areas under the curve. As an example, the probabilities that x lies within 1, 2 and 3 standard deviation(s) about the mean are approximately given as follows:

$$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.6826, P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9544, \text{ and}$$

$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.9980$$

2-5.3 Computation of Probability of Normal Random Variable:

To compute the probability that x lies within the interval $[a, b]$, $P(a \leq x \leq b)$ the normal random variable, x is standardized using the transformation, $Z = \frac{x - \mu}{\sigma}$, called the *Z-score*. The probability distribution function of the standardized random variable, $f(z)$ which has the same shape as $f(x)$ is called the *Standard (or Unit) Normal distribution*. It is defined as

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, -\infty < z < \infty, \text{ or } Z \sim N(0, 1),$$

where the mean and variance are 0 and 1 respectively. Thus,

$$\begin{aligned} P(a \leq x \leq b) &= \int_a^b f(x) dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \int_{a/\mu}^{b/\mu} f(z) dz = \int_{z_a}^{z_b} \frac{1}{\sqrt{2\mu}} e^{-\frac{1}{2}Z^2} dz \\ &= P(Z_a \leq Z \leq Z_b) = P(Z \leq Z_b) - P(Z \leq Z_a) \\ &= \int_{-\infty}^{z_b} f(z) dz - \int_{-\infty}^{z_a} f(z) dz = \Phi(z_b) - \Phi(z_a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

The evaluation of the integral of $f(z)$, $\phi(z)$ for various values of Z have been tabulated in a table called the *Normal table* (See Appendix). In evaluating $\phi(z)$ the following are noted.

- (i) $\Phi(-k) = 1 - \Phi(k) = 1 - P(z \leq k)$
- (ii) $P(z > k) = 1 - P(z \leq k) = 1 - \Phi(k)$
- (iii) $P(-k \leq z \leq k) = \Phi(k) - \Phi(-k) = \Phi(k) - [1 - \Phi(k)] = 2\Phi(k) - 1$

Example 5.4:

5.4(a) Find the following probabilities using the Normal table.

- (i) $P(z \leq -1.95)$
- (ii) $P(-1.18 \leq z \leq 0.48)$
- (iii) $P(0 \leq z \leq 2.58)$
- (iv) $P(z > 2.63)$
- (v) $P(-2.35 \leq z \leq 2.35)$

5.4(b) Suppose that $y \sim N(6, 4)$. what percentage will y fall between 5 and 10?

Solution:

(a) Using the Normal table given in Appendix, the required probabilities are as follows:

- (i) $P(z \leq -1.95) = \Phi(-1.95) = 0.0256$
- (ii) $P(-1.18 \leq z \leq 0.48) = P(z \leq 0.48) - P(z \leq -1.18)$
 $= \Phi(0.48) - \Phi(-1.18) = 0.6844 - 0.1190 = 0.4654$
- (iii) $P(0 \leq z \leq 2.58)$
 $= P(z \leq 2.58) - P(z \leq 0.00)$
 $= \Phi(2.58) - \Phi(0.00) = 0.9951 - 0.5000 = 0.4951$
- (iv) $P(z > 2.63) = 1 - P(z \leq 2.63)$
 $= 1 - \Phi(2.63) = 1 - 0.9957 = 0.0043$
- (v) $P(-2.35 \leq z \leq 2.35)$
 $= P(z \leq 2.35) - P(z \leq -2.35)$
 $= \Phi(2.35) - \Phi(-2.35) \text{ or } 2\Phi(2.35) - 1$
 $= 0.9906 - 0.0094 = 2(0.9906) - 1 = 0.9812$

- (b) Given that $y \sim N(6, 4)$, where $\mu = 6$ and $\sigma = 2$

$$\begin{aligned}
 & P(5 < y < 10) \\
 &= P(y < 10) - P(y < 5) \\
 &= \Phi\left(\frac{10-6}{2}\right) - \Phi\left(\frac{5-6}{2}\right) \\
 &= \Phi(2) - \Phi(-0.5) \\
 &= 0.9772 - 0.3085 = 0.6687 \text{ or } 66.87\%
 \end{aligned}$$

- 5.4(c)** The weekly amount spent for maintenance and repairs in a certain company was observed, over a long period of time, to be approximately normally distributed with a mean of \$400 and a standard deviation of \$20.

- (i) If \$450 is budgeted for the week, what is the probability that the actual costs will exceed the budgeted amount?
- (ii) How much should be budgeted for weekly repairs and maintenance in order for the budgeted amount is exceeded with a probability of 0.1?

