Chapter 1

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 realize 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.* Primacy 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 complied 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 lunched 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 enhance that no essential data are omitted.

3-1.6 List of Available Resources

A wide variety of resources is likely to be required for he 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.

Chapter 2

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.

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,

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

  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.

<u>Example 2.1:</u>

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)	Re*lative 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
Total	-	$n = 50$	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
Total	-	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,

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

$$= 1 + 3.322\log_{10} 65 = 7.0225 \approx 7$$

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 = maximum – minimum observation (weight)

= 80 – 11 = 69

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

Rounded up to 10 to include all the observations.

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

$$LB_1 = (minimum\ observation\ or\ lesser) - \frac{1}{2}(smallest\ unit\ of\ the\ mearements)$$

$$= 11 - \frac{1}{2}(1) = 10.5 \quad \text{or} \quad \left\{10 - \frac{1}{2}(1) = 9.5\right\}$$

$$UB_1 = LB_1 + C$$

= 10.5 + 10 = 20.5 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.

Weight of Tumour	*Tally*	*Frequency*	*Cumulative Frequency*
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
Total	-	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.

<u>Example 2.4:</u>

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	9.0	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

<u>Solution:</u>

(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 (ib/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
Total	-	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:

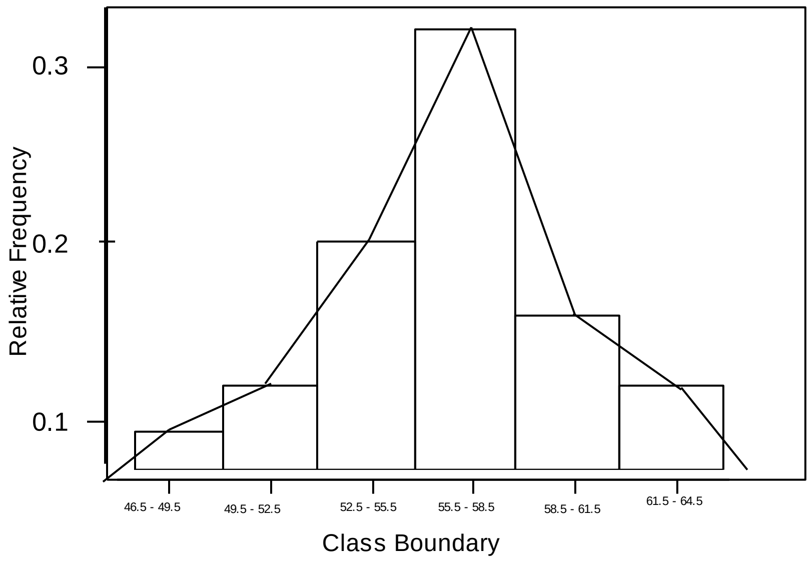
The histogram is the most widely used form of data presentation. It is a graph of frequency or relative frequency distribution where the (relative) frequencies are represented vertically by rectangular bars with no gaps in between them. The horizontal axis takes the observed data using the class boundaries 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):

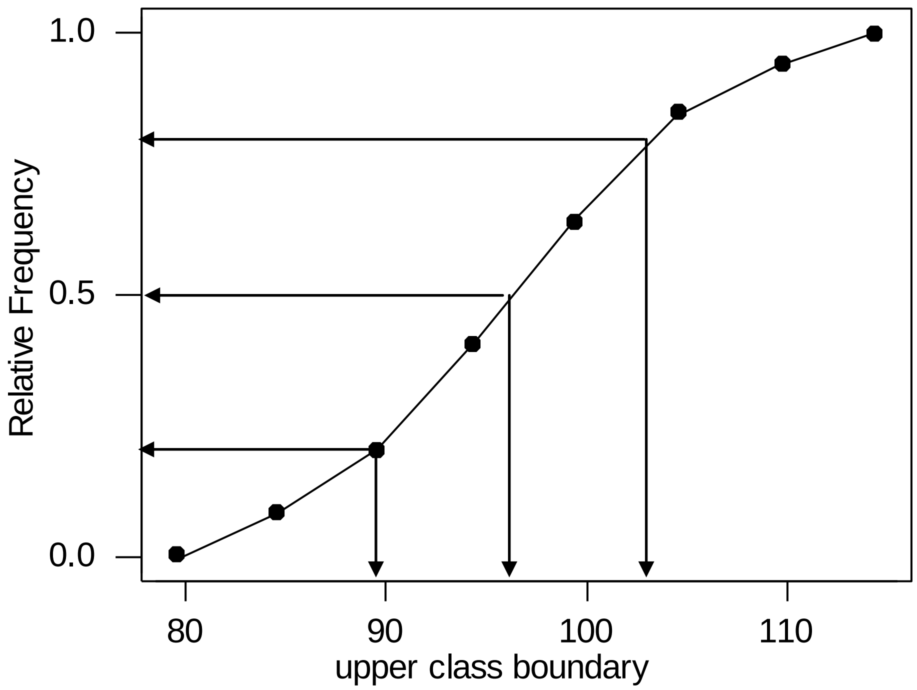
The cumulative frequency distribution shows the number of observations that fall above or below a specified value of observation. The cumulative frequency of a class is observed by cumulating (or summing) all frequencies up to the class. A graph obtained by plotting the cumulative points by smooth curve 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 – 109.5 109.5 – 114.5	7 4	69 + 7 = 76 76 + 4 = 80	0.9500 1.0000
Total	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.



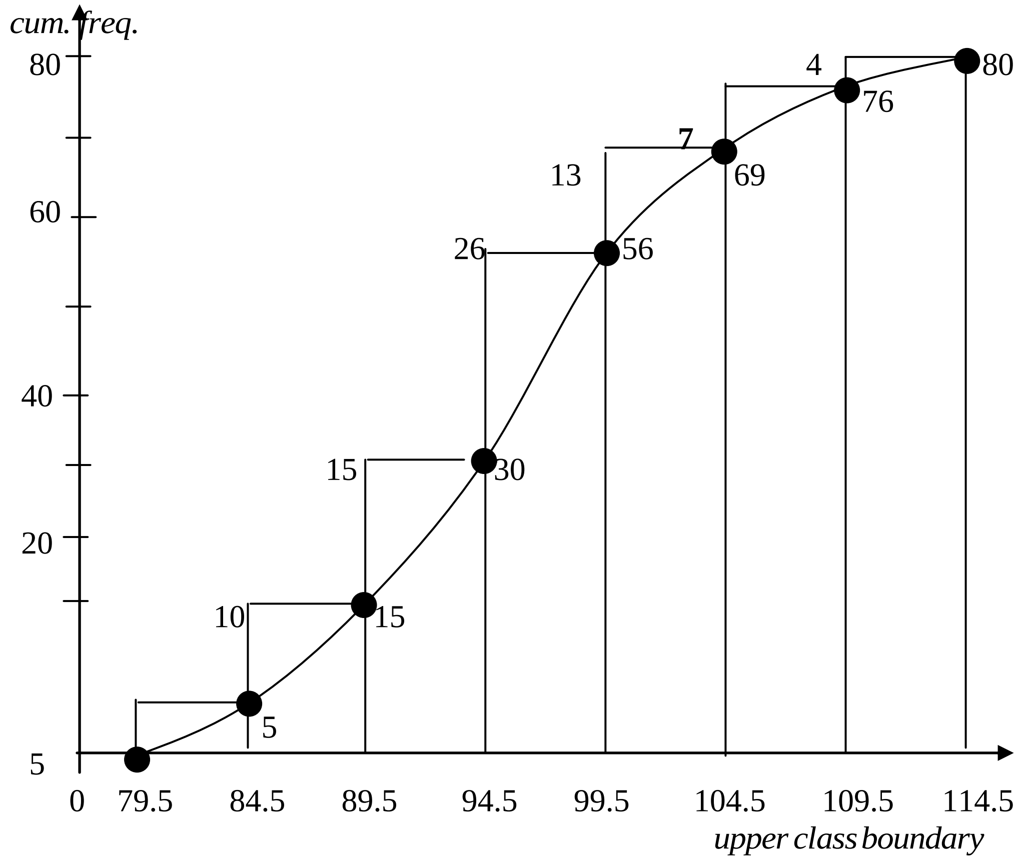
From the curve we have, approximately,

(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 beestings.

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 experience 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 = $\dfrac{16.5-3.8}{6} = 2.12 \approx 2.2$, and the first class boundary:

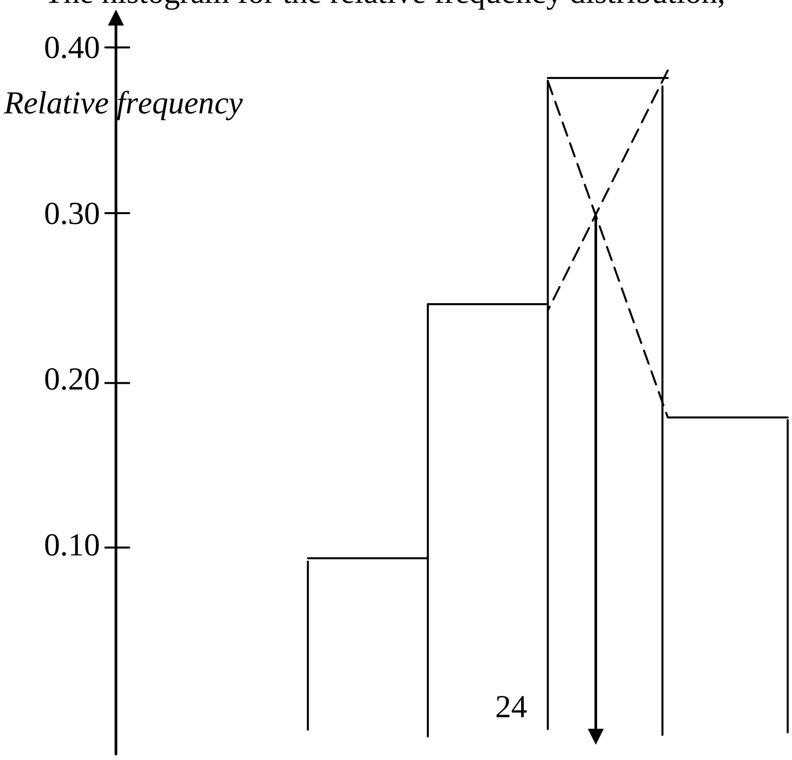
$LB_1 = 3.8 - \dfrac{1}{2}(0.1) = 3.75$

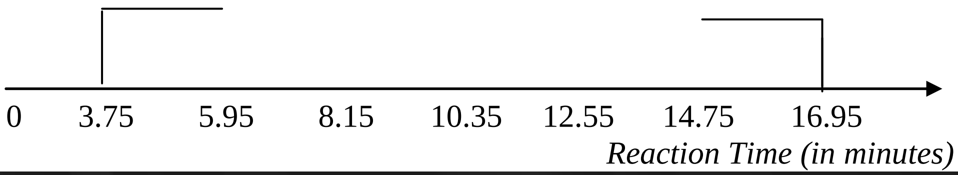
and $UB_1 = 3.75 + 2.2 = 5.95$

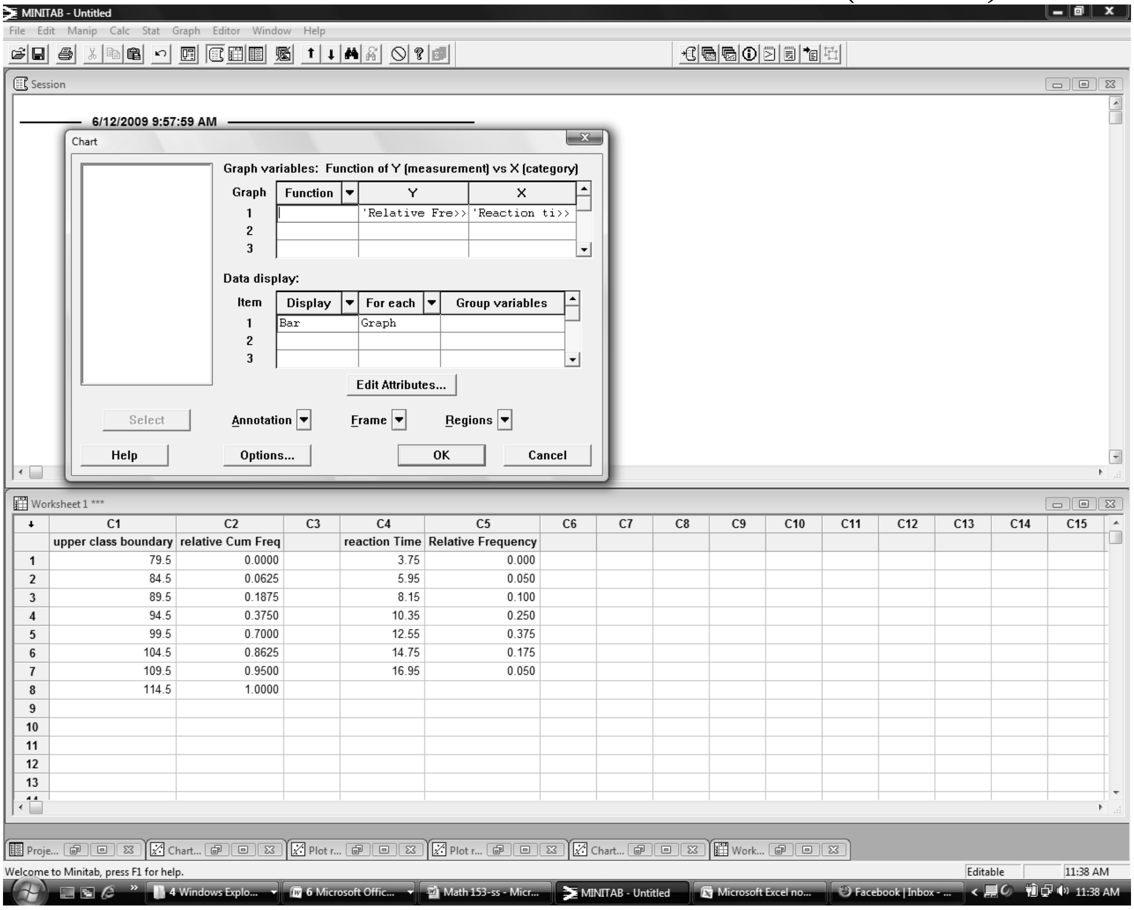
Hence the relative frequency distribution,

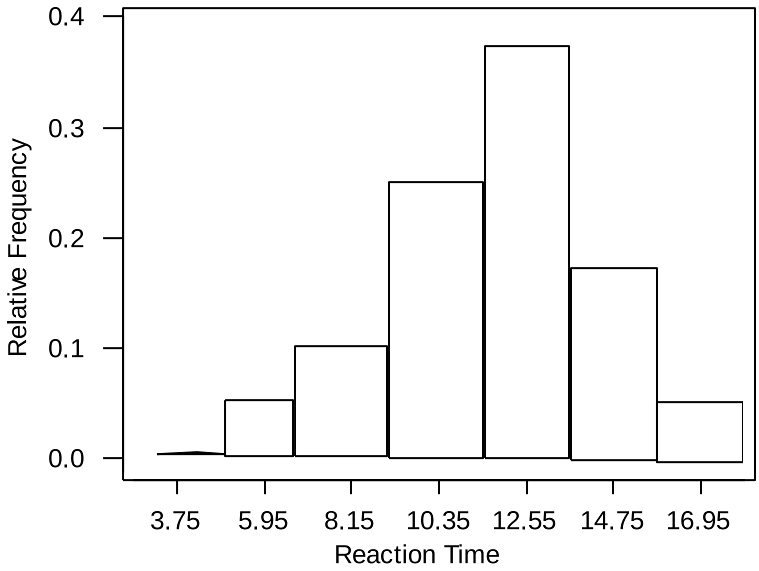
Time	*Tally*	*Frequency*	*Relative Frequency*
3.75 – 5.95	//	2	0.050
5.95 – 8.15	////	4	0.100
8.15 – 10.35	~~////~~ ~~////~~	10	0.250
10.35 – 12.55	~~////~~ ~~////~~ ~~////~~	15	0.375
12.55 – 14.75	~~////~~ //	7	0.175
14.75 – 16.95	//	2	0.050
Total	-	40	1.000

(b) The histogram for the relative frequency distribution,




0
3.75
5.95
8.15
10.35
12.55
14.75
16.95
Reaction Time (in minutes)


MINITAB - Untitled
File Edit Manip Calc Stat Graph Editor Window Help
Session
6/12/2009 9:57:59 AM
Chart
Graph variables: Function of Y (measurement) vs X (category)
Graph Function Y X
1 'Relative Fre>> 'Reaction ti>>
2
3
Data display:
Item Display For each Group variables
1 Bar Graph
2
3
Edit Attributes...
Select Annotation Frame Regions
Help Options... OK Cancel
Worksheet 1 ***
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14 C15
upper class boundary relative Cum Freq reaction Time Relative Frequency
1 79.5 0.0000 3.75 0.000
2 84.5 0.0625 5.95 0.050
3 89.5 0.1875 8.15 0.100
4 94.5 0.3750 10.35 0.250
5 99.5 0.7000 12.55 0.375
6 104.5 0.8625 14.75 0.175
7 109.5 0.9500 16.95 0.050
8 114.5 1.0000
9
10
11
12
13
Proje... Chart... Plot r... Plot r... Chart... Work...
Welcome to Minitab, press F1 for help.
Editable 11:38 AM
4 Windows Explo... 6 Microsoft Offic... Math 153-ss - Micr... MINITAB - Untitled Microsoft Excel no... Facebook | Inbox - ... 11:38 AM

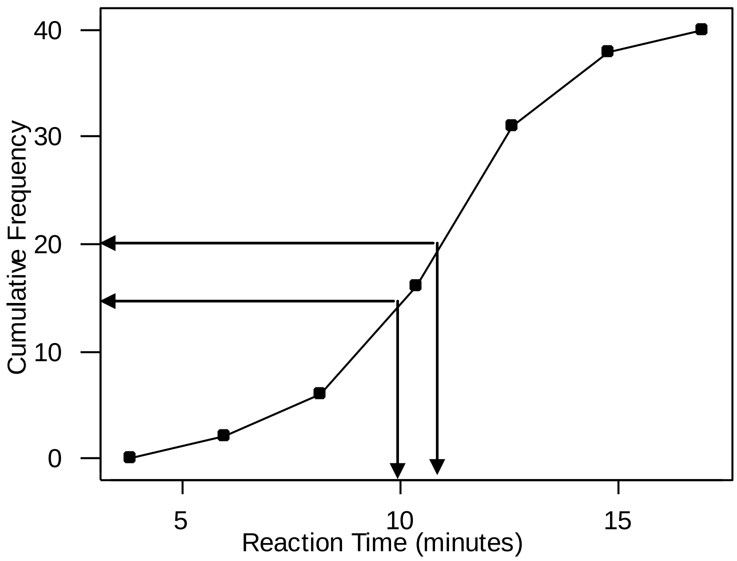


From the histogram, the *mode* = 10.35 + 1 = 11.35 minutes

(b) The relative/cumulative frequency distribution.

Reaction Time up to	Cumulative frequency	*Relative Cum. Frequency*
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
Total	-	-

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							

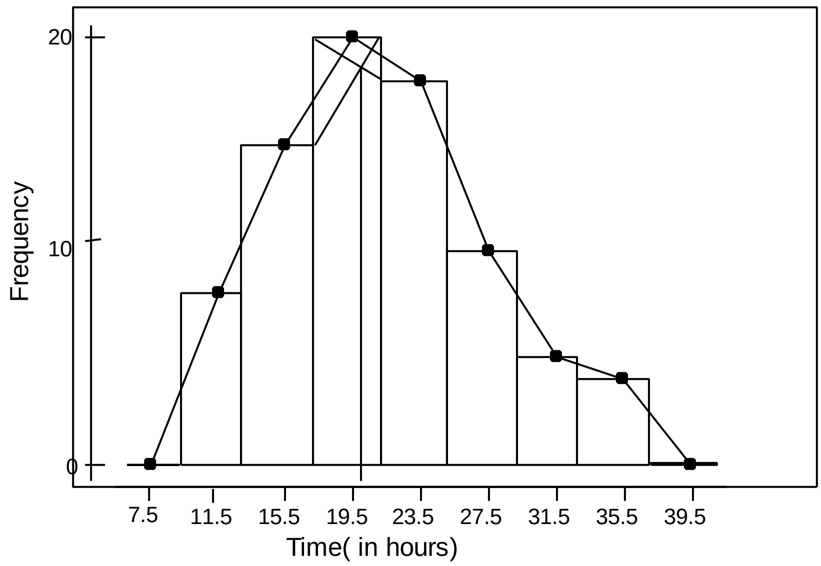
Construct a frequency distribution and histogram, given classes; 10 – 13, 14 – 17, 18 – 21, etc. and use it to construct a histogram.

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
Total	-	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 discusses 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

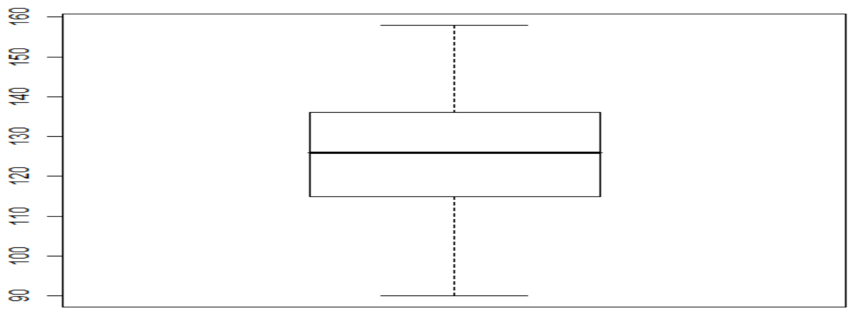
13 | 2 2 3 3 4 6 7 8 9

14 | 0 1 3

15 | 0 8

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

```
> boxplot (p)
```



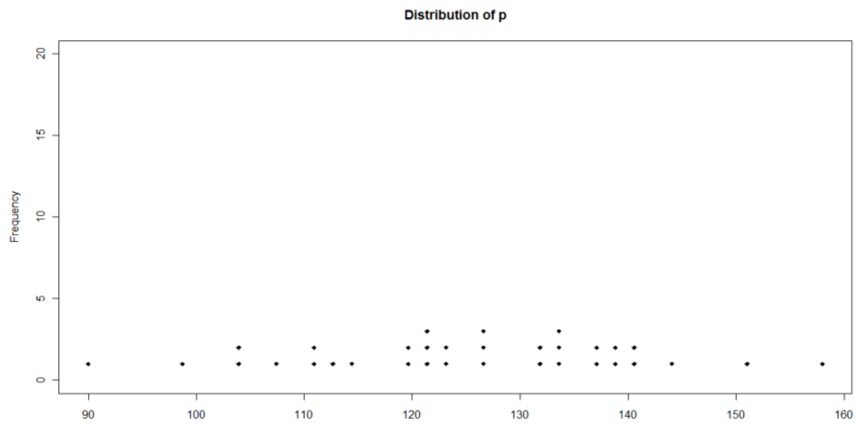
The five-number summary is given by

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 emphasis 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 chat displays the various categories of data as components of the whole set of data.