- 5.4(d)** The nicotine content of a brand of cigarettes is normally distributed with a mean of $2.0mg$ and a standard deviation of $0.25mg$. What is the probability that a cigarette will have nicotine content?

- (i) Of $1.65mg$ or less
- (ii) Between 1.50 and $2.25mg$
- (iii) Of $2.18mg$ or more?

Solution:

- (c) The amount spent on maintenance and repairs, $x \sim N(400, 20^2)$, where $\mu = \$400$ and $\sigma = \$20$,

$$(i) P(x > 450) = 1 - P(x \leq 450)$$

$$\begin{aligned}
 &= 1 - \Phi\left(\frac{450-400}{20}\right) \\
 &= 1 - 0.9938 = 0.0062
 \end{aligned}$$

- (ii) Let the budgeted amount be k dollars.

$$P(x > k) = 0.1$$

$$= 1 - P(x \leq k) = 0.1$$

$$P(x \leq k) = 0.9$$

$$\Phi\left(\frac{k-400}{20}\right) = 0.9$$

$$\frac{k-400}{20} = \Phi^{-1}(0.9) = 1.28 = \phi^{-1}(0.9)$$

$$k = 400 + 20(1.28) = 425.6 \text{ (or } k = \$425.60\text{)}$$

(d) Given that the nicotine content of cigarettes, $y \sim N(2.0, 0.25^2)$,

$$(i) P(y < 1.65) = \Phi\left(\frac{1.65 - 2.0}{0.25}\right)$$

$$= \Phi(-1.40) = 0.0808$$

$$(ii) P(1.50 < y < 2.25)$$

$$= \Phi\left(\frac{2.25 - 2.0}{0.25}\right) - \Phi\left(\frac{1.50 - 2.0}{0.25}\right)$$

$$= \Phi(1.04) - \Phi(-1.92)$$

$$= 0.8508 - 0.0274 = 0.8234$$

$$(iii) P(y > 2.18) = 1 - P(y \leq 2.18)$$

$$= 1 - \Phi\left(\frac{2.18 - 2.0}{0.25}\right)$$

$$= 1 - \Phi(0.72)$$

$$= 1 - 0.7642 = 0.2358$$

2-5.4 The Normal Approximation to Binomial

The Normal distribution provides a good approximation to the binomial distribution when the number of trials, n is large, probability of a success in a trial, p not close to 0 or 1 and both np and $np(1-p)$ are greater than 5. Thus the binomial random variable, x becomes approximately normal random variable with mean, $\mu = np$ and variance, $\sigma^2 = np(1-p)$. The theoretical justification for this approximation is based on the *Central limit Theorem* which is widely applied in inferential analysis of data. The approximation is said to be adequate when the interval $\mu \pm 2\sigma$ falls between 0 and n and also said to be very good if the interval $\mu \pm 3\sigma$ falls between 0 and n .

To improve upon the approximation, a continuity correction may be utilized by adding or subtracting 0.5 to/from x to account for the fact that a discrete distribution is being approximate by a continuous distribution. In this case the standardized random variable

thus becomes $Z = \frac{x \pm 0.5 - np}{\sqrt{np(1-p)}}$

Example 5.5:

5.5(a) Suppose that x has a Binomial distribution with $n = 200$ and $p = 0.4$. Using the continuity correction use the Normal approximation to Binomial to find each of the following probabilities:

- | | |
|-----------------------|---------------------|
| (i) $P(x = 90)$ | (ii) $P(x \leq 95)$ |
| (iii) $P(x > 65)$ | (iv) $P(x < 60)$ |
| (v) $P(70 < x < 100)$ | |

5.5(b) A manufacturer of components for electric motors has found that about 10% of the production will not meet customer specifications. If 500 components are examined,

- (i) Find the expected number of components which did not meet customer specifications.
- (ii) Find the probability that exactly 52 components or more did not meet customer specifications.
- (iii) Find the probability that between 36 and 58 (inclusive) components did not meet customer specifications.

Solution:

- (a) Given that the binomial random variable, x is approximately normally distributed, $x \sim N(\mu, \sigma^2)$, where

$$\mu = 200(0.4) = 80 \text{ and } \sigma = \sqrt{200(0.4)(0.6)} = 6.9282$$

(i) $P(x = 90) = P(89.5 \leq x \leq 90.5)$

$$= \Phi\left(\frac{90.5 - 80}{6.9282}\right) - \Phi\left(\frac{89.5 - 80}{6.9282}\right)$$

$$= \Phi(2.81) - \Phi(-1.37) = 0.9357 - 0.9147 = 0.0210$$

$$(ii) \quad P(x \leq 95) = (x \leq 95.5)$$

$$= \Phi\left(\frac{95.5 - 80}{6.9282}\right)$$

$$= \Phi(2.24) = 0.9875$$

$$(iii) \quad P(x > 65) = 1 - P(x \leq 65.5)$$

$$= 1 - \Phi\left(\frac{65.5 - 80}{6.9282}\right)$$

$$= 1 - \Phi(-2.09) = 1 - 0.0183 = 0.9817$$

$$(iv) \quad P(x < 60) = P(x \leq 59.5)$$

$$= \Phi\left(\frac{59.5 - 80}{6.9282}\right) = \Phi(-2.96) = 0.0015$$

$$(v) \quad P(70 < x < 100) = (70.5 \leq x \leq 99.5)$$

$$= \Phi\left(\frac{99.5 - 80}{6.9282}\right) - \Phi\left(\frac{70.5 - 80}{6.9282}\right)$$

$$= \Phi(2.81) - \Phi(-1.37)$$

$$= 0.9975 - 0.0853 = 0.9122$$

- (b) The number of components that did not meet customer specifications, x is approximately normally distributed where $\mu = np = 500(0.1) = 50$ and $\sigma^2 = np(1-p) = 500(0.1)(0.9) = (6.71)^2$

$$(i) \quad \text{The expected number of components that did not meet customer specifications, } E(x) = \mu = np = 50$$

$$(ii) \quad P(x \geq 52) = 1 - P(x < 51.5)$$

$$= 1 - \Phi\left(\frac{51.5 - 50}{6.71}\right)$$

$$= 1 - \Phi(0.22)$$

$$= 1 - 0.5871 = 0.4129$$

$$(iii) \quad P(36 \leq x \leq 58) = P(35.5 \leq x \leq 58.5)$$

$$= \Phi\left(\frac{58.5 - 50}{6.71}\right) - \Phi\left(\frac{35.5 - 50}{6.71}\right)$$

$$= \Phi(1.27) - \Phi(-2.16)$$

$$= 0.8980 - 0.0197 = 0.8783$$

2-5.5 Linear Combinations of Normal Random Variables:

- Theorem 5.1:

Let $x_1, x_2, x_3, \dots, x_n$ be independent normal random variables and $a_1, a_2, a_3, \dots, a_n$ be constants. Then the linear combination,

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n = \sum_{i=1}^n a_i x_i$$

is also a normal random variable with mean and variance given as follows.

$$E(y) = E\left(\sum_{i=1}^n a_i x_i\right) = \sum_{i=1}^n a_i E(x_i), \text{ and } \text{Var}(y) = \text{Var}\left(\sum_{i=1}^n a_i x_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(x_i)$$

However, if the random variables, $x_1, x_2, x_3, \dots, x_n$ are not independent then

$$\text{Var}(y) = \sum_{i=1}^n a_i^2 \text{Var}(x_i) + 2a_i a_j \sum_i \sum_j \text{Cov}(x_i, x_j)$$

where the double sum is over all pairs (i, j) with $i < j$

- The Sampling Distribution of the Sample Mean, (\bar{x}) :

Let $x_1, x_2, x_3, \dots, x_n$ be random observations drawn from a normally distributed population with a mean of μ and a variance of σ^2 . Then the sample mean

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is normally distribution with mean and variance given as follows:

$$E(\bar{x}) = \frac{1}{n} \sum_{i=1}^n E(x_i) = \frac{1}{n} (n\mu) = \mu$$

$$\text{Var}(\bar{x}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(x_i) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}$$

The sampling distribution of the sample mean is approximately normally distributed, by the Central Limit Theorem.

- *Central Limit Theorem:*

Let a random sample of size n observations be selected from a population with mean μ and variance, σ^2 . The sampling distribution of the sample mean (\bar{x}) will be approximately normally distributed with mean, $\mu_{\bar{x}} = \mu$ and variance, $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ provided n is sufficiently large.