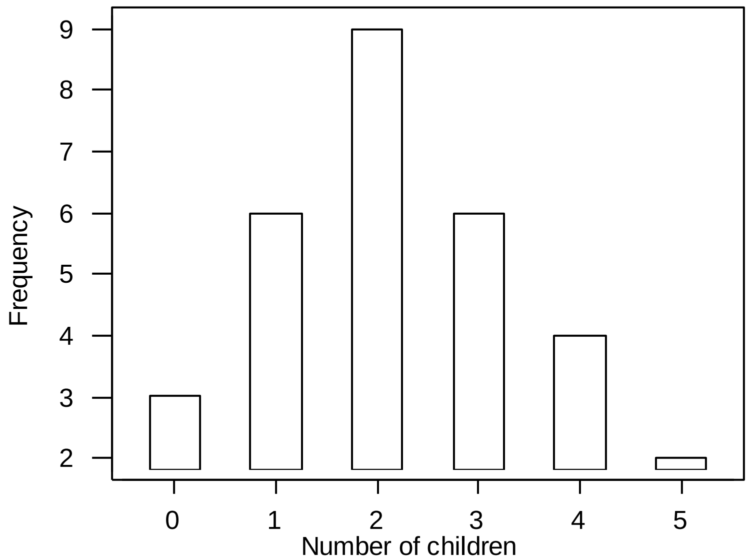
<u>Example 2.7</u> (Simple Bar Charts):

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
Total	30	1.00

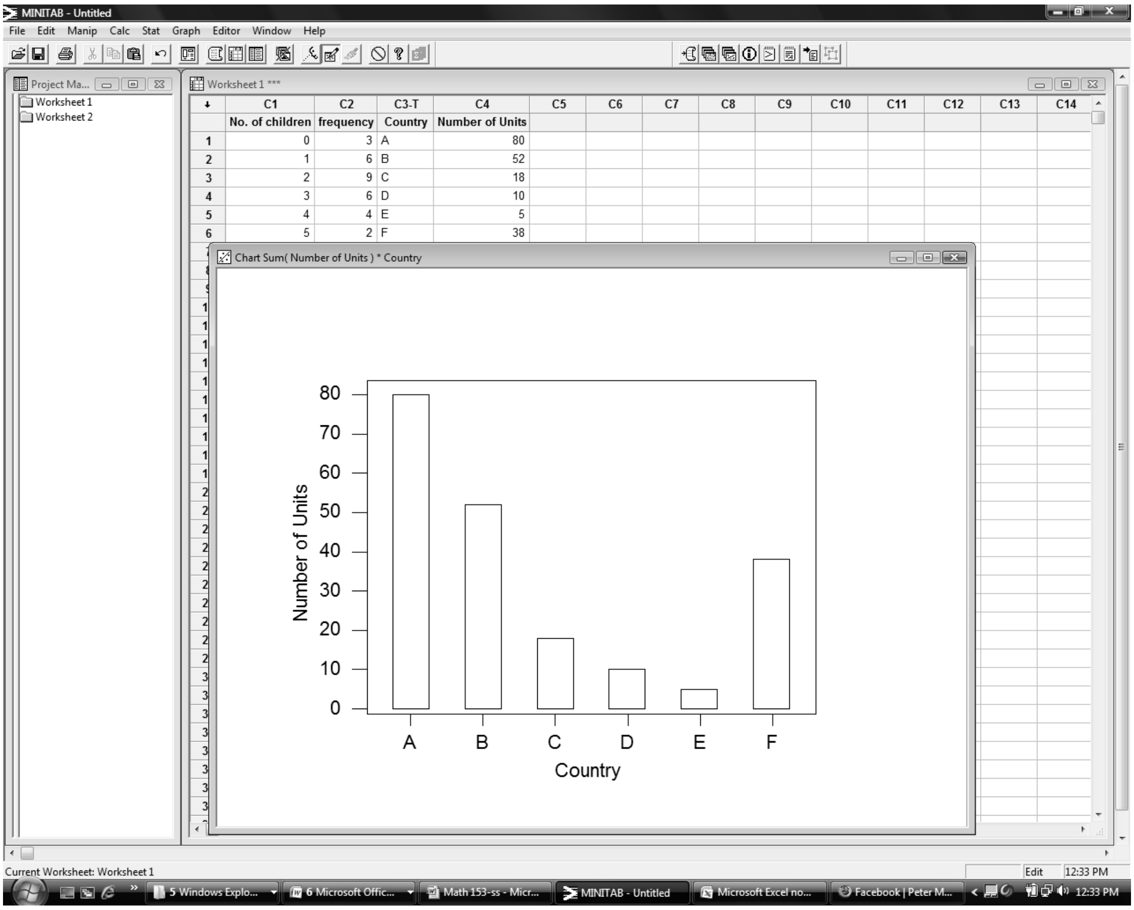
The bar chart for the above distribution show below

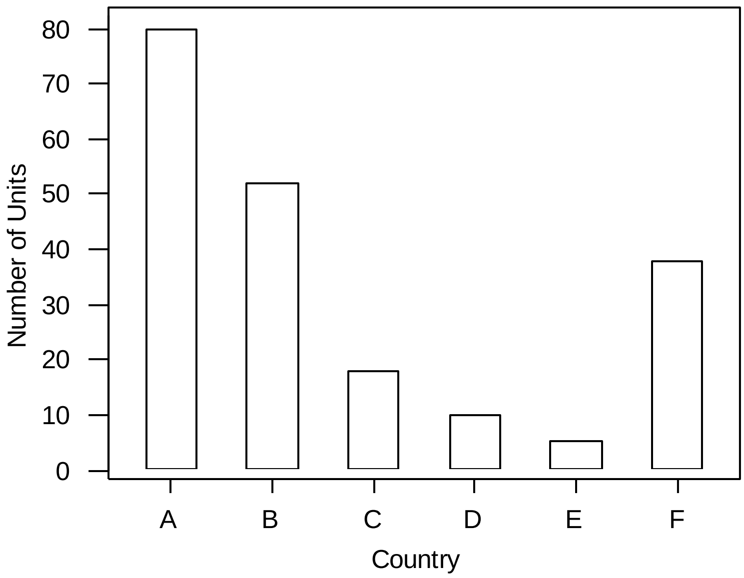


(b) The number of units of nuclear reactors in some countries in 1984 is given in the table below.

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:

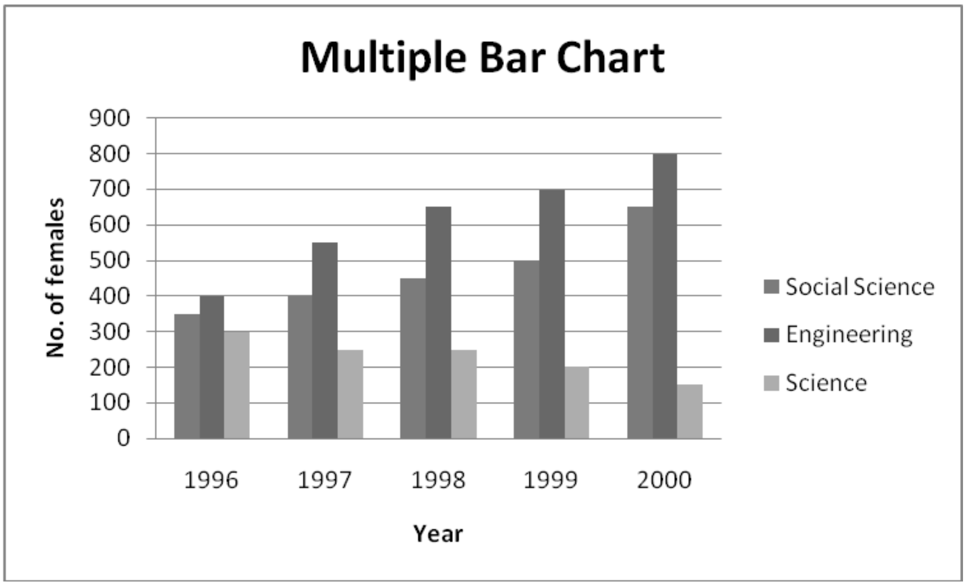




__Example 2.8 (Multiple Bar Charts)__:

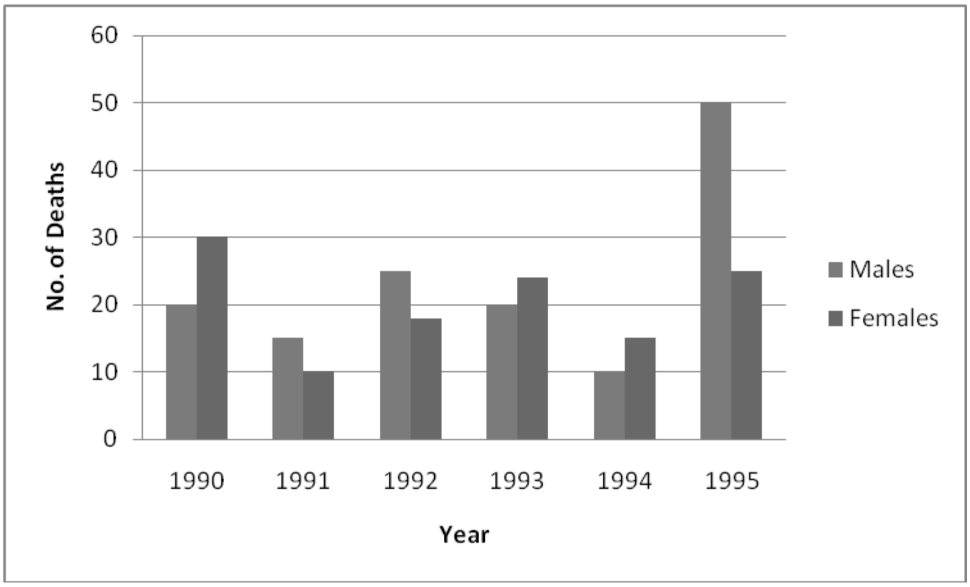
The number of females opted to offer programmes in Social Sciences, Engineering and Science for the period 1996–2000 in KNUST is as in the table below. The given data are displayed in the multiple bar charts, also shown below.

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



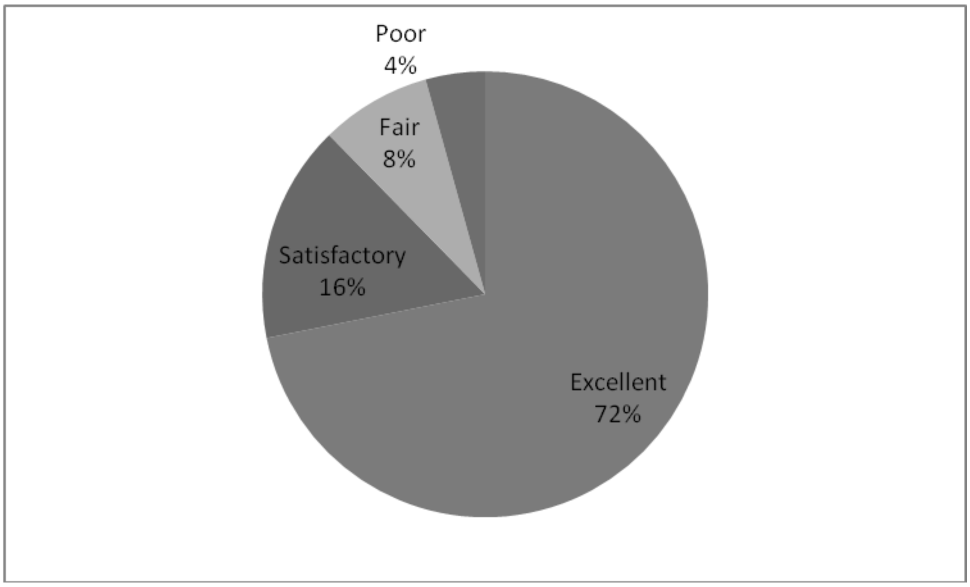
- *Pie Charts:* A pie chart is a circular diagram giving various fractions of section of a given data. The total number of observations of the data is represented by a *pie* which is denoted by a circle. The pie is then cut into slices (sectors) where each slice represents a category of the data. The 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: *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.

Category	*Percentage of income*
Medical Care	5
Clothing	15
Entertainment	5
Housing	40
Food	20
Transportation	15

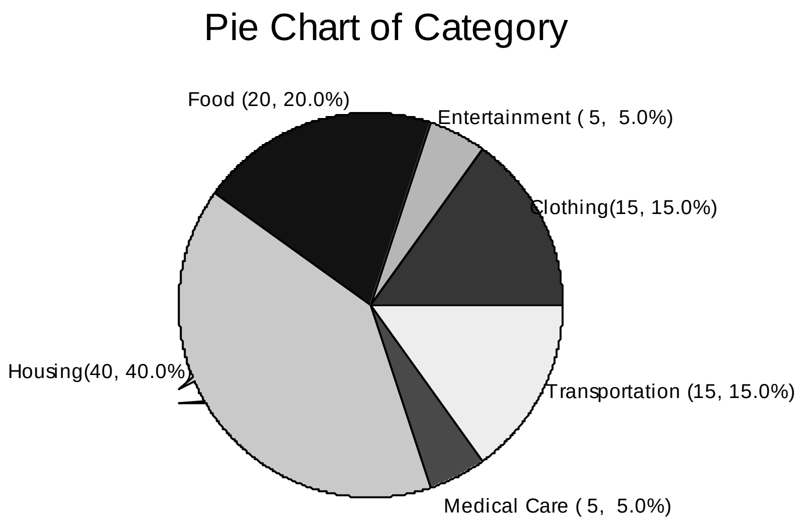
(i) Summarize the information in the form of pie chart.

(ii) What area represents the largest piece of the pie?

<u>*Solution:*</u>

Category	*Percentage of income*	*Angle of Sector*
Medical Care	5	$0.05 \times 360^0 = 18^0$
Clothing	15	$0.15 \times 360^0 = 54^0$
Entertainment	5	$0.05 \times 360^0 = 18^0$
Housing	40	$0.05 \times 360^0 = 144^0$
Food	20	$0.05 \times 360^0 = 72^0$
Transportation	15	$0.05 \times 360^0 = 54^0$
Total	100	360^0

(i) The required pie chart is as

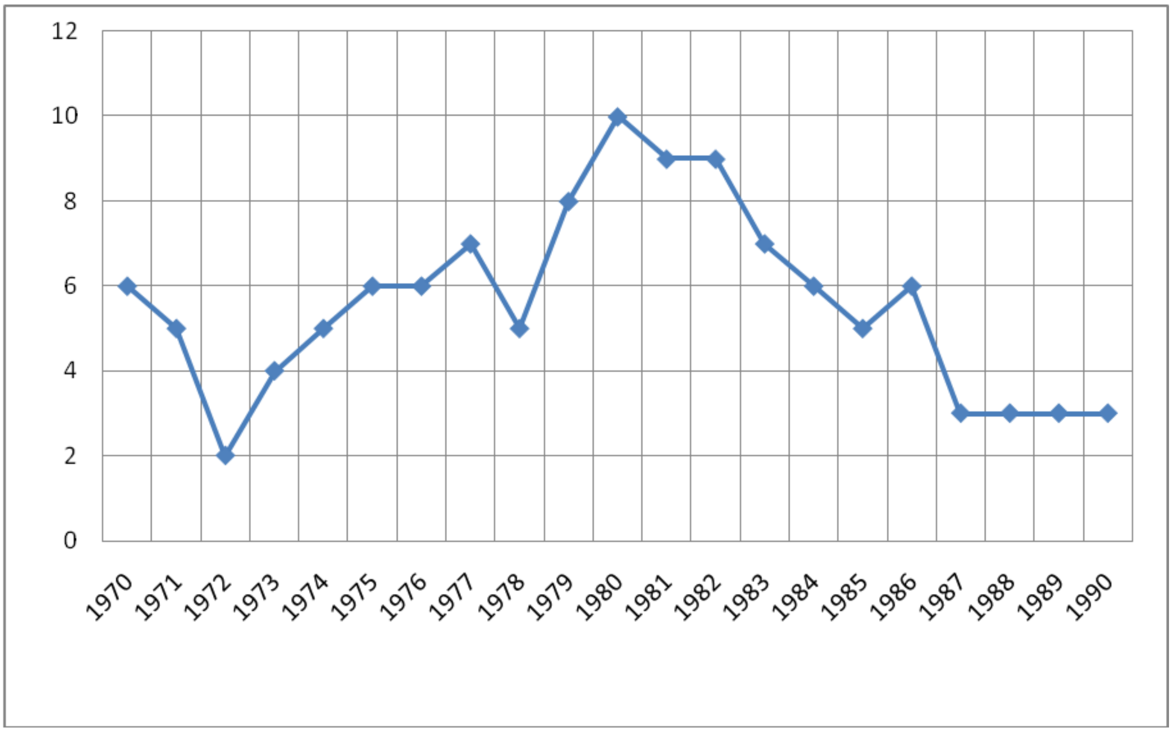


(ii) The largest piece of pie is "*housing*"

- *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 in to 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 a*rithmetic 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 $x_1, x_2, x_3, \ldots, x_n$ 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, $x_1, x_2, x_3, \ldots, x_k$ occur in a frequency distribution with corresponding frequencies, $f_1, f_2, f_3, \ldots, f_k$, then the mean is given by

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n}(x_1 + x_2 + \ldots + x_n), \text{ or}$$
$$= \frac{1}{n}\left(\sum_{i=1}^{k} f_i x_i\right) = \frac{f_1x_1 + f_2x_2 + \ldots + f_kx_k}{f_1 + f_2 + \ldots + f_k},$$

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 on 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, $x_1, x_2, x_3, \ldots, x_n$ form the set of data with weights $w_1, w_2, w_3, \ldots, w_n$ respectively. Then the weighted mean,

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

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

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 *nth* root of the product of the n observations, $x_1, x_2, \ldots, x_n$ forming the data. That is,

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

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., $g_m = anti\log(g_m)$). For $k-$grouped data, the geometric mean is defined as

$$g_m = \sqrt[n]{x_1^{f_1}..x_2^{f_2}.\ \ldots\ ..x_k^{f_k}} = \left(\prod_{i=1}^{k} x_i^{f_i}\right)^{1/k}$$, where x_i'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 mean, $\bar{x} = \frac{1.0+1.1+1.2}{3} = 1.1.$

- For a given set of data, $x_1, x_2, \ldots, x_n$, 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, $x_1, x_2, \ldots, x_n$,

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

For grouped data, the harmonic mean,

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

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. . . 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,

$$M = \begin{cases} x_{\frac{1}{2}(n+11)^{th}} & , \text{ if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\frac{1}{2}n^{th}} + x_{(\frac{1}{2}+1)^{th}}\right) & , \text{ if } n \text{ is even} \end{cases}$$

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, $m = l_m + \left(\frac{n/2 - f_{cm}}{f_m} \right) C_m$, 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 the 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.

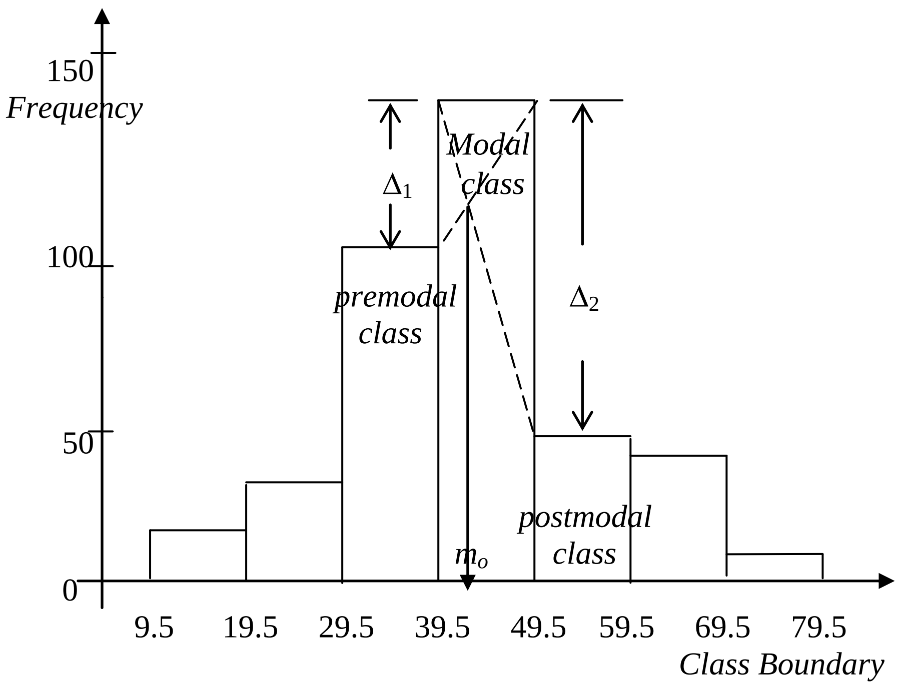
For a grouped data we shall have a set of observations occurring frequently in a particular class, called *model class*. The mode from a grouped data can only be approximately estimated by the following methods, each of which may give different value.

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

  l_o = 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,

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

$$= \frac{10.3+11+13+8.3+5.7+10+11+14+7.5+8.2+7.8+9}{12}$$

$$= \frac{115.8}{12} = 9.65\,years$$

(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
Total	-	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(a) 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

<u>*Solution:*</u>

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

$$\bar{x}_w = \frac{1}{n_w}\sum_{i=1}^{6} w_i$$

$$= \frac{4(56)+3(68)+3(65)+4(70)+3(78)+2(80)}{4+3+3+4+3+2}$$

$$= \frac{1297}{19} = 68.26$$

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

$$\bar{x} = \frac{total\ \cos t}{total\ number\ of\ items}$$

$$= \frac{18+8.5+6+25+2.5}{5}$$

$$= \frac{60.0}{5} = \text{GH¢}12.00$$

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
Total	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 give 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,

$\boxed{4}$	$\boxed{5}$	6	8	9	10	11	12	14	15	15
16	17	18	20,	$\boxed{23}$	$\boxed{26.}$					

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:

$$10\%\ \textit{trimmed mean} = \frac{6+8+9+\ldots+18+20}{13}$$

$$= \frac{161}{13} = 12.385$$

(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,

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

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

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

The harmonic mean,

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

(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

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

  $$g_m = anti\log(4.0108692) = 55.19482506$$

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

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

M

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

M

Hence the median is given by

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

$$= \frac{1}{2}\left(6^{th} \textit{observation} + 7^{th} \textit{ obervation}\right) = \frac{1}{2}(9+11) = 10$$

(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.

Length (mm)	*Frequency*	*Cum. Frequency*
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
Total	40	-

<u>*Solution:*</u>

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:

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

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

(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

<u>*Solution:*</u>

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.

Class Boundary	*Frequency*
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
Total	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, lower limit of modal class

Δ_1 = 54 – 28 = 26, Δ_2 = 54 – 57 = 3, and

C = 4.15 – 4.03 = 0.12, 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 $x_1, x_2, x_3, \ldots, x_n$ 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, $x_1, x_2, x_3, \ldots, x_n$, 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 $x_1, x_2, x_3, \ldots, x_n$ 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 ($\sigma^{2.}$), 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,

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

For grouped data,

- $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}f_1(x_i - \mu)^2 = \frac{1}{N}\left[\sum_{i=1}^{N}f_i x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N}f_i x_i\right)^2\right]$

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}f_1(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}f_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}f_i x_i\right)^2\right]$

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

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

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, $CV = \frac{standard\ deviation\,(s)}{mean\,(\bar{x})} \times 100\%$

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:

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

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{6}\begin{pmatrix}|38-8.5|+|26-28.5|+|14-28.5|+|41-28.5| \\ +|22-28.5|+|30-28.5|\end{pmatrix}$$

$$= \frac{1}{6}(9.5+2.5+14.5+12.5+6.5+1.5) = \frac{1}{6}(47) = 7.83,$$

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:

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

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

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
Total	-	60	2,210	13,873.33	95,275.0

<u>*Solution:*</u>

(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:

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

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

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

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:

$$\bar{x} = A + \sum_{i=1}^{k} f_i d_i$$
$$= 56.5 + \frac{-36}{20} = 54.7\,kg$$

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

(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

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

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

(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 $x_1, x_2, \ldots, x_n$, and the scaled marks, $y_1, y_2, \ldots, y_n$, where $y_i = ax_i + b$

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

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

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

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

$$\sigma_y = a\,\sigma_x, \qquad \ldots\ldots\ldots \quad (2.2)$$

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,

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

$$6 = 3b \qquad \ldots\ldots\ldots \quad (2.4)$$

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 $y = 2(40) - 25 = 55.$

<u>Example 2.13:</u>

(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,

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

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

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.

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

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

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

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

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

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

$$CV_A = \frac{s\tan dard\ deviation}{mean}$$

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

(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 that 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, $OD = \frac{1}{2}(Q_3 - Q_1)$.

- *Deciles*: The deciles are the values that divide the set of data into ten equal parts. They are denoted by $D_1, D_2, D_3, . . . , D_9$ and are such that 10%, 20%, 30% . . . 90% of the observations fall below $D_1, D_2, D_3, . . . , D_9$ respectively.

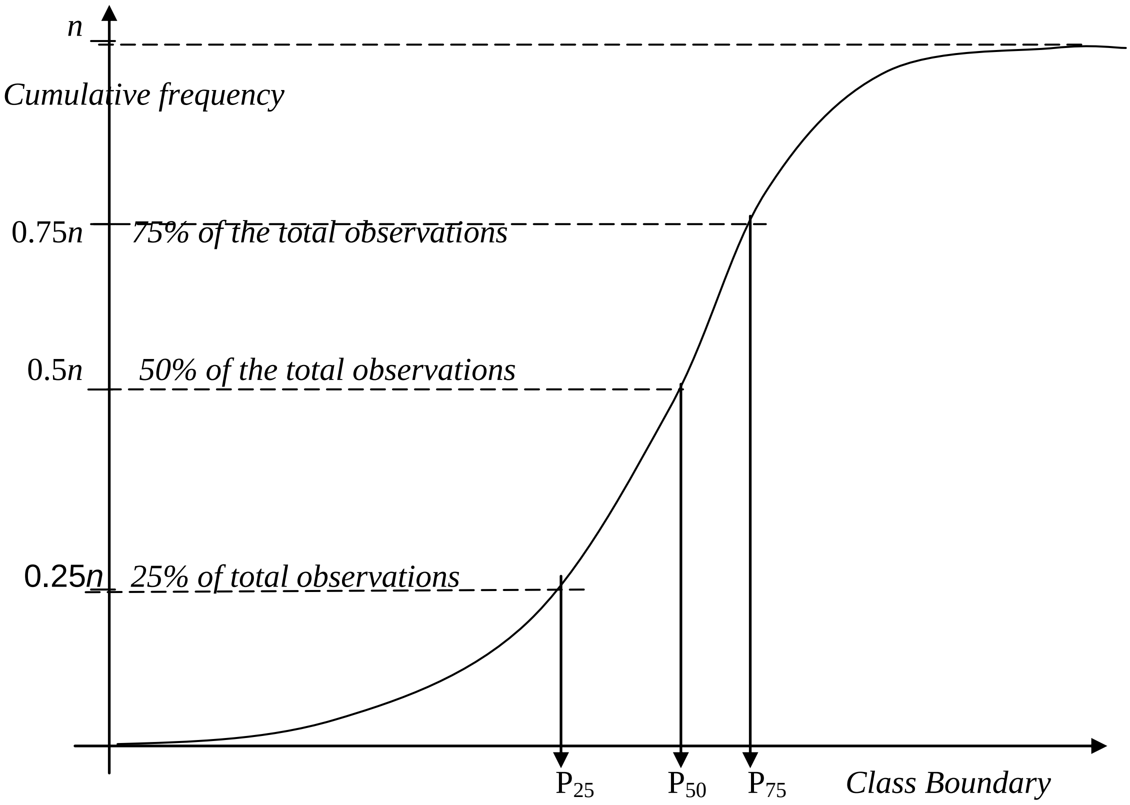
- *Percentiles*: The percentiles divide the data into 100 equal parts. They are denoted by $P_1, P_2, P_3, . . . , P_{99}$ and are such that 1%, 2% . . . 99% of the observations fall below $P_1, P_2, P_3, . . . , P_{99}$ respectively. They are used when dealing with large amount of data. The 25th percentile (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, $x_1, x_2, x_3, \ldots, x_n$, 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 k*urtosis.*

- *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 $S_k = \dfrac{3(mean - median)}{standard\ deviation} = \dfrac{3(\bar{x} - m)}{s}$, which ranges from –3 to 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).

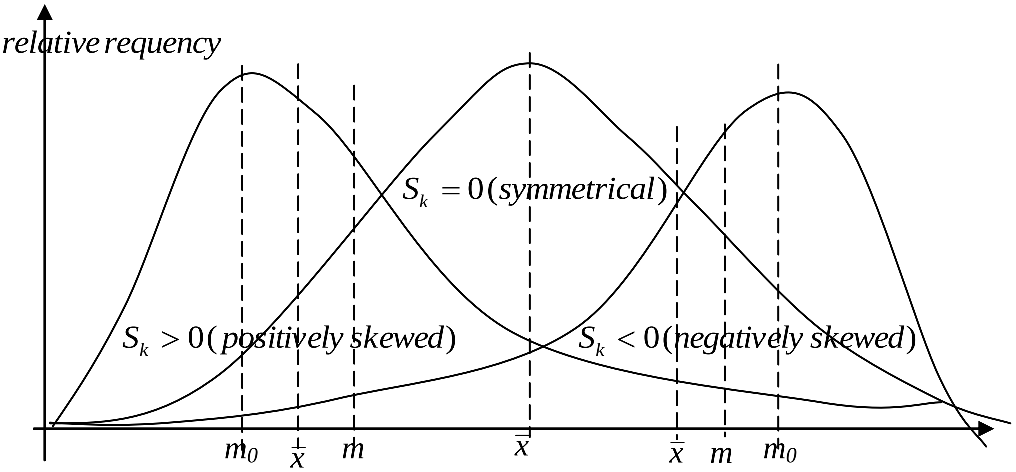


Figure 2.1: Graph of Symmetrical and Non-Symmetrical Distributions

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 $k = \dfrac{\frac{1}{2}(Q_3 - Q_1)}{P_{90} - P_{10}}$, 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.

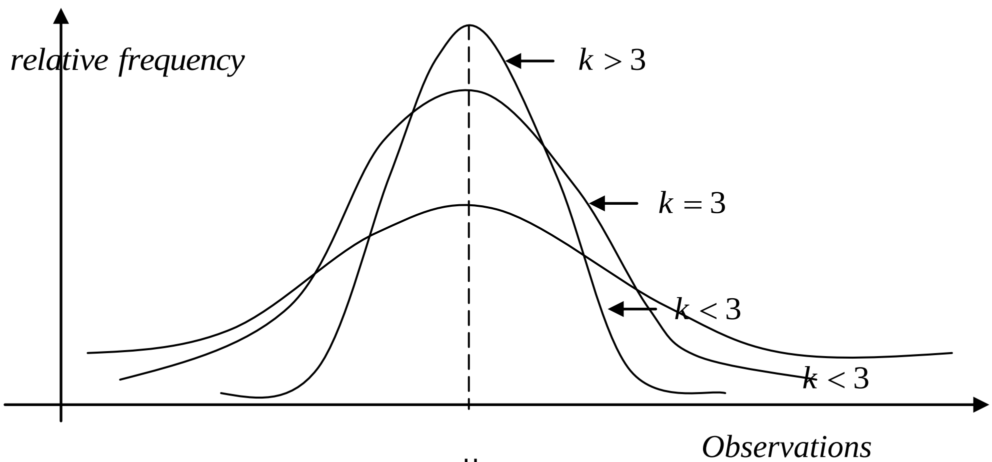


Figure 2.2: Graphs of Distributions indicating their Peakness

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 as $S_k = \dfrac{3 \quad (mean - median)}{s\tan dard \quad deviation} = \dfrac{3(28-25)}{4.2} = 2.14$,

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.

<u>*Solution:*</u>

(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:

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

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

Hence the standard deviation, s = 9.244 years

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

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

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

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

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

$$P_{10} = l_{10} + \frac{c_{10}}{f_{10}}(0.1n - F_{10}) = 9.5 + \frac{5}{11}(6-0) = 10.05 years$$

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

(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.

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

which means that the distribution slightly flattens at the centre than the normal distribution?

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.

Expenditure (x)	58 – 62	63 - 67	68 - 72	73 – 77	78 – 82
Frequency (f)	15	m	10	n	10
fx	900	$65m$	700	$75n$	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.

Chapter 3

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.

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 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 {*win (W), loss(L), 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*.

	Second Toss:					
First Toss	**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	(3,1)	(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:
  - (i) The event of observing exactly two heads *(H's)* in three tosses of a coin,
    A = {*HHT, HTH, THH*}
  - (ii) 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)}
  - (iii) 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,

(i) 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.

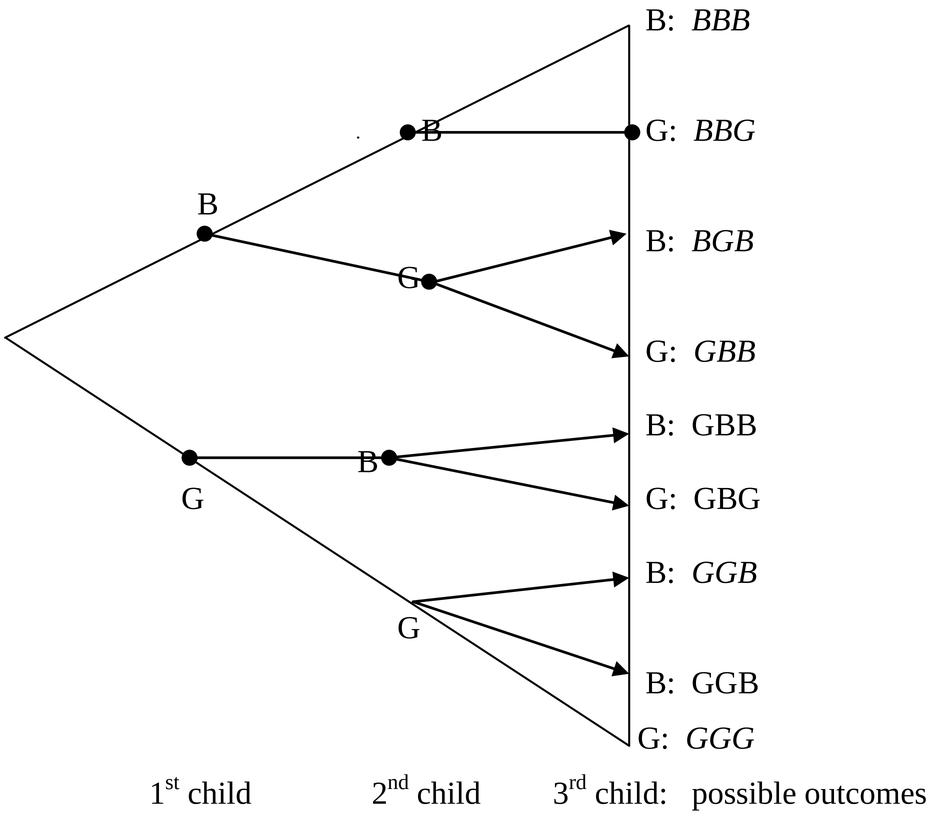


Figure 3.1: Tree Diagram of 3 Births

(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:

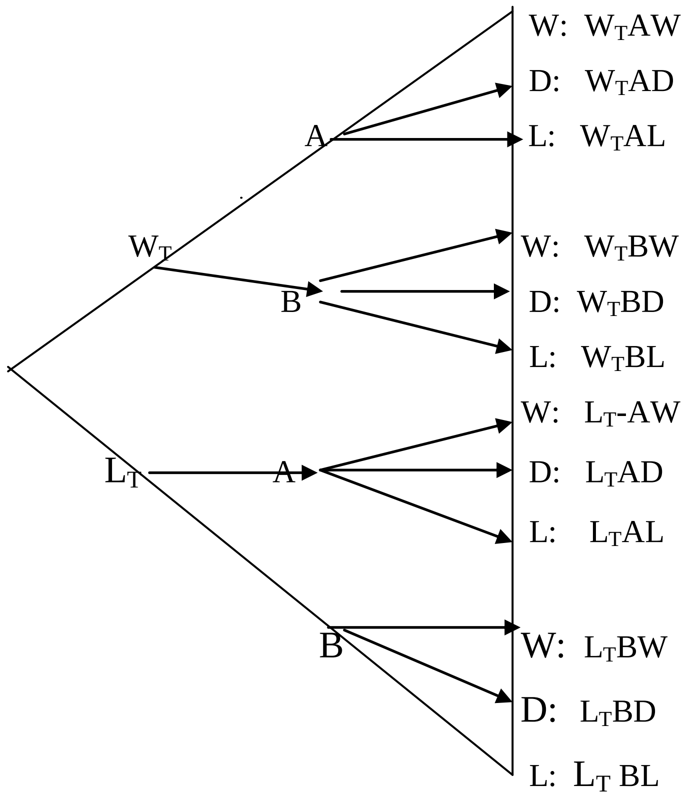


Figure 3.2: Outcomes of Soccer Match

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 $P(\overline{A}) = 1 - P(A)$, 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

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

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,

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

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 $85/120 = 0.708$, 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 = {*playing cards*}, A = {*Heart cards*}, B = {cards numbered 5}, Q = {cards with a picture of queen}. Then $n(S) = 52$, $n(A) = 13$n, $n(B) = 4$ *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 $R = \{red\ balls\}$.

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

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.

(i) The product of the scores is at least 20

<u>*Solution:*</u>

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,

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

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

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

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

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

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

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

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

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

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

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

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

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

  $A \cap B = \phi$ and $P\ (A \cap B) = 0$

- *Exhaustive Events:* Let *A*, *B* and *D* is 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 $P(A \cap B) = P(A).P(B)$.

  However, if *A* and *B* are such that, $P(A \cap B) = P(A).P(B|A),$ 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.

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

$$= \frac{n(A \cap B)}{n(B)},\ P(B) > 0$$

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?

<u>*Solution:*</u>

(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

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

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

(ii) Let B be the event of women with breast cancer and W 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 $P(M) = 1/2$, $P(S) = 2/5$ and $P(D) = 1/5$

(i) The probability they all miss target,

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

$$= \frac{1}{2} \cdot \frac{3}{5} \cdot \frac{4}{5}$$

$$= \frac{6}{25} = 024$$

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

$$P(M \cap \bar{S} \cap \bar{D}) \; or \; \bar{M} \cap S \cap \bar{D} \; or \; \bar{M} \cap \bar{S} \cap D))$$

$$= P(M \cap \bar{S} \cap \bar{D}) + P(\bar{M} \cap S \cap \bar{D}) + P(\bar{M} \cap \bar{S} \cap D)$$

$$= 1/2 \cdot 3/5 \cdot 4/5 + 1/2 \cdot 2/5 \cdot 4/5 + 1/2 \cdot 3/5 \cdot 1/5$$

$$= \frac{12}{50} + \frac{8}{50} + \frac{3}{50}$$

$$= \frac{23}{50} = 0.46$$

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

$$= 1 - P(all \; miss \; t\arg et)$$

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

$$= 1 - \frac{6}{25} = \frac{19}{25} = 0.76, \text{ [from (i)]}$$

(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:

$$P(M/E) = \frac{P(M \cap E)}{P(E)} = \frac{P(M \cap \bar{S} \cap \bar{D})}{P(E)}$$

$$= \frac{1/2 \cdot 3/5 \cdot 4/5}{23/50}$$

$$= \frac{12}{23} = 0.522$$

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

(i) The probability that both are defective,

$$P(D_1D_2) = P(D_1 \cap D_2)$$

$$= P(D_1).P(D_2|D_1)$$

$$= 4/10 . 3/9 = 2/15$$

(ii) The probability that second item is defective,

$$P(D_1D_2 \ or \ \overline{D}_1D_2) = P(D_1D_2) + P(\overline{D}_1D_2)$$

$$= P(D_1 \cap D_2) + P(\overline{D}_1 \cap D_2))$$

$$= P(D_1).\ P(D_2 / D_1) + P(\overline{D}_1).\ P(D_2 / \overline{D}_1)$$

$$= \frac{2}{15} + \frac{6}{10}.\frac{4}{9} = \frac{18}{45} = \frac{2}{5} = 0.4$$

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* $A_1, A_2, A_3, \ldots, A_n$ *is a sequence of n mutually exclusive events, then,*

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

The following theorems arise directly from the above axioms:

- *Theorem 1:* If ϕ is the empty set, the $P(\phi) = 0$.
- *Theorem 2:* If $\bar{A}$ is the complement of an event $\overline{A}$, then $P(\overline{A}) = 1 - P(A)$.
- *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
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- <u>*Corollary:*</u> For any events A_1, A_2 and A_3,

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

3-3.2.2 The Addition Rule:

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

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

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

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

3-3.2.3 The Multiplication Theorem:

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

- $P(A_1 \cap A_2) = P(A_1).P(A_2 | A_1)$
- $P(A_1 \cap A_2 \cap A_3) = P(A_1).P(A_2 | A_1).P(A_3 | A_1 \cap A_2)$
- $P(A_1 \cap A_2 \cap A_3) = P(A_1)..P(A_2 | A_1)..P(A_3 | A_1 \cap A_2) \times \ . \ . \ . \ \times P(A_n | A_1 \cap A_2 \cap \ . \ . \ . \ \cap A_{n-1})$

3-3.2.4 The Total Probability Rule:

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

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

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

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) + \ldots + P(A_n \cap B)$$

$$= P(A_1).P(B|A_1) + P(A_2).P(B|A_2) + P(A_3).P(B|A_3) + \ldots + P(A_n).P(B|A_n)$$

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

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

$$P(B_2) = 1/6, \ P(R|B_1) = 2/10$$

$$P(B_2) = 2/6, \ P(R|B_2) = 3/10$$

$$P(B_3) = 3/6, \ P(R|B_3) = 6/10$$

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

$$P(R) = P(B_1).P(R|B_1) + P(B_2).P(R|B_2) + P(B_3).P(R|B_3)$$

$$= 1/6 \cdot 2/10 + 2/6 \cdot 3/10 + 3/6 \cdot 6/10 = 13/30$$

(ii) The required probability is

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

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

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.

<u>*Solution:*</u>

(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

$$P(B) = P(M).P(B|M) + P(W).P(B|W)$$

$$= (0.60)(0.75) + (0.40)(0.45)$$

$$= 0.63$$

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

$$P(B|M) = \frac{P(M).\,P(B|M)}{P(B)}$$

$$= \frac{(0.60)(0.75)}{0.63} = \frac{5}{7} = 0.71$$

(g) Let A be the event of an increase in AFUF. Then given the probabilities, $P(C_1) = 0.45$, $P(A|C_1) = 0.60$, $P(C_2) = 0.25$, $P(A|C_2) = 0.15$, and $P(C_3) = 0.30$, $P(A|C_3) = 0.45$, , we compute the following:

(i) $$P(A) = P(C_1).P(A|C_1) + P(C_2).P(A|C_2) + P(C_3).P(A|C_3)$$

$$= (0.45)(0.60) + (0.25)(0.15) + (0.30)(0.45)$$

$$= 0.27 + 0.0375 + 0.135 = 0.4425$$

(ii) $$P(C_1|A) = \frac{P(C_1).\,P(A|C_1)}{P(A)} = \frac{P(C_1 \cap A)}{P(A)}$$

$$= \frac{(0.45).(0.60)}{0.4425} = 0.61017,$$

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?

<u>*Solution:*</u>

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

$$P(T_1T_2T_3) = P(T_1).P(T_2).P(T_3)$$

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

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

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

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

$$P(T_1T_2T_3 or T_1\overline{T}_2T_3 or \overline{T}_1\overline{T}_2T_3)$$
$$= P(T_1T_2T_3) + P(T_1\overline{T}_2T_3) + P(\overline{T}_1\overline{T}_2T_3)$$
$$= \frac{30}{50}.\frac{29}{49}.\frac{28}{48} + \frac{30}{50}.\frac{20}{49}.\frac{29}{48} + \frac{20}{50}.\frac{19}{49}.\frac{30}{48}$$
$$= 0.207 + 0.148 + 0.097 = 0.452$$

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

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

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

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

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

$$P(M|E) = \frac{P(M \cap E)}{P(E)}$$

$$= \frac{P(M).P(E|M)}{P(E)}$$

$$= \frac{4/9}{2/3} = \frac{2}{3}$$

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 $n_1 \times n_2 \times n_3 \times \ . \ . \ . \times n_k$ 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 × 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 × 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 × 4 × 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 $n! = n(n-1)(n-2)\times ...\times 3\times 2\times 1$

(ii) n distinct objects taken k at a time is ${}^{n}P_{k}$ or $P(n,k)=\dfrac{n!}{(n-k)!}$, *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 $\dfrac{n!}{n_1!.n_2!.n_3!. . .n_k!}$, where $n = n_1 + n_2 + \ . \ . \ . + n_k$

(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 2*C's*, 3I's, *2L's*, and 1's of the rest of letters $= \dfrac{14!}{2!.3!.2!}$ = 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 selected from *n* distinct objects, irrespective of their order is defined by

$${}^{n}C_{k} \ \ or \ \binom{n}{k} = \frac{n!}{(n-k)!\ k!} = \frac{{}^{n}P_{k}}{k!}$$

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?

<u>*Solution:*</u>

(i) The first, second and third numbers, each can be chosen in 5 ways. The total number of ways = 5 × 5 × 5 = 125

(ii) The number of ways of answering the questions in *Section A*

= 10 × 9 × 8 × 7 × 6 = 30,240

The number of ways of answering the questions in section *B*

= 5 × 4 × 3 = 60

The number of ways of answering the questions in section *C*

= 5 × 4 = 20

Hence the students can answer the questions in the three sections in

= 30,240 × 60 × 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?

<u>*Solution:*</u>

(i) If we regard the 2 girls as a separate persons ($\underline{B_1}\ \underline{B_2}\ \underline{B_3}\ \underline{B_4}\ \underline{G_1 G_2}$), 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.

<u>*Solution*</u>*:*

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

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

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

1B3G, 2B2G or 3B1G

The required number of ways

$$= \binom{6}{1} . \binom{5}{3} + \binom{6}{2} . \binom{5}{2} + \binom{6}{3} . \binom{5}{1}$$

$= 6\ \ (10) + 15\ (10) + 20\ (5) = 130$

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?

<u>*Solution:*</u>

(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

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

(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

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

(γ) The number of ways of forming a 5-member committee = $\binom{12}{5}$

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 than 4 parents = $\binom{12}{5} - \binom{6}{5}$ = 126 – 6 = 120

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

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

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

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

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

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

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

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

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

(ii) The probability that more women than men

$$= (2W\ 1M)+(3W)$$

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

(iii) The probability that a particular women is included

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

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

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

$$= \frac{^{6}C_3}{^{14}C_3} = \frac{6\times5\times4}{14\times13\times12} = \frac{5}{91}$$

(ii) $\Pr(2\,red\ and\ 1\,white\ balls) = \dfrac{{}^{6}C_2 . {}^{3}C_1}{{}^{14}C_3} = \dfrac{45}{364}$

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

(iv) $P(one\,of\ each\,colour)$

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

4-3.4 Trial Questions 1-3:

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

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

$P(\overline{A} \cap \overline{B}) = P(\overline{A}).P(\overline{B})$

(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 $1/4$ and the corresponding probability for his wife is $1/3$. 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 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
- c*onditional 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?

Chapter 4

PROBABILITY DISTRIBUTIONS

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).

- <u>*Definition 4.1:*</u>

  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.