Example 5.6:

5.6(a) If x and y are independent normal random variables with $E(x) = 1$, $Var(x) = 4$, $E(y) = 10$ and $Var(y) = 9$, determine the following:

- $E(2x+3y)$ and $Var(2x+3y)$
- $P(2x+3y < 40)$

5.6(b) The mass of a biscuit is a normal random variable (x) with mean 50 *grams* and a standard deviation of 4 *grams*. If a packet contains 20 biscuits and the mass of the packaging material is also normal random variable with mean 100 *grams* and standard deviation, 3 *grams*, find the probability that the mass of the total packet

- Will exceed 1,047 *grams*
- Lies between 1,050 and 1,200 *grams*

Solution:

(a) Given that $x \sim N(1, 4)$ and $y \sim N(10, 9)$, we let $T = 2x + 3y$, then

- $$\begin{aligned}\mu_T &= E(T) = E(2x + 3y) \\ &= 2E(x) + 3E(y) \\ &= 2(1) + 3(10) = 32\end{aligned}$$

$$\begin{aligned}\sigma_T^2 &= Var(T) = Var(2x + 3y) \\ &= 4Var(x) + 9Var(y) \\ &= 4(4) + 9(9) = 97 = (9.85)^2\end{aligned}$$

(ii) From (i) $T \sim N(32, 97)$

$$\begin{aligned}P(T < 40) &= \Phi\left(\frac{40 - 32}{9.85}\right) \\ &= \Phi(0.81) = 0.7910\end{aligned}$$

- (b) Let the mass of contents of the packet be T_1 and the packaging material T_2 . Then the total mass $T = T_1 + T_2$ is normally distributed, where the mean and variance are given as:

$$\begin{aligned}\mu_T &= E(T) = E(T_1 + T_2) \\ &= E(T_1) + E(T_2) \\ &= E(20x) + E(100) \\ &= 20E(x) + 100 \\ &= 20(50) + 100 = 1,100 \text{ g}\end{aligned}$$

$$\begin{aligned}\sigma_T^2 &= \text{Var}(T) = \text{Var}(T_1 + T_2) \\ &= \text{Var}(20T_1) + \text{Var}(T_2) \\ &= 20^2 \text{Var}(T_1) + \text{Var}(T_2) \\ &= 20^2(16) + 9 = 6,409 = (80.06)^2\end{aligned}$$

Hence, $T \sim N(1,100 + 6,409)$

$$(i) \quad P(T > 1,074) = 1 - P(T \leq 1,074)$$

$$\begin{aligned}&= 1 - \Phi\left(\frac{1,074 - 1,100}{80.06}\right) \\ &= 1 - \Phi(-0.32) \\ &= 1 - 0.3745 \\ &= 0.6255\end{aligned}$$

$$(ii) \quad P(1,050 < T < 1,200)$$

$$\begin{aligned}&= \Phi\left(\frac{1,200 - 1,100}{80.06}\right) - \Phi\left(\frac{1,050 - 1,100}{80.06}\right) \\ &= \Phi(1.25) - \Phi(-0.62) \\ &= 0.8944 - 0.2676 \\ &= 0.6268\end{aligned}$$

- 5.6(c)** A cigarette manufacturer claims that the mean nicotine content in cigarettes is 2 mg with a standard deviation of 0.3 mg.

- (i) If this claim is valid what is the probability that a sample of size 900 cigarettes will yield a mean nicotine content exceeding 2.02 mg?
- (ii) What role does the Central Limit Theorem play in identifying the distribution used in (i)?

2-5.6 Trial Questions 2-5:

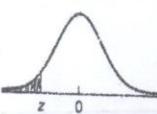
- 1.(a) It is known that the weights of certain group of individuals are approximately normally distributed with a mean of 140 pounds and a standard deviation of 25 pounds.
 - (i) What is the probability that a person picked at random from this group will weigh between 100 and 170 pounds?
 - (ii) If the group is made up of 9,000 people, how many of them would you expect to weigh more than 200 pounds?
 - (b) The prices of estate houses are assumed to be normally distributed with a mean of ₦18 million. It is known that 90 percent of the houses are priced below ₦30 million.
 - (i) Find the standard deviation of the prices of the houses.
 - (ii) What percentage of houses will cost more than ₦20 million?
 - (iii) If the estate is made up of 2,500 houses how many would you expect to be priced less than ₦10 million?
-
- 2.(a) The reaction time of a driver to visual stimulus is normally distributed with a mean of 0.35 second and a standard deviation of 0.15.
 - (i) What is the probability that a reaction time requires more than 0.5 second?
 - (ii) What is the probability that a reaction time requires between 0.28 and 62 seconds?
 - (iii) What reaction time is required to exceed 90 percent of the time?
 - (b) A machine fills millet flour into 500-gram bags. The actual weights of the filled bags vary being approximately normally distributed with a variance of 100 grams.
 - (i) Find the mean weight of the filled bags, if 15% of the filled bags are underweight.
 - (ii) Calculate the proportion of the bags whose weight is between 495 and 530 grams.
 - (iii) If the mean weight is adjusted to 518.8 grams and the standard deviation remains unchanged, what percentage of bags would be sold underweight?