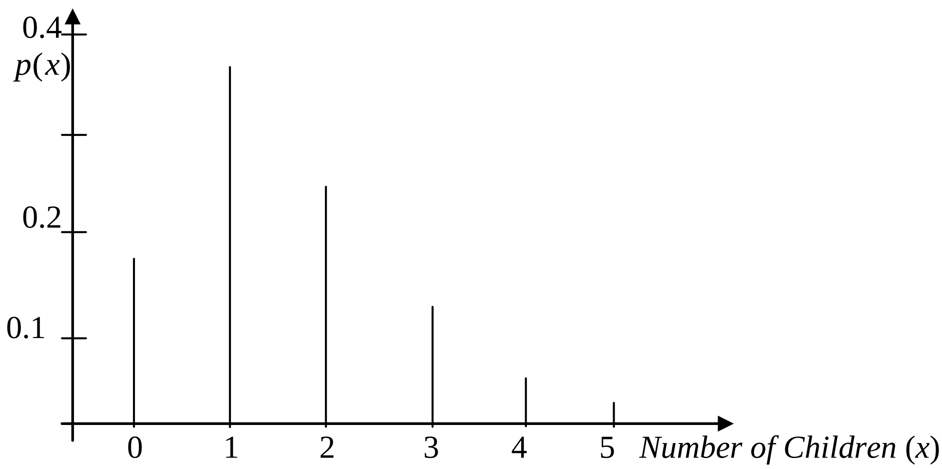
1-4.2.2 Types of Probability Distributions

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:



- <u>*Definition 4.4:*</u>

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).*

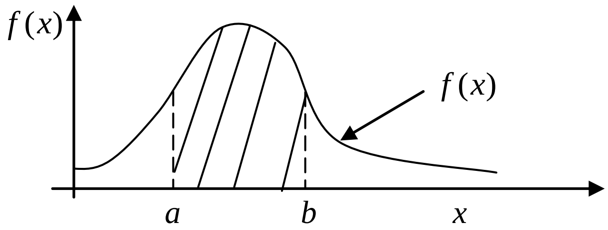


Figure 4.1: Graph of probability density function

That is, from the diagram above,

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

- <u>*Definition 4.4*</u>*:*

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) $f(x) \ge 0,$ *for any value of x.*

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

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

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

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 $F(x) = P(X \le x)$.

  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 $-\infty \leq x \leq \infty$, $f(x) = \dfrac{dF(x)}{dx}$ 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, $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$

(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, TTHH, TTHT, TTHH, TTTT}

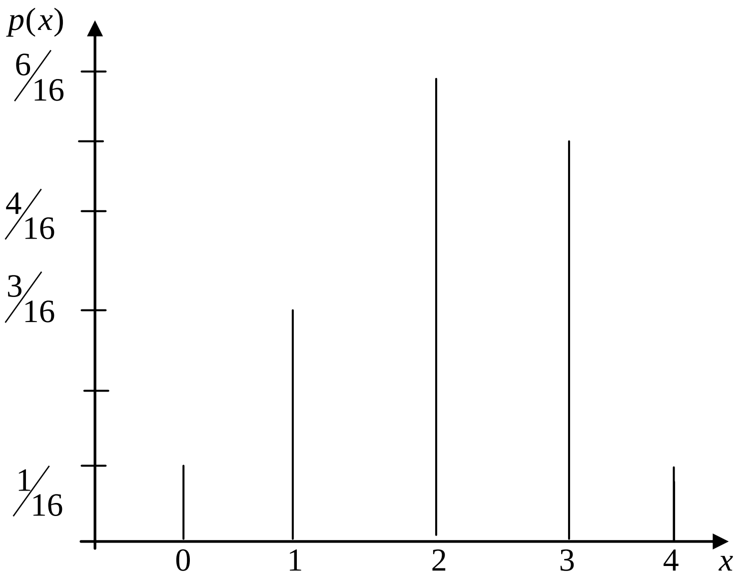
The random variable, X is the number of heads occurring in that experiment which assumes the values, X = 0, 1, 2, 3, 4. The required probability distributions is

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:



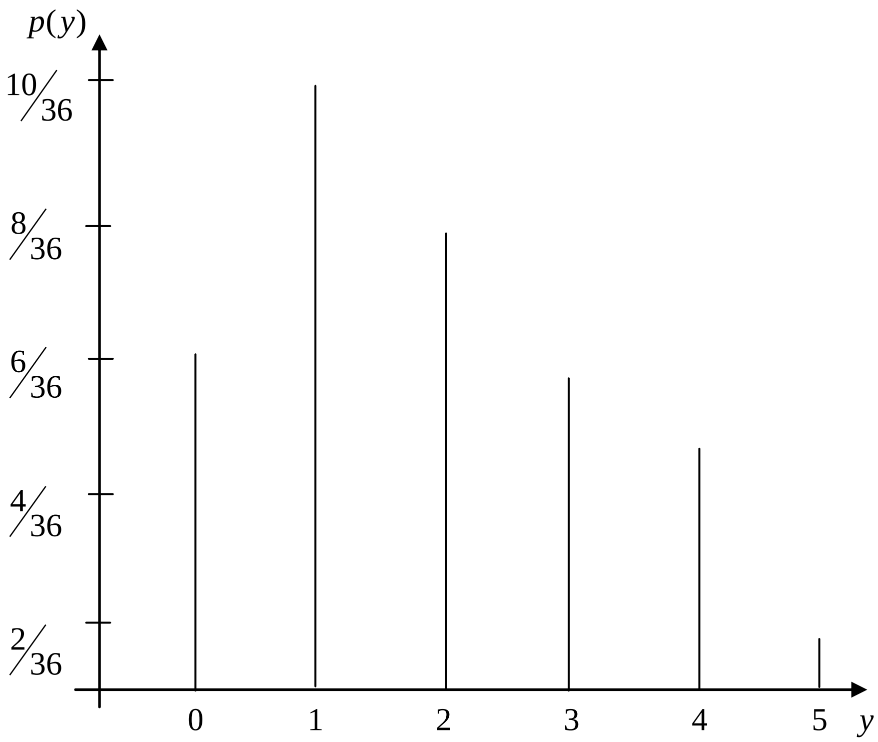
(ii) The sample space for the experiment is given by the table below.

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
$p(y)$	$^6/_{36}$	$^{10}/_{36}$	$^8/_{36}$	$^6/_{36}$	$^4/_{36}$	$^2/_{36}$

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

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

<u>*Solution:*</u>

(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) $P(x \geq 3) = \sum_{x=3}^{6} p(x)$

$= P\ (3) + P\ (4) + P\ (4) + P\ (6)$

$= 0.20 + 0.10 + 0.15 + 0.05 = 0.50$

(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,

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

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

$9k = 1 \Leftrightarrow k = {}^{1}\!/_{9}$

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

$$f(x) = \begin{cases} {}^{1}\!/_{6}\,x + k, & 0 \leq x \leq 3 \\ 0\quad, & elsewhere \end{cases}$$

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

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

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

<u>*Solution:*</u>

(i) Given the probability density function,

$$f(x) = \begin{cases} {}^{1}\!/_{6}\,x + k, & 0 \leq x \leq 3 \\ 0\quad, & elsewhere \end{cases}$$

then,

(i) $\int_0^3 f(x)dx = 1$

$$\int_0^3 \left(\tfrac{1}{6} x + k\right) dx = 1$$

$$\tfrac{1}{12} x^2 + kx\Big|_0^3 = 1$$

$$\left[\tfrac{1}{12} (3)^2 + 3k\right]_0^3 - 0 = 1$$

$$\tfrac{3}{4} + 3k = 1$$

$$3k = \tfrac{1}{4} \Leftrightarrow k = \tfrac{1}{12}$$

Hence, $f(x) = \begin{cases} \tfrac{1}{12}(2x+1), & 0 \le x \le 3 \\ 0 & , \quad elsewhere \end{cases}$

$$P(1 \le x \le 2) = \int_1^2 \tfrac{1}{12}(2x+1)\, dx$$

$$= \tfrac{1}{12}\left[x^2 + x\right]_1^2$$

$$= \tfrac{1}{12}\left[\left(2^2 + 2\right) - \left(1^2 + 1\right)\right]$$

$$= \tfrac{1}{12}(6 - 2) = \tfrac{1}{3}$$

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

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

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

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

$$\frac{9k}{2} + 3k\left[(16-8) - (12 - {}^9/_2)\right] = 1$$

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

$$6k = 1$$

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

Hence, $f(x) = \begin{cases} \tfrac{1}{6}x & , \ 0 \le x \le 3 \\ \tfrac{1}{2}(4-x), & 3 < x \le 4 \\ 0 & , \ elsewhere \end{cases}$

$$P(1 \le x \le 2) = \int_1^2 f(x)dx$$

$$= \int_1^2 \tfrac{1}{6} x\,dx = \left| \tfrac{x^2}{12} \right|_1^2$$

$$= \tfrac{1}{12}\left(2^2 - 1\right) = \tfrac{1}{4}$$

$$P(x > 2) = \int_2^4 f(x)dx$$

$$= \int_2^3 \tfrac{1}{6}\, x\,dx + \int_3^4 \tfrac{1}{2}(4 - x)dx$$

$$= \left| \tfrac{x^2}{12} \right|_2^3 + \tfrac{1}{2}\left| 4x - \tfrac{x^2}{2} \right|_3^4$$

$$= \tfrac{1}{12}(9-4) + \tfrac{1}{2}(16-8) - \tfrac{1}{2}\left(12 - \tfrac{9}{2}\right)$$

$$= \tfrac{5}{12} + \tfrac{1}{4}$$

$$= \tfrac{2}{3}$$

4.2(e) Given the probability mass function,

x	0	1	2	3
$p(x)$	$\tfrac{1}{4}$	$\tfrac{1}{8}$	$\tfrac{1}{2}$	$\tfrac{1}{8}$

Find the cumulative distribution function.

<u>*Solution:*</u>

(a) $$F(x) = P(X \le x) = \sum_{x=0}^{3} p(x)$$

$$F(0) = P(X \le 0) = p(0) = \tfrac{1}{4}$$

$$F(1) = P(X \le 1) = p(0) + p(1) = \tfrac{1}{4} + \tfrac{1}{8} = \tfrac{3}{8}$$

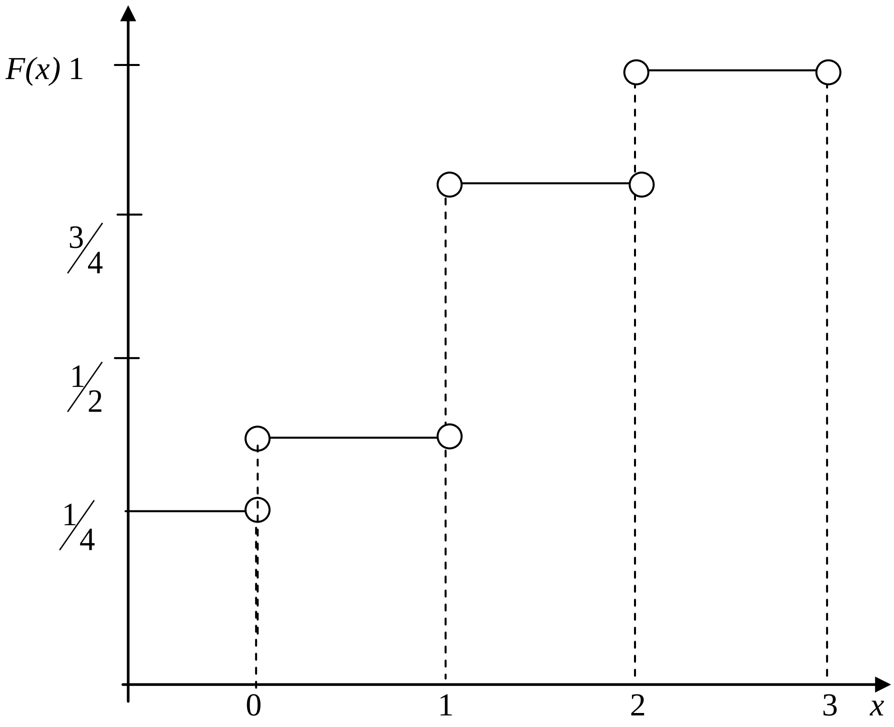
$$F(2) = P(X \le 2) = p(0) + p(1) + p(2) = \tfrac{1}{4} + \tfrac{1}{8} + \tfrac{1}{2} = \tfrac{7}{8}$$

$$F(3) = P(X \le 3) = p(0) + p(1) + p(2) + p(3) = \tfrac{1}{4} + \tfrac{1}{8} + \tfrac{1}{2} + \tfrac{1}{8} = 1$$

Hence the cumulative distribution is

X	0	1	2	3
$F(x)$	$\tfrac{1}{4}$	$\tfrac{3}{8}$	$\tfrac{7}{8}$	1

which is illustrated graphically below:



4.2(f) For a discrete random variable *x* with cumulative distribution function is given by $F(x) = kx, \quad x = 1, 2, 3,$

(i) Find the value of the constant *k*, and

(ii) Evaluate the probability $P(x < 3)$. Hence determine the probability distribution of *x*.

<u>*Solution*</u>

(i) From the properties of cumulative distribution function,

$$F(3) = P(x \le 3) = 1$$

$$3k = 1 \quad \Leftrightarrow \quad k = \tfrac{1}{3}$$

Hence, $F(x) = \tfrac{1}{3}x \quad \textit{for } 1, 2, 3$

(ii) $P(x < 3) = F(2) = \tfrac{1}{3}(2) = \tfrac{2}{3})$

The probability distribution function of *x*, *p(x)* is obtained as follows:

$$P(x = 1) = F(1) - F(0) = \tfrac{1}{3} - 0 = \tfrac{1}{3} = p(1)$$

$$P(x = 2) = F(2) - F(1) = \tfrac{2}{3} - \tfrac{1}{3} = \tfrac{1}{3} = (2)$$

$$P(x = 3) = F(3) - F(2) = 1 - \tfrac{2}{3} = \tfrac{1}{3} = p(3)$$

X	0	1	2
$p(x)$	$1/3$	$1/3$	$1/3$

or $p(x) = \frac{1}{3}x, \quad for\ x = 1, 2, 3.$

4.2(g) Find the cumulative distribution functions of the following probability density functions:

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

(iii) $f(x) = \begin{cases} \frac{1}{2}e^{-x/2}, & x > 0 \\ 0 & , elsewhere \end{cases}$

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

Sketch the graph of *F(x)* in each case.

<u>*Solution:*</u>

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

$$F(x) = \int_0^x f(t)dt = \int_0^x \frac{1}{2}t\,dt$$

$$= \frac{1}{4}x^2,\ for\ 0 \le x \le 2$$

$$= 1 \quad , if\ x > 2\ and\ 0, elsewhere.$$

Hence, $F(x) = \begin{cases} \frac{1}{4}x^2, & 0 \le x \le 2 \\ 1 & , \quad x > 2 \\ 0 & , \quad x < 0 \end{cases}$

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

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

$$F_1(x) = \int_0^x t\ dt = \frac{1}{2}x^2,\ 0 \le x \le 1$$

$$F_2(x) = F_1(1) + \int_1^x \frac{1}{2}dt = \frac{1}{2}x,\ \ 1 < x \le 2$$

$F_3(x) = F_2(2) + 0 = 1, \ \ x > 2$

$$F(x) = \begin{cases} \frac{1}{2}x^2, & 0 \le x \le 1 \\ \frac{1}{2}x, & 1 < x \le 2 \\ 1, & x > 2 \\ 0, & elsewhere \end{cases}$$

(iii) $$f(x) = \begin{cases} \frac{1}{2}e^{-x/2}, & x > 0 \\ 0, & elsewhere \end{cases}$$

$$F(x) = \int_0^x f(t)dt$$

$$= \int_0^x \frac{1}{2}e^{-t/2}\,dt$$

$$= \begin{cases} 1 - e^{-x/2}, & x > 0 \\ 0, & x \le 0 \end{cases} \quad x > 0$$

SESSION 2-4: EXPECTATION OF RANDOM VARIAVLES

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) $\mu = E(x) = \sum_x x\, p(x),$ if *x* is discrete.

  (ii) $\mu = E(x) = \int_{-\infty}^{\infty b} x\, f(x)\, dx,$ if *x* 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

  $\sigma^2 = Var(x) = E[(x-\mu)^2] = E(x^2) - \mu^2,$ where

  (i) $Var(x) = \sum_x (x-\mu)^2 p(x)$

  $= \sum_x x^2 p(x) - \mu^2$, if *x* is discrete.

  (ii) $Var(x) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$

  $= \int_a^b x^2 f(x)\, dx - \mu^2$, if *x* is continuous.

  The *standard deviation* of *x* is the square root *x*... That is,

  σ = $\sqrt{Var(x)}$

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) $\sum_x^m p(x) = \frac{1}{2}$ (or close to it), if *x* is discrete.

  (ii) $\int_a^m f(x)\, dx = \frac{1}{2}$, if *x* 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.

<u>Example 4.3:</u>

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) $$f(x)=\begin{cases}6x(1-x), & 0<x<1\\ 0 & , elsewhere\end{cases}$$

<u>*Solution:*</u>

(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,

$$Var(x) = \sum_{x=1}^{5}(x-\mu)^2\ p(x) = \sigma^2$$

$$= (1-3)^2\ (0.1) + (2-3)^2\ 0.3 + (3-3)^2\ (0.2) + (4-3)^2(0.3)$$

$$+ (5-3)^2(0.1) = 0.4 + 0.3 + 0 + 0.3 + 0.4 = 1.4, \text{ or}$$

$$Var(x) = Var(x) = \sum_{x=1}^{5} x^2\,p(x) - \mu^2 = \sigma^2$$

$$= 1^2(0.1) + 2^2(0.3) + 3^2(0.2) + 4^2(0.3) + 5^2(0.1) - (3)^2$$

$$= 0.1 + 1.2 + 1.8 + 4.8 + 2.5 - 9 = 1.4$$

Hence the standard deviation, $\sigma = \sqrt{1.4} = 1.1832$

(ii) Given the probability density function,

$$f(x)=\begin{cases}6x(1-x)\,, & 0<x<1\\ 0 & , elsewhere\end{cases}$$

The mean of x is $\mu = E(x) = \int_0^1 x\,f(x)dx$, where

$$\int_0^1 xf(x)\,dx = \int_0^1 6x^2(1-x)\,dx$$

$$= \int_0^1 (6x^2 - 6x^3)\,dx$$

$$= \left[\frac{6}{3}x^3 - \frac{6}{4}x^4\right]_0^1$$

$$= 2(1)^3 - \frac{3}{2}(1)^4 - 0$$

$$= 2 - \tfrac{3}{2} = \tfrac{1}{2} = 0.5$$

The variance of x,

$\sigma^2 = Var(x) = E(x^2) - \mu^2$

$$= \int_0^1 x^2 f(x)dx - \mu^2$$

$$= \int_0^1 6x^3(1-x)\,dx - (0.5)^2$$

$$= \int_0^1 \left(6x^3 - 6x^4\right)dx - 0.25$$

$$= \left[\frac{6}{4}x^4 - \frac{6}{5}x^5\right]_0^1 - 0.25$$

$$= \tfrac{3}{2} - \tfrac{6}{5} - 0.25$$

$$= \tfrac{3}{10} - \tfrac{1}{4} = \tfrac{1}{20} = 0.05$$

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,

$$f(x) = \begin{cases} x & , 0 \le x \le 1 \\ \tfrac{1}{2} & , 1 < x \le 2 \\ 0 & , otherwise. \end{cases}$$

(i) Determine the mean, median and the variance.

(ii) Find the mean and median of the distribution,

$$p(x) = \begin{cases} \tfrac{1}{9}(x-1), & x = 3, 4, 5 \\ 0 & , elsewthere \end{cases}$$

Solution:

(i) Given $f(x) = \begin{cases} x \ , 0 \le x \le 1 \\ \frac{1}{2}, \ 1 < x \le 2 \\ 0 \ , \ otherwise \end{cases}$

$$\mu = E(x) = \int_0^2 x\, f(x)\, dx$$

$$= \int_0^1 x^2\, dx + \int_1^2 \tfrac{1}{2} x\, dx$$

$$= \left[\tfrac{1}{3}x^3\right]_0^1 + \left[\tfrac{1}{4}x^2\right]_1^2$$

$$= \tfrac{1}{3}(1)^3 + \tfrac{1}{4}(2)^2 - \tfrac{1}{4}(1)^2$$

$$= \tfrac{1}{3} + 1 - \tfrac{1}{4}$$

$$= \tfrac{13}{12} = 1.083$$

Let m be the median. Then,

$$\int_0^1 x\, dx = \tfrac{1}{2}x^2\Big|_0^1 = \tfrac{1}{2}$$

$$\int_1^m \tfrac{1}{2}\, dx = \tfrac{1}{2} - \int_0^1 x\, dx$$

$$\tfrac{1}{2}x\Big|_1^m = \tfrac{1}{2} - \tfrac{1}{2}$$

$$\tfrac{1}{2}m - \tfrac{1}{2} = 0 \quad \Leftrightarrow \ m = 1$$

The variance, $Var(x)$

$$Var(x) = E(x^2) - \mu^2$$

$$= \int_0^2 x^2\, f(x)\, dx - \mu^2$$

$$= \int_0^1 x^3\, dx + \int_1^2 \tfrac{1}{2}\, dx - \left(\tfrac{13}{12}\right)^2$$

$$= \tfrac{1}{4}x^4\Big|_0^1 + \tfrac{1}{2x}\Big|_1^2 - \tfrac{169}{144}$$

$$= \tfrac{1}{4} + \tfrac{8}{6} - \tfrac{1}{6} - \tfrac{169}{144}$$

$$= \frac{35}{144} = 0.243$$

(ii) $p(x) = \begin{cases} \frac{1}{9}(x-1), & x = 3, 4, 5 \\ 0, & otherwise \end{cases}$

$$\mu = E(x) = \sum_{x=3}^{5} x\, p(x)\, \Sigma\, x\, p(x)$$

$$= \frac{2}{9}(3-1) + \frac{4}{9}(4-1) + \frac{5}{9}(5-1)$$

$$= \frac{1}{9}\{6 + 12 + 20\} = \frac{38}{9} = 4.22$$

For the median, m we have $P(x \leq m) \leq 0.5$ or ≥ 0.5

$P(x \leq 4) = p(3) + p(4)$ or $P(x \geq 5) = p(5)$

$= \frac{1}{9}(2) + \frac{1}{9}(3)$ or $= \frac{1}{9}(4)$

$= \frac{5}{9} = 0.56$ or $= \frac{4}{9} = 0.44$

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,

$$f(y) = \begin{cases} y & , 0 \leq y < \frac{1}{2} \\ \lambda(4-y) & , \frac{1}{2} \leq y \leq 4 \\ 0 & , elsewhere \end{cases}$$

(i) Find the value of λ, and

(ii) Use it to determine the mean, median and the standard deviation.

<u>Solution:</u>

(i) To find λ we have,

$$\int_0^4 f(y)\,dy = 1$$

$$\int_0^{1/2} y\,dy + \lambda \int_{1/2}^4 (4-y)\,dy = 1$$

$$\frac{1}{2}y^2\Big|_0^{1/2} + \lambda\left[4y - \frac{1}{2}y^2\right]_{1/2}^4 = 1$$

$$\frac{1}{8} + \lambda\left\{\left[4(4) - \frac{1}{2}(4)^2\right] - \left[4\left(\frac{1}{2}\right) - \frac{1}{2}\left(\frac{1}{2}\right)^2\right]\right\} = \frac{1}{8} + \lambda\left[(16-8) - \left(2 - \frac{1}{8}\right)\right] = 1$$

$$\frac{49}{8}\lambda = \frac{7}{8} \Leftrightarrow \lambda = \frac{1}{7}$$

$$\text{Hence, } f(y) = \begin{cases} y & , \ 0 \le y < \tfrac{1}{2} \\ \tfrac{1}{7}(y - y) & , \ \tfrac{1}{2} \le y \le 4 \\ 0 & , \ elsewhere \end{cases}$$

(ii) $\mu = E(y) = \int_0^{1/2} y^2 dy + \int_{1/2}^4 \frac{1}{7} y\,(4 - y)\, dy$

$= \frac{1}{3} y^3 \Big|_0^{1/2} + \int_{1/2}^4 \frac{1}{7}\left(4y - y^2\right)dy$

$= \frac{1}{24} + \frac{1}{7}\left|\left(2y^2 - \frac{1}{3} y^3\right)\right|_{1/2}^4$

$= \frac{1}{24} + \frac{1}{7}\left[\left(32 - \frac{64}{3}\right) - \left(\frac{1}{2} - \frac{1}{24}\right)\right]$

$= \frac{256}{168} = \frac{3}{2} = 1.5$

Let m be the median. Then

$\int_0^{1/2} y\, dy = \left[\frac{1}{2} y^2\right]_0^{1/2} = \frac{1}{8}$

$\int_0^{1/2} y\, dy + \int_{1/2}^m \frac{1}{7}(4 - y)dy = \frac{1}{2}$

$\int_{1/2}^m \frac{1}{7}(4 - y)dy = \frac{1}{2} - \frac{1}{8}$

$\frac{1}{7}\left[4y - \frac{1}{2} y^2\right]_{1/2}^m = \frac{3}{8}$

$\frac{1}{7}\left(4m - \frac{1}{2} m^2\right) - \frac{1}{7}\left(2 - \frac{1}{8}\right) = \frac{3}{8}$

$$4m - \frac{1}{2} m^2 = 9 \iff m^2 - 8m + 9 = 0$$

Solving we have $m = 1.35$.

For the standard deviation, σ, we have

$\sigma^2 = Var(x) = E(x^2) - \mu^2$

$= \int_0^{y_2} y^3\, dy + \int_{1/2}^4 \frac{1}{7} y^2 (4 - y)\, dy - (1.5)^2$

$= \left[\frac{1}{4} y^4\right]_0^{1/2} + \frac{1}{7}\left[\left(\frac{4}{3} y^3 - \frac{1}{4} y^4\right)\right]_{1/2}^4 - 2.25$

$= \frac{73}{24} - 2.25 = \frac{19}{24} = 0.79167$

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

$$f(y) = \begin{cases} ky^2 + y & , \ 0 \le y \le 2 \\ 0 & , \ elsewthere \end{cases}$$

(i) Find the value of k. Compute the mean value of y and its standard deviation.

(ii) Find the probability, $P(1 \le y \le 1.5)$.

2.(a) Find the mean and the standard deviation of the following distributions,

(i) $$p(x) = \begin{cases} \frac{k}{2}(x-2) & , \ x = 1, \ 2, \ 3, \ 4 \\ 0 & , \ elsewthere \end{cases}$$

(ii)

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

$$f(y) = \begin{cases} \lambda y^2(1-y) & , \ 0 \le y \le 1 \\ 0 & , \ elsewhere \end{cases}$$

(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,

$$f(x) = \begin{cases} kx & , \ 0 \le x \le 2 \\ 2k(3-x) & , \ 2 < x \le 3 \\ 0 & , \ elsewhere \end{cases}$$

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,

$$f(x) = \begin{cases} \frac{1}{2}(x-2), & 2 \le x \le 3 \\ a & ,3 < x \le 5 \ and \ a > 0 \\ 2-bx & ,5 < x \le 6 \ and \ b > 0 \\ 0 & ,elsewhere \end{cases}$$

(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.

Chapter 5

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

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

n *where* $0 < p < 1$ *and the mean and variance are respectively,*

$$\mu = E(x) = np \ \ and \ \ \sigma^2 = Var(x) = np(1-p)$$

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.

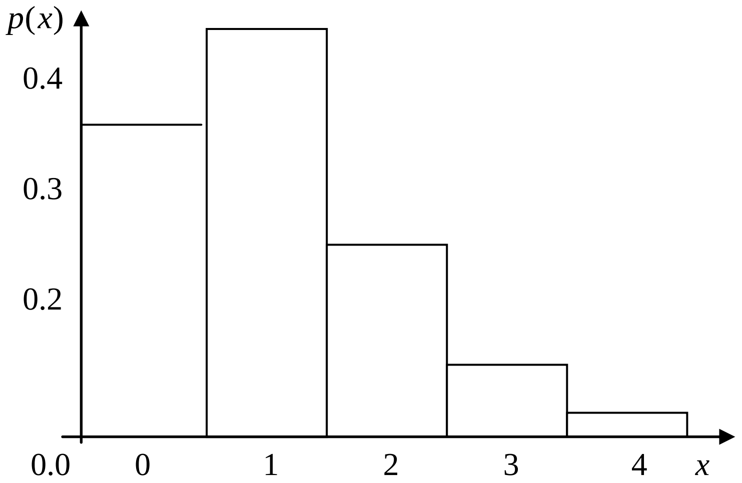


Figure 5.1: *Binomial distribution for* $n = 10$ *and* $p = 0.10$

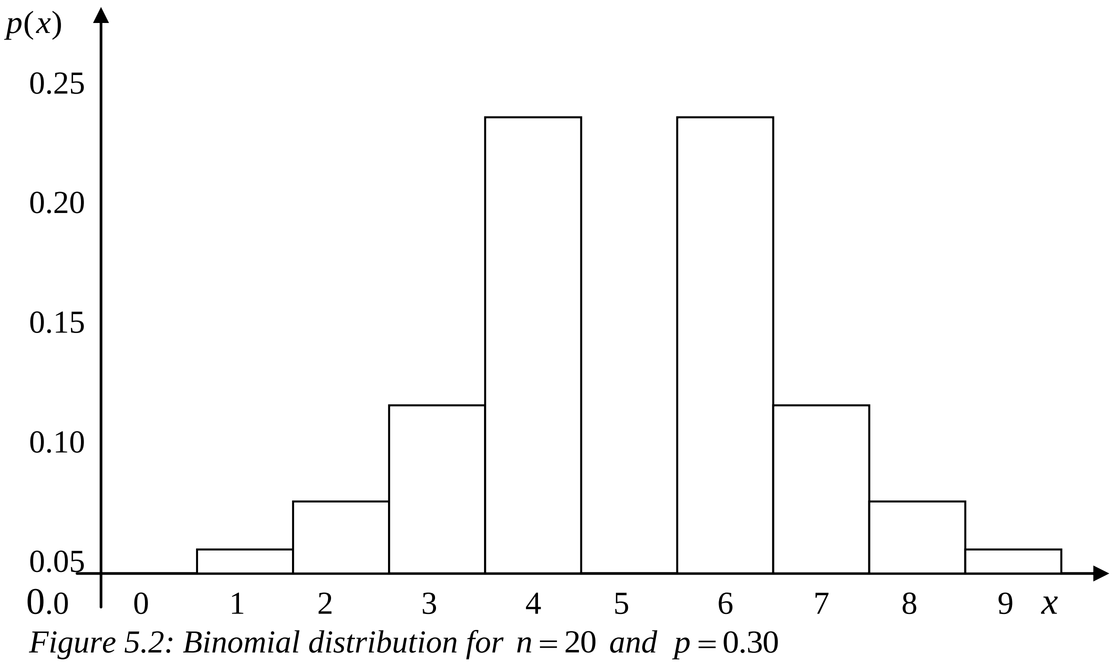


Figure 5.2: Binomial distribution for $n = 20$ *and* $p = 0.30$

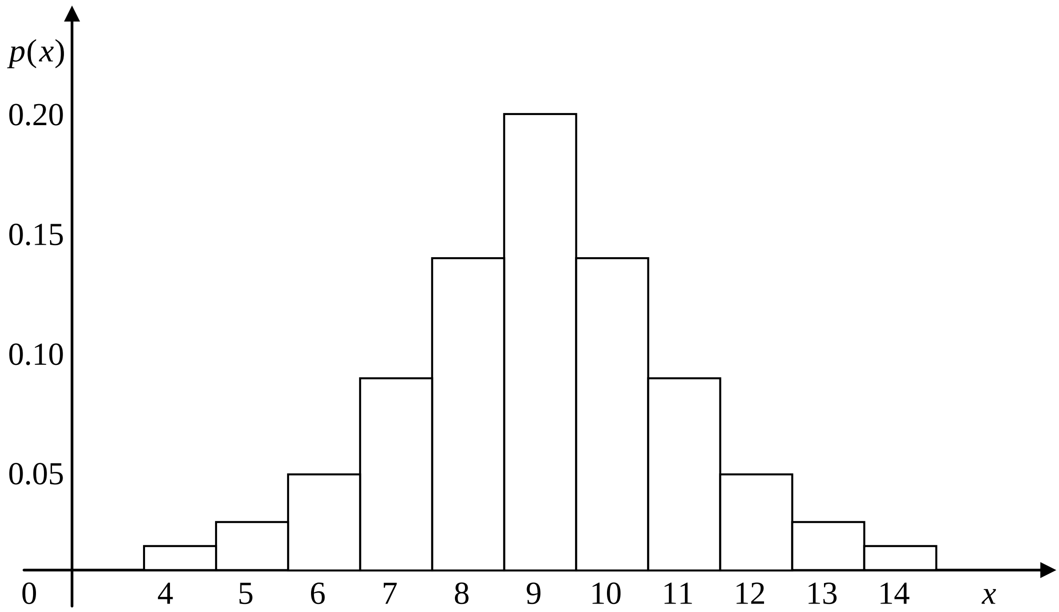


Figure 5.3: Binomial distribution for $n = 50$, $p = 0.50$

The term *binomial experiment* is derived from the fact that the probabilities, *p(x)* at *x* = 0, 1, 2. . . *n* are terms of the binomial expansion,

$$(p+q)^n = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}$$

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 is 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.

<u>Example 5.1:</u>

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 $x =$ number of persons with blood group *A*. Thus $p(x) = \binom{20}{x}(0.25)^x(0.75)^{20-x},\ x = 0, 1, 2, \ldots, 20.$

(i) $$p(2) = \binom{20}{2}(0.25)^2(0.75)^{18} = 190(0.0625)(0.00563771) = 0.066947808$$

(ii) $$P(x \geq 3) = 1 - P(x \leq 2)$$
$$= 1 - \sum_{x=0}^{2}\binom{20}{x}(0.25)^x(0.75)^{20-x}$$
$$= 1 - \left\{\binom{20}{0}(0.25)^0(0.75)^{20} + \binom{20}{1}(0.25)^1(0.75)^{19} + \binom{20}{2}(0.25)^2(0.75)^{18}\right\}$$
$$= 1 - 0.09126043 = 0.90873957$$

(iii) $$P(x < 2) = \sum_{x=0}^{1}\binom{20}{x}(0.25)^x(0.75)^{20-x}$$
$$= \binom{20}{0}(0.25)^0(0.75)^{20} + \binom{20}{1}(0.25)^1(0.75)^{19} = 0.024312625$$

(b) The number of correct answers, $x \sim B\left(15, \frac{1}{5}\right)$. Then required probability is

$$P(x \leq 3) = \sum_{x=0}^{3}\binom{15}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{15-x}$$
$$= \binom{15}{0}\left(\frac{1}{5}\right)^0\left(\frac{4}{5}\right)^{15} + \binom{15}{1}\left(\frac{1}{5}\right)^1\left(\frac{4}{5}\right)^{14} + \binom{15}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^{13}\binom{15}{3}\left(\frac{1}{5}\right)^3\left(\frac{4}{5}\right)^{12}$$
$$= \left(\frac{4}{5}\right)^{15} + 15\left(\frac{1}{5}\right)\left(\frac{4}{5}\right)^{14} + 105\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^{13} 455\left(\frac{1}{5}\right)^3\left(\frac{4}{5}\right)^{12}$$
$$= 0.03518 + 0.13194 + 0.23090 + 0.25014$$
$$= 0.64816$$

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?

<u>*Solution:*</u>

(c) Let the number of defective devices among the 20 randomly selected from a shipment be y and so $y \sim B(n = 20, p = 0.03)$.

(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

$$P(y \le 2) = \sum_{y=0}^{2}\binom{20}{y}(0.03)^y(0.97)^{20-y}$$

$$= \binom{20}{0}(0.03)^0(0.97)^{20} + \binom{20}{1}(0.03)^1(0.97)^{19} + \binom{20}{2}(0.03)^2(0.97)^{18}$$

$$= (0.97)^{20} + 20(0.03)(0.97)^{19} + 190(0.0009)(0.97)^{18}$$

$$= 978991643$$

(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

$C = 10,000 + 25,000y,$ (in dollars), where the expected cost is

$$\begin{aligned} E(C) &= E(10,000 + 25,000y) \\ &= 10,000 + 25,000E(y) \\ &= 10,000 + 25,000(5)(0.15) = \$28,750 \end{aligned}$$

(iii) The probability that 0 or 1 successful exploration is made,

$$\begin{aligned} P(x = 0\, or\, x = 1) &= P(x = 0) + P(x = 1) \\ &= \binom{5}{0}(0.15)^0(0.85)^5 + \binom{5}{1}(0.15)^1(0.85)^4 \\ &= (0.85)^5 + 5(0.15)(0.85)^4 = 0.8352 \end{aligned}$$

$$= (0.85)^5 + 5(0.15)(0.85)^4 = 0.4437 + 0.3915 = 0.8352$$

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:*

$$p(y) = \binom{y-1}{k-1} p^k (1-p)^{y-k}, \quad y = k,\, k+1,\, k+2, \ldots \;;\; 0 < p < 1$$

where mean and variance given as follows: $E(y) = \dfrac{k}{p}$ *and* $Var(y) = \dfrac{k(1-p)}{p^2}$

(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*

$$p(y) = p(1-p)^{y-1},\ \ y = 1, 2, 3, \ldots;\ 0 < p < 1$$

where mean and variance become, $E(y) = \dfrac{1}{p}$ and $Var(y) = \dfrac{1-p}{p^2}$

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 member 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

$$P(y=6) = \binom{y-1}{k-1} p^k (1-p)^{y-k},\ \textit{where } k = 3$$

$$= \binom{6-1}{3-1}(0.2)^3(0.8)^3 = 10(0.2)^3(0.8)^3 = 0.04096$$

(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,

$$P(x=6) = \binom{6-1}{3-1}.(0.9)^3(0.1)^3 = 0.00729$$

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

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}, \; x = 0, 1, 2, \ldots, \textit{ and } \; \mu > 0,$$

where the mean and variance are the same. That is, $E(x) = \mu = Var(x)$.

The distribution of x *may simply be denoted as* $x \sim P(\mu)$.

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 $(n \to \infty)$ 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 integer* x, $\lim_{n \to \infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{\mu^x e^{-\mu}}{x!}$, *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	*Binomial,* $p(x) = \binom{5}{x} (0.2)^x (0.8)^{5-x}$	*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	*Binomial,* $p(x) = \binom{100}{x}(0.01)^x(0.99)^{100-x}$	*Poisson,* $p(x) = \dfrac{1^x e^{-1}}{x!}$
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.

<u>Example 5.3:</u>

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) Fever than three accidents will occur within a period six weeks?

<u>Solution:</u>

(a) The random variable, x is the number of telephone calls received with an average of μ = 4 per minute. Hence x will have the Poisson distribution,

$$p(x) = \frac{4^4 e^{-4}}{x!},\ x = 0, 1, 2, \ldots .$$

(i) $P(x = 0) = p(0) = \dfrac{4^0 e^{-4}}{0!}$ = e^{-4} = 0.018316

(ii) $P(x \geq 2) = 1 - P(x \leq 1)$

$$= 1-\{P(x=0)+P(x=1)\}$$

$$= 1-\left(\frac{4^0\ e^{-4}}{0!} + \frac{4^1\ e^{-4}}{1!}\right)$$

$= 1 – (e^{-4} + 4e^{-4}) = 1 – 5e^{-4} \approx 0.908422$

(iii) $\mu = 4\times 2 = 8$ Calls received in two minutes.

$$P(x\geq 3) = 1-P(x\leq 2)$$

$$= 1-\{P(x=0)+P(x=1)+P(x=2)\}$$

$$= 1-\left(\frac{8^0 e^{-8}}{0!} + \frac{8^1 e^{-8}}{1!} + \frac{8^2 e^{-8}}{2!}\right)$$

$= 1 – (e^{-8} + 8e^{-8} + 32e^{-8}) = 1 – 41e^{-8} \approx 0.986246$

(b) Given that y approximately have the Poisson distribution,

$$p(y) = \frac{1.5^y e^{-1.5}}{y!},\ y = 0, 1, 2, \ldots, \text{ then}$$

(i) $$P(y>3) = 1-P(y\leq 2) = 1-\sum_{y=0}^{2}\frac{1.5^y\ e^{-1.5}}{y!}$$

$$= 1-\left(\frac{1.5^0 e^{-1.5}}{0!} + \frac{1.5^1\ e^{-1.5}}{1!} + \frac{1.5^2\ e^{-1.5}}{2!}\right)$$

$= 1 – (e^{-1.5} + 1.5\ e^{-1.5} + 1.125\ e^{-1.5})$

$= 1 – 3.25\ e^{-1.5} \approx 0.191153$

(ii) $\mu = 1.5\times {}^3\!/_2 = 2.25$ accidents in six weeks,

$$P(y<3) = \sum_{y=0}^{2}\ \frac{2.25^{-y}\ e^{-2.25}}{y!}$$

$$= \frac{2.25^0\ e^{-2.25}}{0!} + \frac{2.25^1\ e^{-2.25}}{1!} + \frac{2.25^2\ e^{-2.25}}{2!}$$

$= e^{-2.25} + 2.25\ e^{-2.25} + 2.53125\ e^{-2.25}$

$= 5.78125\ e^{-2.25} \approx 0.609339$

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.*

<u>*Solution:*</u>

(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,

$$P(x \geq 1) = 1 - P(x = 0)$$

$$= 1 - \frac{3^0 e^{-3}}{0!} = 1 - e^{-3} = 0.950213$$

(d) (i) By the Binomial distribution, n = 300 and p = 0.015. The required probability,

$$P(x \leq 1) = P(x = 0) + P(x = 1)$$

$$= \binom{300}{0}(0.015)^0(0.985)^{300} + \binom{300}{1}(0.015)^1(0.985)^{299}$$

$$= (0.985)^{300} + 300\,(0.015)\,(0.985)^{299}$$

$$= 0.010737 + 0.049051$$

$$\approx 0.059788$$

(ii) Using the Poisson Approximation, $\mu = np$ =300(0.015) = 4.5

$$P(x \leq 1) = P(x = 0) + P(x = 1)$$

$$= \frac{4.5^0 e^{-4.5}}{0!} + \frac{4.5\,e^{-4.5}}{1!}$$

$$= e^{-4.5} + 4.5\,e^{-4.5}$$

$$= 5.5\,e^{-4.5} \approx 0.061099$$

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 $x \sim B\left(n = 10, p = 0.70\right)$.

(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 drive 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 a 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

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, & -\infty < x < \infty \\ 0, & elsewhere \end{cases}$$

where $\sigma > 0,\ -\infty < \mu < \infty$ *and the mean and variance of the measurements,* x *are*

$E(x) = \mu$ and $Var(x) = \sigma^2$

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.

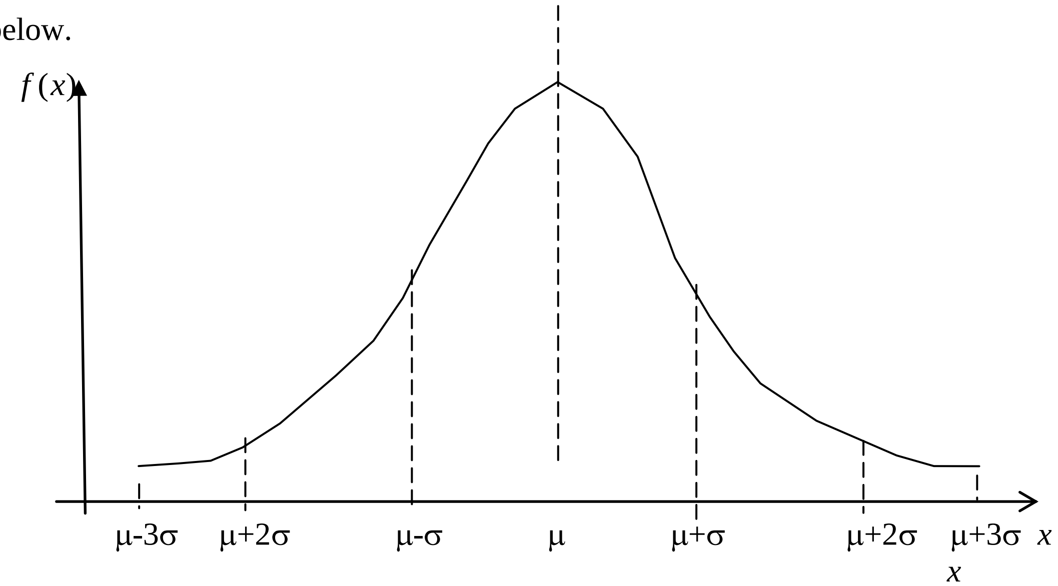


Figure 5.4: The Normal Curve

2-5.2 Properties of Normal Curve:

The normal curve has the following desirable properties, accounting for the wide-spread 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:

  $P(\mu-\sigma \le x \le \mu+\sigma) \approx 0.6826$, $P(\mu-2\sigma \le x \le \mu+2\sigma) \approx 0.9544$, *and*

  $P(\mu-3\sigma \le x \le \mu+3\sigma) \approx 0.9980$

2-5.3 Computation of Probability of Normal Random Variable:

To compute the probability that x lies within the interval [*a*, *b*], $P(a \le x \le b)$ the normal random variable, x is standardized using the transformation, $Z = \dfrac{x-\mu}{\sigma}$, 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,

$$P(a \le x \le b) = \int_a^b f(x)\, dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx$$

$$= \int_{a_a}^{z_b} f(z)\, dz = \int_{z_a}^{z_b} \frac{1}{\sqrt{2\mu}}\, e^{-\frac{1}{2}Z^2} dz$$

$$= P(Z_a \le Z \le Z_b) = P(Z \le Z_b) - P(Z \le Z_a)$$

$$= \int_{-\infty}^{z_b} f(z)\,dz - \int_{-\infty}^{z_a} f(z)\,dz = \Phi(z_b) - \Phi(z_a) = \Phi\left(\frac{b-\mu}{\mu}\right) - \Phi\left(\frac{a-\mu}{\mu}\right)$$

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) $\Phi(-k) = 1 - \Phi(-k) = 1 - P(z \le k)$

(ii) $P(z > k) = 1 - P(z \le k) = 1 - \Phi(k)$

(iii) $P(-k \le z \le k) = \Phi(k) - \Phi(-k) = \Phi(k) - [1 - \Phi(k)] = 2\Phi(k) - 1$

Example 5.4:

5.4(a) Find the following probabilities using the Normal table.

(i) $P(z \le -1.95)$ (ii) $P(-1.18 \le z \le 0.48)$

(iii) $P(0 \le z \le 2.58)$ (iv) $P(z > 2.63)$

(v) $P(-2.35 \le z \le 2.35)$

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) $P(z \le -1.95) = \Phi(-1.95) =$ 0.0256

(ii) $P(-1.18 \le z \le 0.48)$ $= P(z \le 0.48) - P(z \le -1.18)$

$= \Phi(0.48) - \Phi(-1.18)$ = 0.6844 – 0.1190 = 0.4654

(iii) $P(0 \le z \le 2.58)$)

$= P(z \le 2.58) - P(z \le 0.00)$

$= \Phi(2.58) - \Phi(0.00)$ = 0.9951 – 0.5000 = 0.4951

(iv) $P(z > 2.63) = 1 - P(z \le 2.63)$

$= 1 - \Phi(2.63)$ = 1 – 0.9957 = 0.0043

(v) $P(-2.35 \le z \le 2.35)$

$= P(z \le 2.35) - P(z \le -2.35)$

$= \Phi(2.35) - \Phi(-2.35)$ 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 $\sigma = 2$

$$P(5 < y < 10)$$
$$= P(y < 10) - P(y < 5)$$
$$= \Phi\left(\frac{10-6}{2}\right) - \Phi\left(\frac{5-6}{2}\right)$$
$$= \Phi(2) - \Phi(-0.5)$$
$$= 0.9772 - 0.3085 = 0.6687 \text{ or } 66.87\%$$

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 2.0*mg* and a standard deviation of 0.25*mg*. What is the probability that a cigarette will have nicotine content?