3. (a) (i) State the normal distribution and its properties
(ii) Under what conditions and how would you use the Normal Approximation to the Binomial?
- (b) It is believed that 45% of a large population of registered voters favour a particular candidate for the constituency. A public opinion poll used a random selected sample of voters and asked each person polled to indicate his/her preference for the candidate. What is the probability that a weekly poll based on 150 responses of registered voters will show at least 65% of the voters favouring the candidate?
- 4.(a) The blood pressure of a group of female industrial workers are normally distributed with a mean of 130 mmHg and a standard deviation of 15 mmHg .
(i) If 500 female workers are randomly selected, how many would you expect to have blood pressure (α) of at least 100 mmHg and (β) between 110 and 150 mmHg ?
(ii) If 68% of the workers have a systolic blood pressure of at most $k \text{ mmHg}$, find the value of k .
- (b) In a collection of plants, it is found that 20% have heights greater than 36.3 cm and 67% have heights greater than 29.9cm. Suppose the heights are normally distributed in this collection.
(i) Find the mean and the standard deviation of the heights of the plants.
(ii) If the collection is made up 500 plants, how many of them would you expect to have heights exceeding 25.8 cm but less than 37.5 cm?
- 5.(a) An experiment was conducted to test for the presence or absence of fungus on tobacco plants. 400 plants were observed to have been inflected by fungus.
(i) Does this appear to meet the requirements of a binomial experiment?
(ii) Previous experience suggests that the fungus affect 60% of the planting of tobacco seedlings. Is it probable that the observed number of infected plants could be larger than 245? Explain.
(iii) Suppose the characteristics of a binomial experiment are satisfied, what interpretation can you give to $p = 0.60$?

- (b) The manufacturing of semi-conductor chips produces 2% defective chips. Assume that the chips are independent and that a lot contains 1000 chips. Use the continuity correction to approximate the probability that
- (i) Exactly 20 chips are defective, and
 - (ii) Between 20 and 30 chips in the lot are defective.
 - (iii) Determine the number of defective chips, x such that the probability of obtaining x defective chips is greatest.
- 6.(a) Let the random variable, x be normally distributed with mean 30 and standard deviation 4. If $D = 90 - 2x$:
- (i) find the mean and variance of the distribution of D
 - (ii) Compute the probability, $P(30 \leq D \leq 36)$.
- (b) Let x and y be independent random variables where $x \sim N(2, 9)$ and $y \sim N(3, 16)$. The expectation and the standard deviation of the linear combination, $x + y$ are
- (c) A chartered airliner agency is asked to carry regular loads of 100 cartons of an item. The plane available for this work has a carrying capacity of 5,000 kg . If it is known that the weight of a carton of the item is normally distributed with a mean of 40 kg and standard deviation 9 kg . Can the agency take the order?
- 7.(a) The speeds of cars on a two-lane highway were found to be normally distributed with a mean of 58.6 mph and a standard deviation of 9.0 mph². What is the percentage of cars are observing the 55 mph speed limit on this highway?
- (c) A company pays its employees an average wage of GH¢15.90 an hour with a standard deviation of GH¢1.50. If the wages are approximately normally distributed and paid to the nearest pesewas,
- (i) what percentage of the workers receive wages between GH¢13.75 and GH¢16.22 an hour inclusive?
 - (ii) what amount needs to be exceeded for the highest 5 percent of the employee hourly wages?
 - (ii) If the standard deviation is adjusted to GH¢2.50, find the proportion of the employees whose hourly wages will not be less than GH¢16.22.

APPENDIX A STATISTICAL TABLES

TABLE II
Areas under the
standard normal curve

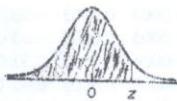


z	Second decimal place in z									
	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
-3.9										0.0000†
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005
-3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	0.0007
-3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	0.0010
-3.0	0.0010	0.0010	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	0.0013
-2.9	0.0014	0.0014	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
-2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
-2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
-2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
-2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
-2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0445
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
-1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
-0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
-0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
-0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
-0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
-0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
-0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
-0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
-0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
-0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
-0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000†

† For $z \leq -3.90$, the areas are 0.0000 to four decimal places.

APPENDIX A STATISTICAL TABLES

TABLE I (cont.)
Areas under the standard normal curve



z	Second decimal place in z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000 [†]									

[†] For $z \geq 3.90$, the areas are 1.0000 to four decimal places.