(i) Of 1.65*mg* or less

(ii) Between 1.50 and 2.25*mg*

(iii) Of 2.18*mg* or more?

<u>*Solution:*</u>

(c) The amount spent on maintenance and repairs, $x \sim N(400, 20^2)$, where $\mu = \$400$ and $\sigma = \$20$,

(i) $P(x > 450) = 1 - P(x \le 450)$

$$= 1 - \Phi\left(\frac{450-400}{20}\right)$$
$$= 1 - 0.9938 = 0.0062$$

(ii) Let the budgeted amount be *k* dollars.

$$P(x > k) = 0.1$$
$$= 1 - P(x \le k) = 0.1$$

$$P(x \le k) = 0.9$$

$$\Phi\left(\frac{k-400}{20}\right) = 0.9$$

$$\frac{k-400}{20} = \Phi^{-1}(0.9) = 1.28 = \phi^{-1}(0.9)$$

$$k = 400 + 20(1.28) = 425.6 \quad (\text{or } k = \$425.60)$$

(d) Given that the nicotine content of cigarettes, $y \sim N(2.0,\, 0.25^2)$,

(i) $P(y < 1.65) = \Phi\left(\frac{1.65 - 2.0}{0.25}\right)$

$= \Phi(-1.40) = 00808$

(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) $P(y > 2.18) = 1 - P(y \le 2.18)$

$= 1 - \Phi\left(\frac{2.18 - 2.0}{0.25}\right)$

$= 1 - \Phi(0.72)$

= 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 $Z = \dfrac{x \pm 0.5 - np}{\sqrt{np(1-p)}}$

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) $P(x \le 95)$

(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, $x \sim N\left(\mu, \sigma^2\right)$, where

$\mu = 200(0.4) = 80$ and $\sigma = \sqrt{200(04)(06)} = 6.9282$

(i) $P(x = 90) = P(89.5 \le x \le 90.5)$

$$= \Phi\left(\frac{90.5 - 80}{6.9382}\right) - \Phi\left(\frac{89.5 - 80}{6.9382}\right)$$

$$= \Phi(2.81) - \Phi(-1.37) = 0.9357 - 0.9147 = 0.0210$$

(ii) $P(x \le 95) = (x \le 95.5)$

$= \Phi\left(\frac{95.5-80}{6.9282}\right)$

$= \Phi(2.24) = 0.9875$

(iii) $P(x > 65) = 1 - P(x \le 65.5)$

$= 1 - \Phi\left(\frac{65.5-80}{6.9282}\right)$

$= 1 - \Phi(-2.09) = 1 - 0.0183 = 0.9817$

(iv) $P(x < 60) = P(x \le 59.5)$

$= \Phi\left(\frac{59.5-80}{6.9282}\right) = \Phi(-2.96) = 0.0015$

(v) $P(70 < x < 100) = (70.5 \le x \le 99.5)$

$= \Phi\left(\frac{99.5-80}{6.9282}\right) - \Phi\left(\frac{70.5-80}{6.9282}\right)$

$= \Phi(2.81) - \Phi(-1.37)$

$= 0.9975 - 0.0853 = 0.9122$

(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) The expected number of components that did not meet customer specifications, $E(x) = \mu = np = 50$

(ii) $P(x \ge 52) = 1 - P(x < 51.5)$

$= 1 - \Phi\left(\frac{51.5-50}{6.71}\right)$

$= 1 - \Phi(0.22)$

$= 1 - 0.5871 = 0.4129$

(iii) $P(36 \le x \le 58) = P(35.5 \le x \le 58.5)$

$$= \Phi\left(\frac{58.5-50}{6.71}\right) - \Phi\left(\frac{35.5-50}{6.71}\right)$$

$$= \Phi(1.27) - \Phi(-2.16)$$

= 0.8980 – 0.0197 = 0.8783

2-5.5 Linear Combinations of Normal Random Variables:

- *Theorem5.1:*

Let $x_1, x_2, x_3, \ldots, x_n$ be independent normal random variables and $a_2, a_2, a_3, \ldots, a_n$ be constants. Then the linear combination,

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \ldots + a_nx_n = \sum_{i=1}^{n} a_i x_i$$

is also a normal random variable with mean and variance given as follows.

$$E(y) = E\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} a_i E(x_i),\ and\quad Var(y) = Var\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} a_i^2\, Var(x_i)$$

However, if the random variables, $x_1, x_2, x_3, \ldots, x_n$ are not independent then

$$Var(y) = \sum_{i=1}^{n} a_i^2\, Var(x_i) + 2a_i a_j \sum_i \sum_j Cov(x_i, x_j)$$

where the double sum is over all pairs (i, j) with $i < j$

- *The Sampling Distribution of the Sample Mean,* $(\bar{x})$:

Let $x_1, x_2, x_3, \ldots, x_n$ 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:

$$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}(n\mu) = \mu$$

$$Var(\bar{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

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, $\mu_x = \mu$ and variance, $\sigma_{\bar{x}}^2 = \sigma^2/n$ 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:

(i) $E(2x+3y)$ and $Var(2x+3y)$

(ii) $P(2x+3y < 40)$

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

(i) Will exceed 1,047 *grams*

(ii) 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

(i) $\mu_T = E(T) = E(2x + 3y)$

$$= 2E(x) + 3E(y)$$
$$= 2(1) + 3(10) = 32$$

$$\sigma_T^2 = Var(T) = Var(2x + 3y)$$
$$= 4Var(x) + 9Var(y)$$
$$= 4(4) + 9(9) = 97 = (9.85)^2$$

(ii) From (i) $T \sim N(32, 97)$

$$P(T < 40) = \Phi\left(\frac{40 - 32}{9.85}\right)$$
$$= \Phi(0.81) = 0.7910$$

(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:

$$\begin{aligned}\mu_T &= E(T) = E(T_1 + T_2)\\ &= E(T_1) + E(T_2)\\ &= E(20x) + E(100)\\ &= 20E(x) + 100\\ &= 20(50) + 100 = 1{,}100g\end{aligned}$$

$$\begin{aligned}\sigma_T^2 &= Var(T) = Var(T_1 + T_2)\\ &= Var(20T_1) + Var(T_2)\\ &= 20^2 Var(T_1) + Var(T_2)\\ &= 20^2(16) + 9 = 6{,}409 = (80.06)^2\end{aligned}$$

Hence, $T \sim N$ (1,100 + 6,409)

(i) $P(T > 1{,}074) = 1 - P(T \le 1{,}074)$

$$\begin{aligned}&= 1 - \Phi\left(\frac{1{,}074 - 1{,}100}{80.06}\right)\\ &= 1 - \Phi(-0.32)\\ &= 1 - 0.3745\\ &= 0.6255\end{aligned}$$

(ii) $P(1{,}050 < T < 1{,}200)$

$$\begin{aligned}&= \Phi\left(\frac{1{,}200 - 1{,}100}{80.06}\right) - \Phi\left(\frac{1.050 - 1{,}100}{80.06}\right)\\ &= \Phi(1.25) - \Phi(-0.62)\\ &= 0.8944 - 0.2676\\ &= 0.6268\end{aligned}$$

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 second?

(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 $k\,mmHg$, 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 detective, 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 is known that the weight of a carton of the item is normally distributed with a mean of 40 kg and standard and 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 mph^2. 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 pesewa,

(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.

JOINT PROBABILITY DISTRIBUTIONS

Introduction

The study of probability distributions in MATH 166 centred on a single random variable of either discrete or the continuous type. Such random variables are called *univariate*. Often, problems do arise in which two or more random variables are to be studied simultaneously. For example, we might wish to study the yield of a chemical reaction in conjunction with the temperature at which the reaction is run. Typical questions to ask are: Is the yield independent of the temperature?" or "What is the average yield if the temperature is 40^{o}C?" To answer questions of this type, we need to study the joint behaviour of these random variables. In this unit we present the basic theoretical concepts underlying these random variables for both discrete and continuous. For simplicity, we begin by considering random experiments in which only two random variables are studied. In the later sections, we generalize the presentation to the joint probability distribution of more than two random variables. The joint probability distributions indicate the joint behaviour (or interaction) of the random variables. These concepts form the basis for the study of regression and correlation analysis (MATH 369), a course to be studied in the third year of the programme.

Learning Objectives

After studying this unit, you will be able to:

- Cite various situations which lead to the modelling of joint behaviour of two or more random variables.
- Explain the properties of bivariate probability distribution (a special type of joint probability distribution) and obtain the marginal and conditional distributions as well.
- Compute conditional mean and variance, and the correlation coefficient between two random variables.
- Determine the probability distribution of function of random variable(s).
- Determine moment generating function of bivariate distributions.
- Find the probability distributions of order statistics drawn from a probability distribution.

SESSION 1-1: BIVARIATE PROBABILITY DISTRIBUTIONS

1-1.1 Introduction: Joint Probability Distributions

In most scientific or experimental investigations several types of observations are made on the experimental units. This provides crucial information since the manner in which variables act together may indicate the most factors in explaining the outcome of the experiment. The interactions revealed in such studies are often of great value than the effects of the individual variables alone in investigations which are becoming increasingly rare. There are various situations which lead to the modelling of joint behaviour of two or more random variables. Some of these situations are as given below:

- An actuary may be interested in studying the joint behaviour of age of an insured vehicle in an accident and the length of time the vehicle has been insured as at the time of accident.
- A physicist may wish to study the effects of transmissions in a fibre optic cable when transmission rates and the composition of the cable are varied.
- A physician may be interested in studying the joint behaviour of blood pressure, weight, height, and body temperature of patients.
- In sample survey study, respondents are usually asked several questions, creating a random variable for each question.
- A sample piece of a wire may have these properties measured; tensile, torsion and diameter, where their joint behaviour may be of much interest.
- The Academic Board of KNUST studying student performance may relate it to the entry grades, attendance of lectures, studying time, end of semester examination score, etc.
- Tossing a coin n times and observing the number of heads in the first k tosses and number of heads in the last $(n - k)$ tosses results two random variables which we may wish to investigate their joint behaviour.

Each of the above examples suggests a sample on which more than one random variable is defined. The joint behaviour of these random variables shows how they interact or are correlated with each other. The probability distribution involving two or more random variables is known as *joint (or multivariate) probability distribution*. A joint probability distribution involving two random variables is called a *bivariate probability distribution*. We begin the study with some properties of bivariate distributions, and generalize the presentation for more than two random variables in the later sections.

1-1.2 Basic Definitions and Properties

Let X and Y be two random variables defined over the sample space, S. Then the set of values $(x, y) \in S$ forms an event with the probability distribution, $f(x, y)$ that defines their simultaneous or joint behaviour, called *bivariate distribution*. The bivariate distribution describes the joint probability, $P\left(X = x, Y = y\right)$, corresponding to some particular values (x, y) for the two random variables *X* and *Y*.

We now present the following definitions and properties of joint probability distribution involving two random variables.

- ***Probability Mass Function***: Let *x* and *y* be discrete random variables defined on the sample space, *S* with joint probability distribution, $f(x, y)$. The joint probability distribution, $f(x, y)$ is said to be *probability mass function*, if the following probabilities are satisfied:
  - (i) $f(x, y) \geq 0, \textit{ for all } (x, y)$
  - (ii) $\sum_{\forall x}\sum_{\forall y} f(x, y) = 1$
  - (iii) $P(X = x, Y = y) = f(x, y)$

- ***Probability Density Function***: Let *X* and *Y* be continuous random variables defined on the sample space, *S* with joint probability distribution, $f(x, y)$. The joint distribution, $f(x, y)$ is said to be *probability density function* if the following properties hold;
  - (i) $f(x, y) \geq 0, \textit{ for all } (x, y)$
  - (ii) $\int_{Rx}\int_{Rx} f(x, y)\, dxdy = 1$
  - (iii) $P(X \leq a, Y \leq b) = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x, y)\, dxdy$

- ***Cumulative Distribution Functions****:* Suppose *X* and *Y* are two random variables defined on the sample space, *S*. The joint cumulative distribution functions of *x* and *y* is defined as:
  - (i) $F(x,\ y) = \sum_{u}^{x}\sum_{v}^{y} f(u,\ v)$, where *x* and *y* are discrete.
  - (ii) $F(x,\ y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(x,\ y)$, where *x* and *y* are continuous

- ***Marginal Probability Distributions****:* The probability distributions of the individual random variables *x* and *y* are referred to as the *marginal probability distributions*. The marginal distributions of *x* and *y* with joint probability distribution function $f(x, y)$ are defined by:
  - (i) $f(x) = \sum_{\forall y} f(x, y),$ $\qquad f(y) = \sum_{\forall x} f(x, y)$

    where *x* and *y* are discrete random variables.
  - (ii) $f(x) = \int_{Ry} f(x, y)dy$, $\qquad f(y) = \int_{Rx} f(x, y)dx$

    where *x* and *y* are continuous random variables.

- ***Independent Random Variables***: The two random variables *x* and *y* are said to be *independent* if $f(x, y) = f(x).f(y)$.

- ***Conditional Distributions****:* Let x and y be random variables with joint probability distribution function, $f(x, y)$. We define the conditional probability distribution of:

  (i) y given $X = x$ as: $f(y \mid X = x) = \dfrac{f(x,\ y)}{f(X = x)}$, where $f(x) > 0$.

  (ii) x given $Y = y$ as: $f(x \mid Y = y) = \dfrac{f(x,\ y)}{f(Y = y)}$, where $f(y) > 0$.

Illustrative Example 1.1

For each of the following joint probability distribution functions, we find the marginal and conditional distributions, and related information.

(a) $$f(x, y) = \begin{cases} x^2 + \frac{8}{3}xy, & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0, & elsewhere \end{cases}$$

(b) $$f(x, y) = \begin{cases} k\left(\frac{1}{x} + \frac{1}{y}\right), & for\ x = 1, 2, 3\ and\ y = 2, 3. \\ 0, & elsewhere \end{cases}$$

(c) $$f(x, y) = \begin{cases} ke^{-2x-y}, & x \ge 0,\ y \ge 0 \\ 0, & elsewhere \end{cases}$$

(d) A joint distribution involving *x* and *y* is as shown below.

x	0	1	1	0	2	2
y	1	0	1	2	0	2
$f(x, y)$	$1/8$	$1/16$	$1/4$	$1/16$	$1/16$	$7/16$

where $f(x, y) = 0$ *for* $(0, 0), (1, 2)$ *and* $(2, 1)$.

<u>*Solution*</u>

(a) $$f(x, y) = \begin{cases} x^2 + \frac{8}{3}xy, & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0, & elsewhere \end{cases}$$

(i) The marginal distributions of x and y are respectively

$$f(x) = \int_0^1 f(x, y)dy$$

$$= \int_0^1 \left(x^2 + \frac{8}{3}xy\right)dy$$

$$= \left[x^2 y + \frac{4}{3}xy^2\right]_0^1 = x^2 + \frac{4}{3}x$$

i.e. $f(x) = \begin{cases} x(2+x), & 0 < x < 1 \\ 0 & , \textit{elsewhere} \end{cases}$

$$f(y) = \int_0^1 f(x,y)dx$$

$$= \int_0^1 \left(x^2 + \frac{8}{3}xy \right) dx$$

$$= \left[\frac{1}{3}x^3 + \frac{4}{3}x^2 y \right]_0^1 = \frac{1}{3} + \frac{4}{3}y$$

i.e. $f(y) = \begin{cases} \frac{1}{3}(1+4y), & 0 < y < 1 \\ 0 & , \textit{elsewhere} \end{cases}$

(ii) To show that x and y are independent:

$$f(x).f(y) = \left(x^2 + \frac{4}{3}x \right)\left(\frac{1}{3}(1+4y) \right)$$

$$= \frac{1}{3}\left(x^2 + 4x^2 y + \frac{4}{3}x + \frac{16}{3}xy \right)$$

$$\neq f(x,y) = x^2 + \frac{8}{3}xy$$

Hence x and y are not independent.

(iii) The conditional distribution of x given $Y = 1$,

$$f(x \mid y) = \frac{f(x,y)}{f(y)}, \; f(y) > 0$$

$$= \frac{x^2 + {}^{8}\!/_{3}\, xy}{{}^{1}\!/_{3}\,(1+4y)}$$

$$f(x \mid Y = 1) = \frac{x^2 + {}^{8}\!/_{3}\, x}{{}^{1}\!/_{3}\,(5)}$$

$$= \frac{x}{5}(3x+8), \; 0 \le x \le 1$$

(iv) The probability,

$$P\left(y < \tfrac{1}{2} \mid X = 1\right) = \int_0^{1/2} f\left(y \mid X = 1\right) dy$$

$$= \int_0^{1/2} \frac{f(x,y)}{f(X=1)} dy$$

$$= \int_0^{1/2} \left(\frac{(1 + \tfrac{8}{3} y)}{\tfrac{7}{3}} \right) dy$$

$$= \int_0^{1/2} \frac{1}{7}\left(3 + 8y\right) dy$$

$$= \frac{1}{7}\left[3y + 4y^2\right]_0^{1/2}$$

$$= \frac{1}{7}\left[\tfrac{3}{2} + 4\left(\tfrac{1}{2}\right)^2\right] = \tfrac{5}{14}$$

(b) $$f(x,y) = \begin{cases} k\left(\dfrac{1}{x} + \dfrac{1}{y}\right), & x = 1, 2, 3\,;\ \ y = 2, 3 \\ 0 & , \ elsewhere \end{cases}$$

(i) We first find k by assuming $f(x,\ y)$ is probability mass function.

$$\sum_{x=1}^{3}\sum_{y=2}^{2} f(x,y) = 1$$

$$\sum_{x=1}^{3}\sum_{y=2}^{2} k\left(\frac{1}{x} + \frac{1}{y}\right) = 1$$

$$k\sum_{x=1}^{3}\left[\left(\frac{1}{x} + \frac{1}{2}\right) + \left(\frac{1}{x} + \frac{1}{3}\right)\right] = k\sum_{x=1}^{3}\left(\frac{2}{x} + \frac{5}{6}\right) = 1$$

$$k\left\{\left(\frac{2}{1} + \frac{5}{6}\right) + \left(\frac{2}{2} + \frac{5}{6}\right) + \left(\frac{2}{3} + \frac{5}{6}\right)\right\}$$

$$= \frac{37k}{6} = 1 \Leftrightarrow k = \frac{6}{37}$$

i.e. $$f(x,y) = \begin{cases} \dfrac{6}{37}\left(\dfrac{1}{x} + \dfrac{1}{y}\right), & x = 1, 2, 3;\ \ y = 2, 3 \\ 0 & , \ elsewhere \end{cases}$$

(ii) The marginal distributions of x and y are

$$f(x) = \sum_{y=2}^{3} f(x, y)$$

$$= \sum_{y=2}^{3} \frac{6}{37}\left(\frac{1}{x} + \frac{1}{y}\right)$$

$$= \frac{6}{37}\left[\left(\frac{1}{x} + \frac{1}{2}\right) + \left(\frac{1}{x} + \frac{1}{3}\right)\right]$$

$$= \frac{6}{37}\left(\frac{2}{x} + \frac{5}{6}\right), \; x = 1, 2, 3$$

$$f(y) = \sum_{x=1}^{3} f(x, y)$$

$$= \sum_{x=1}^{3} \frac{6}{37}\left(\frac{1}{x} + \frac{1}{y}\right)$$

$$= \frac{6}{37}\left[\left(\frac{1}{1} + \frac{1}{y}\right) + \left(\frac{1}{2} + \frac{1}{y}\right) + \left(\frac{1}{3} + \frac{1}{y}\right)\right]$$

$$= \frac{6}{37}\left(\frac{11}{6} + \frac{3}{y}\right), \; y = 2, 3$$

(iii) $P(x \geq y) = f(2, 2) + f(3, 2) + f(3, 3)$

$$= \frac{6}{37}\left[\left(\frac{1}{2} + \frac{1}{2}\right) + \left(\frac{1}{3} + \frac{1}{2}\right) + \left(\frac{1}{3} + \frac{1}{3}\right)\right]$$

$$= \frac{6}{37}\left(1 + \frac{5}{6} + \frac{2}{3}\right) = \frac{6}{37}\left(\frac{15}{6}\right) = \frac{15}{37}$$

(c) $$f(x, y) = \begin{cases} ke^{-2x-y} & , \; x \geq 0, \; y \geq 0 \\ 0 & , \; elsewhere \end{cases}$$

(i) To determine k we assume that $f(x, y)$

$$\int_0^\infty \int_0^\infty ke^{-(y+2x)} dxdy = 1$$

$$k\int_0^\infty \left[-\tfrac{1}{2} e^{-(y+2x)}\right]_0^\infty dy = 1$$

$$k\int_0^\infty \tfrac{1}{2} e^{-y} dy = 1$$

$$k\left(-\tfrac{1}{2} e^{-y}\right)\Big|_0^\infty = \frac{1}{2}k = 1$$

$$k = 2$$

i.e $f(x, y) = 2e^{-(2x+y)}, \; x \geq 0, \; y \geq 0$

(ii) The marginal distributions are

$$f(x) = \int_0^{\infty} 2e^{-(2x+y)}dy$$
$$= \left[-2e^{-(2x+y)}\right]_0^{\infty}$$
$$= 2e^{-2x},\ x \geq 0$$

$$f(y) = \int_0^{\infty} 2e^{-(2x+y)}dx$$
$$= \left[-e^{-(2x+y)}\right]_0^{\infty}$$
$$= e^{-y},\ y \geq 0,$$

from which we note that x and y are independent. Thus,

$$f(x).f(y) = 2e^{-2x}.\ e^{-y}$$
$$= 2e^{-(2x+y)}$$
$$= f(x,y)$$

(d) The given discrete distribution can also be put in a tabular form as;

x	y = 0	y = 1	y = 2	$f(x)$
0	0	$1/8$	$1/16$	$3/16$
1	$1/16$	$1/4$	0	$5/16$
2	$1/16$	0	$7/16$	$1/2$
$f(y)$	$1/8$	$3/8$	$1/2$	1

where $f(x,y) = 0,$ *for* $(0, 0),\ (1, 2)$ *and* $(2, 1)$

(i) The marginal distributions of x and y are given by the row and totals $f(x)$ and $f(y)$ as shown by the tables below.

x	0	1	2
$f(x)$	$1/8$	$5/16$	$1/2$

y	0	1	2
$f(y)$	$1/8$	$3/8$	$1/2$

(ii) The expectation of x and y is as follows:

$$E(x) = \sum_{x=0}^{2}\sum_{y=0}^{2} x\, f(x,y) \ or \ \sum_{x=0}^{2} x\, f(x)$$

$$= \sum_{x=0}^{2} x\big[[\, f(x,0) + f(x,1) + f(x,2)\big]$$

$$= 0.f(0,0) + 0.f(1,1) + 0.f(0,2) + 1.f(1,0) + 1.f(1,1)$$
$$+1.f(1,2) + 2.f(2,0) + 2.f(2,1) + 2.f(2,2)$$

$$= \tfrac{1}{16} + \tfrac{1}{4} + 2\left(\tfrac{1}{16}\right) + 2\left(\tfrac{7}{17}\right) = \frac{21}{16} = 1.3125$$

$$E(y) = \sum_{x=0}^{2}\sum_{y=0}^{2} y\, f(x,y) \ \ or \ \ \sum_{y=0}^{2} y\, f(y)$$

$$= \sum_{y=0}^{2} y[\, f(0,y) + f(1,y) + f(2,y)]$$

$$= 0.f(0,0) + 0.f(1,0) + 0.f(2,0) + 1.f(0,1) + 1.f(1,1)$$
$$+1.f(2,1) + 2f(0,2) + 2.f(1,2) + 2.f(2,2)$$

$$= \frac{1}{8} + \frac{1}{4} + 2\left(\tfrac{1}{16}\right) + 2\left(\tfrac{7}{16}\right)$$

$$= \frac{11}{8} = 1.375$$

(iii) To find the probability

$$P(x \le 2, y = 1) = f(0,1) + f(1,1) + f(2,1)$$

$$= \tfrac{1}{8} + \tfrac{1}{4} + 0 = \tfrac{3}{8}$$

$$= 0.375$$

1-1.3 Covariance and Correlation Coefficient

If the random variables x and y have a joint probability distribution, there may exist a linear relationship or not between them. The *covariance* of x and y indicates the association between x and y while the strength of the association is determined by *correlation coefficient*.

- The covariance of x and y denoted by, $Cov(x, y)$ is defined as

$$Cov(x, y) = E\big[(x - \mu_x)(y - \mu_y)\big] = E(xy) - \mu_x\mu_y \quad . \ . \ . \ . \qquad (1.1)$$

where $\mu_x = E(x)$ *and* $\mu_y = E(y)$

- The correlation coefficient between x and y, denoted by $Corr(x, y)$ *or* ρ_{xy}, is defined as $\rho_{xy} = \dfrac{Cov(x,\ y)}{\sigma_x\sigma_y} = \dfrac{E(xy) - \mu_x\mu_y}{\sigma_x\sigma_y}$ . . . . (1.2)

from which we have

$$E(xy) = \rho_{xy}\sigma_x\sigma_y + \mu_x\mu_y \quad . \ . \ . \ . \qquad (1.3)$$

where $\sigma_x^2 = Var(x)$ *and* $\sigma_y^2 = Var(y)$ *and* $-1 \le \rho_{xy} \le 1$

- If x and y are independent, then $E(xy) = E(x)\,E(y) = \mu_x\mu_y$. Hence,

$$Cov(x, y) = E(xy) - \mu_x\mu_y$$
$$= \mu_x\mu_y - \mu_x\mu_y = 0$$

1-1.4 Conditional Mean and Variance

Let the random variables, x and y be jointly distributed with probability distribution function, $f(x, y)$.

- The conditional mean y given $X = x$ is defined by:

(i) $\mu_{y|x} = E(y \mid X = x) = \sum_{\forall y} y\, f(y \mid X = x)$

$$= \sum_{\forall y} y\, \frac{f(x, y)}{f(x)},\ f(x) > 0, \text{ if } x \text{ and } y \text{ are discrete} \quad . \ . \ . \ . \quad (1.4)$$

(ii) $\mu_{y|x} = E(y \mid X = x) = \int_{Ry} y\, f(y \mid X = x)dy$

$$= \int_{Ry} y\, \frac{f(x, y)}{f(x)} dy,\ f(x) > 0, \text{ if } x \text{ and } y \text{ are continuous.} \ . \ . \ . \ . \quad (1.5)$$

- The conditional variance of y given $X = x$ is given by:

(i) $\sigma_{y|x}^2 = Var(y \mid X = x) = E\left[(y - \mu_{y|x})^2\right]$

$$= \sum_{\forall y} (y - \mu_{y|x})^2\ f(y \mid y), \quad . \ . \ . \ . \quad (1.6)$$

where $\mu_{y|x}$ is the conditional mean of y given x and x and y are discrete.

(ii) $\sigma_{y|x}^2 = Var(y \mid X = x) = E\left[(y - \mu_{y|x})^2\right]$

$$= \int_{Ry} (y - \mu_{x|y})^2\ f(\mid X = x)dy \quad . \ . \ . \ . \quad (1.7)$$

where $\mu_{y|x}$ is the conditional mean of y given x, and x and y are continuous.

The conditional mean of y given x is a linear function of the form:

$$\Phi(x) = a + bx \quad . \ . \ . \ . \quad (1.8)$$

i.e. $\Phi(x) = E(y \mid x) = \int_{Ry} y\ f(y \mid x)dy = \int_{Ry} y \frac{f(x, y)}{f(x)} dy$

$$\Phi(x).f(x) = \int_{Ry} y\ f(x, y)dy \quad . \ . \ . \ . \quad (1.9)$$

Integrating both sides with respect to x

$$\int_{Ry} (a + bx)\ f(x)dx = \int_{Rx}\left[\int_{Ry} y\, f(x, y)dy\right]dx$$

$$a\int_{Rx} f(x)\,dx + b\int_{Rx} x\, f(x)dx = \int_{Ry} y\left[\int_{Rx} f(x, y)dx\right]dy$$

$$a + bE(x) = E(y)$$

$$a + b\mu_x = \mu_y \qquad \text{. . . . (1.10)}$$

Multiplying both sides of equation (1.9) by x and integrating with respect to x.

$$\int_{Rx} (a + bx)\, x.f(x)dx = \int_{Rx}\int_{Ry} xy\, f(x,y)dydx$$

$$a \int_{\forall x} x.f(x)dx + b\int_{\forall x} x^2 f(x)dx = E(xy)$$

$$aE(x) + bE(x^2) = E(xy)$$

$$a\mu_x + b(\sigma_x^2 + \mu_x^2) = E(xy)$$

$$a\mu_x + b(\sigma_x^2 + \mu_x^2) = \rho_{xy}\sigma_x\sigma_y + \mu_x\mu_y, \; [\mathit{from}\ (1.3)] \qquad \text{. . . . (1.11)}$$

Solving equations (1.10) and (1.11) simultaneously we obtain,

$$a = \mu_y - \rho_{xy}\frac{\sigma_y}{\sigma_x}\mu_x \quad \text{and} \quad b = \rho_{xy}\frac{\sigma_y}{\sigma_x}$$

Hence the conditional mean of y given x, $E(y \mid x) = a + bx$, becomes,

$$\mu_{y|x} = \mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x - \mu_x) \qquad \text{. . . . (1.12)}$$

Similarly, the conditional mean of x given y, $E(x \mid y)$, which is a linear function of y given by

$$E(x \mid y) = \Phi(y) = c + dy = \mu_{x|y}$$

$$\mu_{x|y} = \mu_x + \rho_{xy}\frac{\sigma_x}{\sigma_y}(y - \mu_y) \qquad \text{. . . . (1.13)}$$

- ***Theorem 1.1: The Limits of*** ρ_{xy}

The correlation coefficient, ρ_{xy} ranges between – 1 and 1, inclusive. That is,

$-1 \le \rho_{xy} \le 1.$

Proof of Theorem

By equation 6.7, the conditional variance

$$Var(y|x) = E\left[(y-\mu_{y|x})^2\right] \geq 0$$

$$= E\left[\left(y-\mu_y-\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)^2\right] \geq 0$$

$$= E\left[\left((y-\mu_y)^2 - 2\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\mu_x)(y-\mu_y) + \rho_{xy}^2\frac{\sigma_y^2}{\sigma_x^2}(x-\mu_x)^2\right)\right] \geq 0$$

$$= E(y-\mu_y)^2 - 2\rho_{xy}\frac{\sigma_y}{\sigma_x}E[(x-\mu_x)(y-\mu_y)] + \rho_{xy}^2\frac{\sigma_y^2}{\sigma_x^2}E(x-\mu_x)^2 \geq 0$$

$$= \sigma_y^2 - 2\rho_{xy}^2\sigma_y^2 + \rho_{xy}^2\sigma_y^2 \geq 0$$

from which we have obtain the desired result, $-1 \leq \rho_{xy} \leq 1$ . . . .

(1.15)

Illustrative Example 1.2

1.(a) Let X and Y be random variables whose joint probability distribution and marginal distributions are given in the table below:

Y	X = 1	X = 2	$f(y)$
1	0.15	0.40	0.40
2	0.35	0.50	0.60
$f(x)$	0.50	0.50	1.00

(i) Find the expectations of X, Y and the product, XY.

(ii) Find the Variances of X and Y

(iii) Find the covariance and the correlation coefficient of X and Y.

(iv) Find the variance of the sum of X and Y, $Var(X+Y)$.

(b) Let X and Y be random variables whose joint probability density function is given by

$$f(x,y) = \begin{cases} \dfrac{4(1-xy)}{3}, & for\ 0 \leq x \leq 1\ and\ 0 \leq y \leq 1 \\ 0 & , elsewhere \end{cases}$$

(i) Find $P(x+y \leq 1)$

(ii) Find expectations of X, Y, the sum, $X+Y$, and the product, XY

(iii) Find the variances of X and Y, and the covariance.

(iv) Compute the variance of sum, $X+Y$, and the correlation coefficient.

<u>Solution</u>

(a) (i) The expectations of X, Y, and XY are:

$$E(X) = 1(0.5) + 2(0.50) = 1.5$$

$$E(Y) = 1(0.4) + 2(0.60) = 1.6$$

$E(XY) = 1(0.15) + 2(0.25) + 2(0.35) + 4(0.25) = 2.35$

(ii) $Var(X) = E(X^2) - \left(E(X)\right)^2 = 1(0.5) + 4(0.50) - (1.5)^2 = 0.25$

$Var(Y) = E(Y^2) - \left(E(Y)\right)^2 = 1(0.4) + 4(0.60) - (1.6)^2 = 0.24$

(iii) $Cov(X,Y) = E(XY) - E(X)E(Y) = 2.35 - (1.5)(1.6) = -0.05$

$$\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-0.05}{\sqrt{(0.25)(0.24)}} = -0.20412$$

(iv) $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

$= 0.25 + 0.24 + 2(-0.05) = 0.39$

(b) (i) $P(x + y \le 1) = \int_0^1 \int_0^{1-y} f(x,y)dxdy$

$$= \int_0^1 \int_0^{1-y} \frac{4}{3}(1 - xy)dxdy$$

$$= \frac{4}{3}\int_0^1 \left(x - \frac{x^2 y}{2} \right)\Bigg|_o^{1-y} dy$$

$$= \frac{2}{3}\int_0^1 (2 - 3y + 2y^2 - y^3)dy$$

$$= \frac{2}{3}\left(2y - \frac{3y^2}{2} + \frac{2y^3}{3} - \frac{y^4}{4} \right)\Bigg|_0^1 = \frac{11}{18}$$

(ii) $E(X+Y) = \int_0^1 \int_0^1 (x+y) f(x,y)dxdy$

$$= \frac{4}{3}\int_0^1 \int_0^1 (x + y - x^2 y - xy^2)dxdy$$

$$= \frac{4}{3}\int_0^1 \left(\frac{x^2}{2} + xy - \frac{x^3 y}{3} - \frac{x^2 y^2}{2} \right)\Bigg|_0^1 dy$$

$$= \frac{4}{3}\int_0^1 \left(\frac{1}{2} + \frac{2y}{3} - \frac{y^2}{2} \right)dy$$

$$= \frac{4}{3}\left(\frac{y}{2} + \frac{y^2}{3} - \frac{y^3}{6} \right)\Bigg|_0^1 = \frac{16}{18} = \frac{8}{9}$$

$$E(XY) = \int_0^1 \int_0^1 xy\, f(x,y)dxdy$$

$$= \frac{4}{3}\int_0^1 \int_0^1 (xy - x^2y^2)dxdy$$

$$= \frac{4}{3}\int_0^1 \left(\frac{x^2y}{2} - \frac{x^3y^2}{3} \right)\Bigg|_0^1 dy$$

$$= \frac{4}{3}\int_0^1 \left(\frac{y}{2} - \frac{y^2}{3} \right) dy$$

$$= \frac{4}{3}\left(\frac{y^2}{4} - \frac{y^3}{9} \right)\Bigg|_0^1 = \frac{5}{27}$$

$$E(X) = \int_0^1 \int_0^1 x\, f(x,y)dxdy$$

$$= \frac{4}{3}\int_0^1 \int_0^1 (x - x^2y)dxdy = \frac{4}{9}$$

$$E(Y) = \int_0^1 \int_0^1 y\, f(x,y)dxdy$$

$$= \frac{4}{3}\int_0^1 \int_0^1 (y - xy^2)dxdy = \frac{4}{9}$$

(iii) $$E(X^2) = \int_0^1 \int_0^1 x^2 f(x,y)dxdy = \frac{4}{9}$$

$$E(Y^2) = \int_0^1 \int_0^1 y^2 f(x,y)dxdy = \frac{4}{9}$$

$$Var(X) = E(X^2) - \left(E(X)\right)^2 = \frac{13}{162}$$

$$Var(Y) = E(Y^2) - \left(E(Y)\right)^2 = \frac{13}{162}$$

$$Cov(x,y) = E(XY) - E(X)E(Y) = -\frac{1}{81}$$

(iv) $$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) = \frac{11}{81}$$

$$\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$= \frac{-1/81}{\sqrt{\left(13/162\right)\left(13/162\right)}} - 2/13 = -0.15385$$

Self Assessment Questions (Exercise 1.1)

1.(a) Compute the correlation coefficient for each of the following probability densities:

(i) $f(x,y)=\begin{cases}\frac{1}{3}(x+y), \ for \ 0\le x\le 1; \ 0\le y\le 2, \ and \ 0, elsewhere\end{cases}$

(ii) $f(x,y)=\begin{cases}\frac{1}{22}(x+2y), \ for \ (1,1),(1,3),(2,1),(2,3), \ and \ 0, \ elsewhere\end{cases}$

(b) Suppose x and y have the joint distribution,

$$f(x,\ y)=\begin{cases}k(x+y^2), & 0\le x\le 2, \quad 0\le y\le 1\\ 0 & , \ elsewhere\end{cases}$$

(i) Find k and the marginal distribution functions of x and y.

(ii) Compute the probabilities, $P\left(x>1,\ y<\frac{1}{2}\right)$ and $P\left(y>\frac{1}{2}\middle| x=1\right)$

(iii) Find the conditions, and hence $E\left(y\,|\,x\right)$ and $P\left(x+y\le 2\right)$

(c) Compute the correlation coefficient for the following probability mass function:

(x,y)	(1, 1)	(1, 2)	(1, 3)	(2, 1)	(2, 2)	(2, 3)
$f(x,y)$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{4}{15}$

2.(a) Let x and y have the joint probability distribution function,

$$f(x,y)=\begin{cases}k\left(16-4x-4y+xy\right), & for \ x=1,\ 2,\ 3; \ y=1,\ 2,\ 3\\ 0 & , \ elswhere\end{cases}$$

(i) What are the value of k and the marginal distributions of x and y?

(ii) Compute the probability, $P(x \le y)$.

(b) A construction company wins two road rehabilitation projects, *RH1* and *RH2* which are to be executed simultaneously. The scheduled completion times of the projects follow the distribution as shown below.

Project RH1		*Project RH2*	
Time, t_1 (in months)	*Probability, $P(t_1)$*	*Time, t_2 (in months)*	*Probability, $P(t_2)$*
24	0.30	18	0.30
30	0.60	24	0.50
36	0.10	30	0.20

Assuming that the completion times of the projects are independent, find the probability that:

(i) the two projects will be completed at the same time,

(ii) both projects will be completed in less than 30 months, and

(iii) project *RH1* takes longer time to complete than project *RH2.*

(iv) Find the expected completion time for each project and interpret your results.

(c) Once a fire is reported to a fire to an insurance company, the company makes an initial estimate, X, of the amount it will pay to the claimant for the fire loss. When the claim is finally settled, the company pays an amount, Y, to the claimant. The company has determined that x and y have the joint density function,

$$f(x,y) = \begin{cases} \dfrac{2}{x^2(x-1)} y^{-(2x-1)/(x-1)}, & for\ x > 1,\ y > 1 \\ 0 & ,\ otherwise \end{cases}$$

Given that the initial claim estimated by the company is 2, determine value of the marginal distribution of X and the probability that the final settlement is between 1 and 3.

3.(a) The School of Engineering of KNUST is planning to buy a new machine of either type A or B for one of its laboratories. The number of daily repairs x_1 required to maintain a machine of type A is a random variable with mean and variance both equal to $0.12t,$ where t denotes the number of hours of daily operation. The number of daily repairs x_2 required to maintain a machine of type B is a random variable with mean and variance both equal to $0.15t$. The daily cost of operating A is $C_A(t) = 10t + 30x_1^2,$ while that of B is

$C_B(t) = 8t + 30x_2^2$. Assume that the repairs take negligible time and that each night the machines are tuned so that they operate essentially like new machines at the start of the next day.

(i) Find the expected cost in terms of t for each machine.

(ii) Which machine gives the minimum expected daily cost if a work day consists of 10 hours?

(b) Suppose the random variables, x and y have the bivariate probability function,

$$f(x, y) = \begin{cases} \frac{1}{6}(x+2y), \ for \ (0, 0), (0, 1), (1, 0), (1, 1), \\ 0 \qquad , elsewhere \end{cases}$$

(i) What are the marginal distributions of x and y?

(ii) Compute the probability, $P(x \le y)$.

(c) (i) State the properties of the joint probability distribution, $f(x, y)$, where x and y are discrete random variables.

(ii) Define the correlation coefficient of the two random variables x and y.

4.(a) Let x and y be continuous random variables with means, μ_x and μ_y and variances σ_x^2 and σ_y^2. If the conditional mean, $E(y/x)$ is expressed of the form; $\phi(x) = \alpha + \beta x$, determine the values of α and β, and show that the conditional variance of y given x, $\sigma_{y/x}^2 = \sigma_y^2(1 - \rho_{xy}^2)$, where ρ_{xy} is correlation coefficient.

(b) Let the random variables x and y have the joint probability distribution,

$$f(x, y) = \frac{4}{4xy}, \ for \ x = 1, 2 \ and \ y = 2, 3$$

(i) Find the correlation coefficient between x and y.

(ii) Compute the probability, $P(x = y)$

(c) Suppose the joint probability density function for random variables x and y is given by $f(x, y) = \begin{cases} k(3x^2 + xy), & 0 \le x \le 1, \ 0 \le y \le 2 \\ 0 & , elsewhere \end{cases}$,

where $k > 0$ is a constant.

(i) Determine the value of k and the marginal distributions of x and y.

(ii) Find the conditional mean, $E(y|x = 1)$, and variance, $Var(y|x = 1)$

(ii) Compute the probabilities $P(x + y \ge 1)$ and $P(y > 1|x = 0..5)$.

(d) An insurance policy reimburses a loss up to a benefit limit of 10. The policy holder's loss (Y) follows a distribution with density function,

$$f(x, y) = \frac{2}{y^3}, \ for \ y > 1, \ and \ 0, \ elsewhere.$$

What is the expected value of the benefit paid under the insurance policy?

SESSION 2-1: FUNCTIONS OF RANDOM VARIABLES

The theory of functions of random variables considers the transformation of the random variable, $Y = h(X_1, X_2, X_3, \ldots, X_k),$ where $k \geq 1$ and finds its probability distribution function, given the joint probability distribution functions of the random variables, $X_1, X_2, X_3, \ldots, X_k$. We will first consider $Y = h(X)$, a function of only one random variable (X) and find its distribution functions (the *pdf* and *cdf*), $f(y)$ and $F(y)$, and also consider the multivariate transformations, $Y_i = h(X_1, X_2, X_3, \ldots, X_k),$ and find the joint of $Y_i = h(X_1, X_2, X_3, \ldots, X_k),$ $i = 1, 2, \ldots, k.$

2-1.1 One Dimensional Transformations

Given a random variable X, we can, generally, find the probability density function of $Y = h(x)$, following the steps:

(i) Find the cumulative distribution function (cdf), $F_y(y) = P\left(h(x) \leq y\right)$ of Y.

(ii) Find $F_y(y) = F_y^1(y) = \dfrac{F_y(y)}{dy} = f(y)$.

If the function h is one-to-one, we may use the following theorem:

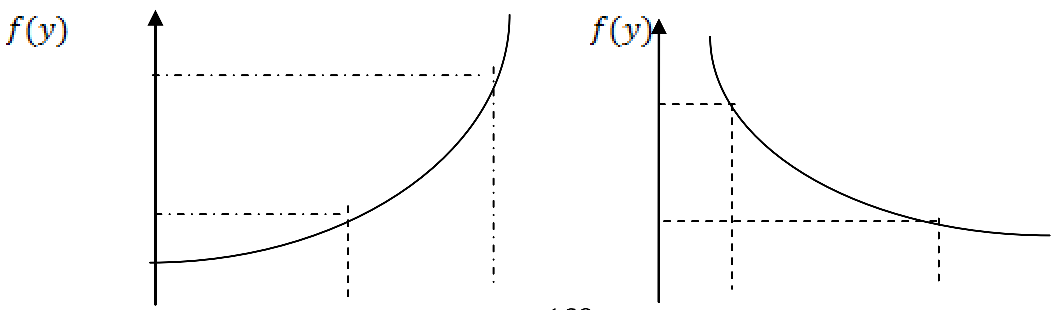
- ***Theorem 1.2***

  Let X be a continuous random variable with probability density function (pdf), $f(x)$, for $a < x < b$. Let $Y = h(x)$, which is strictly monotonic and differentiable function, then the probability density function of Y is given by

$$f_y(y) = f_x\left(h^{-1}(y)\right)\left|\frac{dh^{-1}(y)}{dy}\right|, \; c < x < d$$

(1.16)

<u>*Proof of Theorem*</u>



y y

increasing $Y = h(X)$ decreasing $Y = h(X)$

(i) Assume that $Y = h(x)$, is strictly increasing and continuous functions of X, then the cdf of Y is

$$F_y(y) = P(Y \le y) = P\left(h(y) \le y\right) = P\left(X \le h^{-1}(y)\right) = \int_{-\infty}^{h^{-1}(y)} f_x(x)dx = F_x(x)$$

Differentiating both sides with respect to y

$$F_y^{1}(y) = f_y(y) = f_x\left(h^{-1}(y)\right)\frac{dh^{-1}(y)}{dy}$$

(1.17)

(ii) If $y = h(x)$ is strictly decreasing,

$$F_y(y) = P(Y \le y) = P\left(h(y) \le y\right) = P\left(X \ge h^{-1}(y)\right)$$

$$= \int_{h^{-1}(y)}^{\infty} f_x(x)dx = 1 - P\left(X < h^{-1}(y)\right) = 1 - F_x\left(h^{-1}(y)\right)$$

Differentiating both sides with respect to y

$$F_y^{1}(y) = f_y(y) = -\frac{dF_x(X)}{dy} = -f_x\left(h^{-1}(y)\right)\frac{dh^{-1}(y)}{dy}$$

(1.18)

where $\dfrac{dh^{-1}(y)}{dy} < 0$.

Combining equations (1.17) and (1.18) we have

$$f_y(y) = f_x\left(h^{-1}(y)\right)\left|\frac{dh^{-1}(y)}{dy}\right|$$

Note

If $Y = h(x)$ is not a monotonic function of X, we cannot apply the above theorem. In this case we use the general method. Suppose that X is a discrete random variable with distribution function $f(x)$. Then the probability distribution function of $Y = h(x)$ is

$$f_y(y) = P(Y = y) = P\left(X = h^{-1}(y)\right) = f_x\left(h^{-1}(y)\right)$$

- ***Theorem 1.3***

Suppose that X is a discrete random variable with probability mass function (*pmf*), $f(x)$. Let $Y = h(x)$ define a one-to-one transformation between the values of X and Y so that the equation $Y = h(x)$ has the unique solution in terms of Y, $X = h^{-1}(y)$. Then the probability distribution of Y is $f_y(y) = f\left(h^{-1}(y)\right)$

Illustrative Example 1.3

1. Suppose X is a random variable with the following $pdf's$ and the corresponding defined transformations, $Y = h(x)$:

(a) $f_x(x) = \begin{cases} 2x, \ 0 < x < 1 \\ 0 \ , \textit{elsewhere} \end{cases}$ and $Y = 3x + 6$

(b) $f_x(x) = \begin{cases} 2e^{-2x}, \ x > 0 \\ 0 \quad , \textit{elsewhere} \end{cases}$ and $Y = e^x$.

<u>Solution</u>

(a) Given $f_x(x) = 2x, \ 0 < x < 1$ and $Y = 3x + 6$, which is strictly increasing and differentiable. Then both the general method and ***Theorem 1.1*** are applicable and so we have

$$X = h^{-1}(y) = \frac{1}{3}(y - 6) \text{ and } \left|\frac{dx}{dy}\right| = \left|\frac{dh^{-1}(y)}{dy}\right| = \frac{1}{3}$$

(i) Then by the theorem,

$$f_y(y) = f_x\left(dh^{-1}(y)\right)\left|\frac{dx}{dy}\right|$$

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

$$f_y(y) = \begin{cases} \frac{2}{9}(y - 6), \ 6 < y < 9 \\ 0 \qquad , \textit{elsewhere} \end{cases}$$

(ii) By the general method

$$F_x(x) = \int_0^x f(t)dt$$
$$= \int_0^x 2t\,dt = t^2\Big|_0^x$$
$$= x^2,\ 0 < x < 1$$
$$F_y(y) = F_x(x) = \left(\frac{y-6}{3}\right)^2$$
$$F_y(y) = F_y^1(y) = 2\left(\frac{y-6}{3}\right)\left(\frac{1}{3}\right)$$
$$f_y(y) = \begin{cases} \frac{2}{9}(y-6),\ 6 < y < 9 \\ 0 \qquad\qquad ,\ elsewhere \end{cases}$$

(b) Given the *pdf*, $f_x(x) = 2e^{-2x},\ x > 0$ and $Y = e^x$, then $X = \ln y$:

(i) By the general method,

$$F_x(x) = \int_0^x f(t)dt = \int_0^x 2e^{-2t}dt = -e^{-2t}\Big|_0^x = 1 - e^{-2x},\ x > 1$$
$$F_y(y) = F_x(x) = 1 - e^{-2\ln y} = 1 - y^{-2},\ x > 1$$
$$f_y(y) = \begin{cases} 2y^{-3},\ y > 1 \\ 0 \quad\ ,\ elsewhere \end{cases}$$

(ii) By the theorem,

$$f_y(y) = f_x(X)\left|\frac{dx}{dy}\right|$$
$$= 2e^{-2\ln y}\left(\frac{1}{y}\right) = \frac{2}{y^3}$$
$$f_y(y) = \begin{cases} 2y^{-3},\ y > 1 \\ 0 \quad\ ,\ elsewhere \end{cases}$$

2.(a) Let X be a discrete random variable with distribution function: $f(-1) = 0.2$, $f(0) = 0.5$ and $f(1) = 0.3$. Determine the probability function of y if

(i) $Y = 2x$ and

(ii) $Y = x^2$.

(b) Let X be a Geometric random variable with the distribution,

$$f(x) = \frac{3}{4}\left(\frac{1}{4}\right)^{x-1}, \; x = 1, 2, 3, \ldots$$

Find the probability distribution of the random variable $Y = x^2$.

(c) The amount of damage, X $(in\ GH¢\ m)$ caused by fire outbreak in a region has the probability distribution

X	50	250	500	1000
$f(x)$	0.2	0.3	0.4	0.1

The District Assembly of the region is considering whether or not to buy an insurance policy to cover fire damage. The policy will pay 80% of the damage in excess of GH¢100m, subject to a maximum total payment of GH¢500m. Compute the expected payment from the policy.

Solution

(a)(i) If $Y = 2x$, then the transformation is one-to one and it is simple to see that:

$Y = 2x$	-2	0	2
$f(y)$	0.2	0.5	0.3

(ii) If $Y = x^2$, then $(-1)^2 = 1^2 = 1$, $f(0) = 0.5$, $f\left((-1)^2\right) + f\left(1^2\right)$, $f(0) = 0.5$, $f\left((-1)^2\right) + f\left(1^2\right) = 0.2 + 0.3 = 0.5$. This transformation is not one-to-one and by the general application,

$Y = x^2$	0	1
$f(y)$	0.5	0.5

The cdf of Y can be defined in a similar way as

$$F_y(y) = \mathrm{P}\left(Y \le y\right) = \sum_{x_i : h(x_i) \le y} f_x(x_i)$$

(b) The given transformation, $Y = x^2$, defines a one-to-one correspondence between X and Y values. $Y = x^2$ and $X = \sqrt{x}$. Hence

$$f(y) = \begin{cases} f\left(\sqrt{x}\right) = \dfrac{3}{4}\left(\dfrac{1}{4}\right)^{\sqrt{y}-1}, & y = 1, 4, 9, \ldots \\ 0, & elsewhere \end{cases}$$

(c) Let Y be the payment from the policy. So if $X = 0$, then $Y = 0$, since the damage amount does not exceed GH¢100m. For $X > 50$, we have $Y = \min\left\{0.8(x-100), 500\right\}$. The probability distribution of the payments is

X	50	250	500	1000
Y	0	120	320	500

$f(y)$	0.2	0.3	0.4	0.1

The expected payment is

$$E(Y) = \sum_{\forall y} yf(y)$$
$$= 0(0.2) + 120(0.3) + 320(0.4) + 500(0.1)$$
$$= 214 \text{ (or GH¢214m)}$$

2-1.2 Multivariate Transformations

Suppose that the random variables $X_1, X_2, \ldots, X_k$ have a joint probability distribution, $f(X_1, X_2, \ldots, X_k)$. If the random variables $Y_1, Y_2, \ldots, Y_k$ are defined in terms of $X_1, X_2, \ldots, X_k$, then we derive the method of finding the joint probability distribution of $Y_1, Y_2, \ldots, Y_k$ in terms of the joint distribution of $X_1, X_2, \ldots, X_k$

- ***Theorem 1.4***

Let $X_1, X_2, \ldots, X_k$ be continuous random variables having joint probability density function, $f(x_1, x_2, \ldots, x_k)$. Let the random variables, $Y_1, Y_2, \ldots, Y_k$ defined by $Y_i = h_i(X_1, X_2, \ldots, X_k),\ i = 1, 2, \ldots, k$ be a one-to-one transformation between the points $(X_1, X_2, \ldots, X_k)$ and $(Y_1, Y_2, \ldots, Y_k)$ and further assume that:

(i) The equations $Y_i = h_i(X_1, X_2, \ldots, X_k),\ i = 1, 2, \ldots, k$ have a unique solution, $X_i = h_i^{-1}(Y_1, Y_2, \ldots, Y_k)$, and

(ii) The partial derivative $\dfrac{\partial y_i}{\partial x_j}$ exists and it is continuous.

Then the joint probability density function of $Y_1, Y_2, \ldots, Y_k$ is given by

$$f(y_1, y_2, ..., y_k) = f\left(X_1 = h_1^{-1}(y_1, y_2, ..., y_k), ..., X_k = h_k^{-1}(y_1, y_2, ..., y_k)\right)\left|J(x_1, x_2, ..., x_k)\right|^{-1}$$

where

$$J(x_1, x_2, \ldots, x_k) = \begin{vmatrix} \dfrac{\partial h_1}{\partial x_1} & \dfrac{\partial h_1}{\partial x_2} & \cdots & \dfrac{\partial h_1}{\partial x_k} \\ \dfrac{\partial h_2}{\partial x_1} & \dfrac{\partial h_2}{\partial x_2} & \cdots & \dfrac{\partial h_2}{\partial x_k} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial h_k}{\partial x_1} & \dfrac{\partial h_k}{\partial x_2} & \cdots & \dfrac{\partial h_k}{\partial x_k} \end{vmatrix} \neq 0 \qquad (1.19)$$

is called the *jacobian* of the transformation $Y_i = h_i(X_1, X_2, \ldots, X_k)$, and x_i's are defined implicitly in terms of y_i's.

Equation (1.19) should easily be seen as a special case of the equation (1.16) with $f(x)$ corresponding to $f(x_1, x_2, \ldots, x_k)$, $h^{-1}(y)$ playing the role of the inverse transformation, $X_i = h_i^{-1}(Y_1, Y_2, \ldots, Y_k)$, and $\left|\frac{dh^{-1}(y)}{dy}\right|$, being equivalent to the jacobian of the inverse transformation $X_i = h_i^{-1}(Y_1, Y_2, \ldots, Y_k)$. In particular, if for example, that $h_1(x_1, x_2)$ and $h_2(x_1, x_2)$ are one-to-one functions and x_1 and x_2 are unique solutions to the transformation $Y_1 = h_1(x_1, x_2)$ and $Y_2 = h_2(x_1, x_2)$, then the joint probability density function of Y_1 and Y_2 is

$$f(y_1, y_2) = f\left(X_1 = h_1^{-1}(y_1, y_2), X_2 = h_2^{-1}(y_1, y_2)\right)\left|J(x_1, x_2)\right|^{-1}$$

and the jacobian of the transformation is the determinant,

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} \tag{1.20}$$

- ***Theorem 1.5***

If X is a vector of discrete random variable with joint probability distribution function, $f_x(x)$ and $Y = h(x)$ defines a one- to- one transformation, then the joint probability mass function Y is

$$f_y(y_1, y_2, \ldots, y_k) = f_x(x_1, x_2, \ldots, x_k) \tag{1.21}$$

where $X_1, X_2, \ldots, X_k$ is the unique solution of $Y = h(x)$ and consequently depends on $y_1, y_2, \ldots, y_k$. If the transformation is not one to one, and if a partition exists, say $A_1, A_2, \ldots, A_k$ such that the equation $Y = h(x)$ has a unique solution $x = X_j$ or $X_j = (X_{1j}, X_{2j}, \ldots, X_{kj})$ over A_j, then the probability mass function of Y is

$$f_y(y_1, y_2, \ldots, y_k) = \sum_j f_x(x_{1j}, x_{2j}, \ldots, x_{kj}) \tag{1.22}$$

Illustrative Example 1.4

1.(a) Suppose that X_1 and X_2 are independent standard normal random variables and that $Y_1 = X_1$ and $Y_2 = X_1 + X_2$. Then the joint distribution of Y_1 and Y_2 is a bivariate normal distribution. The Jacobian of the transformation

$$J = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial x_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1$$

The inverse transformations are $X_1 = Y_1$ and $X_2 = Y_1 + Y_2$. The joint pdf of Y_1 and Y_2 is

$$f(y_1, y_2) = \frac{1}{2\Pi}\exp\left\{-\frac{1}{2}\left(y_1^2 + (y_2 - y_1)^2\right)\right\}(1)^{-1} = \frac{1}{2\Pi}\exp\left\{-\frac{1}{2}\left(2y_1^2 + y_2^2 - 2y_1y_2\right)\right\}$$

(b) Let X_1 and X_2 the joint distribution

$$f(x_1, x_2) = \begin{cases} \frac{3}{8}\left(x_1 + 2x_2\right), \ 0 \le x_1 \le 1, \ 0 \le x_2 \le 2 \\ 0, \qquad\qquad , \ elsewhere \end{cases}$$

Consider the transformation, $Y_1 = X_1 - X_2$ and $Y_2 = X_1 + X_2$, where

$X_1 = \frac{1}{2}\left(Y_1 + Y_2\right)$ and $X_2 = \frac{1}{2}\left(Y_2 - Y_1\right)$

The jacobian of the transformation $Y = h\left(x\right)$ is

$$J = \begin{vmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{vmatrix} = 1/2$$

The joint probability density function of Y_1 and Y_2 is

$$f(y_1, y_2) = f_x(x_1, x_2)\left|J\right|$$

$$= \frac{3}{8}\left[\frac{1}{2}\left(y_1 + y_2\right) + \frac{1}{2}2\left(y_2 - y_1\right)\left(\frac{1}{2}\right)\right]$$

$$= \begin{cases} \frac{3}{32}\left(3y_2 - y_1\right), \ -2 \le y_1 \le 1, \ 0 \le y_2 \le 3 \\ 0 \quad , \ elsewhere \end{cases}$$

2. A satellite has two independent power systems. The lifetime of the primary system is distributed exponentially with mean 10 years. The lifetime of the backup system is distributed exponentially with mean 5 years. Obtain the probability density function of Y, the total time that the power system will remain functional, if the backup system is operated continuously as soon as the primary system fails.

Solution

Let $y = x_1 + x_2$, where x_1 and x_2 are the lifetimes of the primary and backup systems. The joint density of x_1 *and* x_2 is

$$f(y_1, y_2) = f(x_1).\, f(x_2)$$

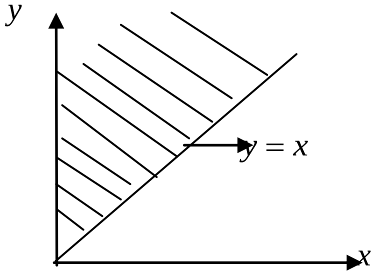
$$= \left(0.10e^{-0.x_1}\right)\left(0.20e^{0.20x_2}\right) = \begin{cases} 0.02e^{0.1x_1 - x_2}, x_1, \ x_2 > 0 \\ 0 \quad , \ elsewhere \end{cases}$$

Let $x = x_1$ and hence $y = x + x_2 \Leftrightarrow x_1 = x$ and $x_2 = y - x.$

We find the joint pdf of x and y as follows:

The jacobian of the transformation $x = x_1$ and $y = x_1 + x_2$

$$J = \begin{vmatrix} 1 & 0 \\ 1 & -1 \end{vmatrix} = -1$$



$$f(x, y) = f(x_1, x_2)\left|J\right|^{-1}$$

$$= \begin{cases} 0.02e^{-0.3x-0.2y}, & 0 < x < y, 0 < y < \infty \\ 0 & , elsewhere \end{cases}$$

Hence we find the marginal distribution of y:

$$f(y) = \int_0^y f(x, y)dx$$

$$= \int_0^y 0.02e^{-0.3x-0.2y}dx$$

$$= \left(0.02e^{-2y}\right)\left(-e^{-0.3y}\right)\Big|_0^y$$

$$= \begin{cases} 0.02\left(e^{-2y} - e^{-0.5y}\right), & y > 0 \\ 0 & , elsewhere \end{cases}$$

2-1.3 Common Transformations in the Insurance Industry

We will look at three transformations with common applications in insurance. The random variable here is continuous, representing the loss to an insured party.

- ***Inflation***: We may wish to model the cost of insurance claims if we assume that losses increase in each year due to inflation.
- ***Deductible***: We may wish to model the cost of insurance claims if we apply a deductible to each loss. This means that the insurance company will pay zero if a loss is below the deductible, and will otherwise make a claim payment equal to the loss less the deductible.
- ***Policy Limit***: We may model the cost of insurance claims if we apply a policy limit to each loss. This means that every claim payment made by the insurance company is restricted to a maximum of the policy limit.
- ***Claim Inflation***: Suppose that losses are subject to $100r\%$ inflation next year. For example, if losses are subjected to 5% inflation, $r = 0.05.$ let y represent the loss distributions for next year. The transformation is thus $y = (1+r)x,$ which is one-to-one and differentiable. So, we can find the probability density function for year's losses using the method of transformations. In this case, we let $y = g(x) = (1+r)x,$ then $x = g^{-1}(y) = \dfrac{y}{(1+r)}.$ Then the probability density function is calculated as follows:

$$f_y(y) = f_x\left(g^{-1}(y)\right)\left|\frac{dx}{dy}\right| = f_x\left(g^{-1}\left(\frac{y}{(1+r)}\right)\right)\left|\frac{dx}{dy}\right|.$$

***Illustrative Example* 1.5**

- <u>*Inflation*</u>

For example, in 2005, a random loss, *X*, is distributed uniformly on the interval [0, 1000]. If insurance losses are expected to rise by 4% inflation per year, then the probability density function of the loss, y is obtained as follows:

The probability density function x is $f_x(x) = \frac{1}{1{,}000}, \; for \; 0 \le x \le 1{,}000.$

The loss in 2006 is the transformation, $y = 1.04x$, which is one–to–one and differentiable. Then $x = \frac{y}{1.04}, \; \frac{dx}{dy} = \frac{1}{1.04}$ and the probability density function is

$$\begin{aligned} f_y(y) &= f_x(x)\left|\frac{dx}{dy}\right| \\ &= \frac{1}{1{,}000}.\frac{1}{1.04} = \frac{1}{1{,}040}, \; 0 \le y \le 1{,}040. \end{aligned}$$

- <u>*Deductible*</u>

Let y represent the claim payment made by an insurance company that applies a deductible d to each loss x. Then:

$$y = (x-d)_+ = \max\{(x-d), 0)\} = \begin{cases} x-d \, , \; x > d \\ 0 \quad\;\; , \; 0 < x \le d \end{cases}$$

The expected value of the claim payments are computed as computed follows:

$$E(y) = \int_{R_x} y f_x(x)dx = \int_0^d 0.\,f_x(x)dx + \int_d^\infty (x-d)\,f_x(x)dx = \int_d^\infty (x-d)\,f_x(x)dx$$

$$E(y^2) = \int_{R_x} y^2 f_x(x)dx = \int_0^d 0^2.\,f_x(x)dx + \int_d^\infty (x-d)^2 f_x(x)dx = \int_d^\infty (x-d)^2 f_x(x)dx$$

Hence the variance of y is

$$Var(y) = E(y^2) - \left(E(y)\right)^2 = \int_d^\infty (x-d)^2 f_x(x)dx - \left(\int_d^\infty (x-d)\,f_x(x)dx\right)^2$$

For example, if a random loss x is distributed uniformly on the interval $[0, 1{,}000]$. To calculate the expected claim payment (y), if a deductible of 100 is applied to each loss:

$$\begin{aligned} E(y) &= \int_{100}^{1{,}000} (x-100)\,f_x(x)dx \\ &= \int_{100}^{1{,}000} (x-100)\frac{1}{1{,}000}dx \\ &= \frac{1}{2{,}000}\left(x-100\right)^2\Big|_{100}^{1{,}000} = \frac{900^2}{2{,}000} = 405 \end{aligned}$$

- *Policy Limit*

Let y represent the claim payment made by an insurance company that applies a policy limit L to each loss $x.$ Then

$$y = X \wedge L = \min(X,\ L) = \begin{cases} x,\ 0 < x < L \\ L,\ x \geq L \end{cases}$$

We can calculate the expected value of the claim payments as:

$$E(y) = \int_0^L x\ f_x(x)dx + \int_L^\infty L\ f_x(x)dx = \int_0^L x\ f_x(x)dx + L\left(1 - F_x(L)\right)$$

$$E(y^2) = \int_0^L x^2\ f_x(x)dx + \int_L^\infty L^2\ f_x(x)dx$$

Hence the variance of y is

$$Var(y) = E(y^2) - \left(E(y)\right)^2$$

$$= \int_d^\infty x^2\ f_x(x)dx + L^2\left(1 - F(L)\right) - \left(\int_0^L xf_x(x)dx + L\left(1 - F(L)\right)\right)^2$$

For example if a random loss x is distributed uniformly don the interval [0, 1000]. Then the expected claim payment (y) if a policy limit of 50 is applied to each loss is

$$E(y) = \int_0^{500} x.\frac{1}{1{,}000}.dx + \int_{500}^{1{,}000} 500.\frac{1}{1{,}000}dx$$

$$= \left.\frac{x^2}{2{,}000}\right|_0^{500} + 500\left(1 - F(500)\right)$$

$$= \frac{500^2}{2{,}000} + 500\left(1 - \frac{500}{1{,}000}\right)$$

$$= 125 + 250 = 375$$

Unit 4

ESTIMATION OF POPULATION PARAMETERS

Usually parameters, the numerical characteristics of a population, are unknown. They are estimated using random sample observations to suggest the likely value or set of values for each unknown parameter. The process of determining the values of such parameters using a randomly obtained sample observations is called *estimation.* Estimation problem leads to two types of estimates, *point estimates* and *interval estimates*. When a single value is suggested for an unknown parameter, this single value is called point estimate. For example, if a random sample of ten observations whose values are 8.7, 7.6, 10.2, 8.2, 9.6, 8.0, 8.8, 9.6, 8.0 and 8.2, is selected from a population distribution whose mean, μ is to be estimated. A common procedure is to use the sample mean, $\bar{x} = 8.69$ as the estimate of the population mean, μ. Sometimes, it is possible to determine the interval, (8.1, 9.3) within which it can be established that the unknown parameter, μ will lie with some specified degree of confidence. The interval (8.1, 9.3) is an interval estimate of μ. The error made in estimating for a parameter is also of great interest.

Briefly, the unit gives the some desirable properties of point estimators and a detailed outline for confidence intervals estimation.

Learning Objectives:

After studying the unit, students will be able to obtain estimates of population parameters and develop the understanding of concepts of goodness of these estimates. Specifically, they should be able to:

- Determine whether a point estimator is unbiased and efficient, and also show that it is consistent.
- Compute and interpret confidence intervals for the population mean, population proportion, and population variance.
- Compute and interpret confidence intervals for the difference between two population means, difference between two population proportions, and ratio of two population variances.
- Compute the size of a sample given an error of estimation and the confidence coefficient.

SESSION 1-4: POINT ESTIMATION OF PARAMETERS

1-4.1 Introduction

The estimation of a parameter is done by using an appropriate statistic. A statistic used to estimate the parameter, θ is called *point estimator*, denoted $\hat{\theta}$. The parameters which are frequently estimated are the population mean (μ), proportion (P), variance or standard deviation $(\sigma^2 \text{ or } \sigma)$, the difference between two means $(\mu_1 - \mu_2)$, and proportions $(\hat{p}_1 - \hat{p}_2)$. The statistics of these parameters are defined as follows:

- The sample mean, $\bar{x} = \frac{1}{n}\sum_{i=}^{n} x_i$, which is the point estimator for μ.
- The sample proportion, $\hat{p} = \frac{x}{n}$, the point estimator for P.

- The sample variance, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, the point estimation for σ^2.
- The mean and proportion sample differences, $(\bar{x}_1 - \bar{x}_2)$ and $(\hat{p}_1 - \hat{p}_2)$ are point estimators for $(\mu_1 - \mu_2)$ *and* $(p_1 - p_2)$ respectively.
- The ratio of two sample variances s_1^2 / s_2^2, the point estimator for σ_1^2 / σ_2^2

These point estimators are statistics (or rules) which specify how to use the sample observations to estimate the unknown parameters. The numerical values assumed by these statistics when evaluated for the given sample observations are called point estimates. For example, if the mean income of the sample of n = 5,000 persons, estimated by $\bar{x} = 1/n\sum_{i=}^{n} x_i$, yielded a value, ¢1,200,000 per month, this value is the point estimate for μ, the exactly mean income of the entire 50,000 persons.

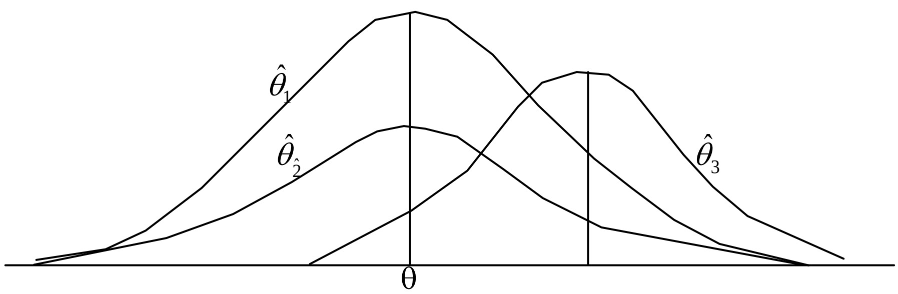
The error made in estimating for a parameter is also of great interest. Suppose we wish to determine the mean income, μ of a population of, say, 50,000 persons. If we know the income of each number of the population, we could be able to calculate μ exactly. Unfortunately, we will usually not have this complete knowledge because we may not have the resources to seek the information personally from all individuals in the population, or from their employers. What is done is to take a sufficiently large sample subject to available resources and to use the mean of this sample as an estimate the unknown population mean. If we select 5,000 persons (that is, a sample of 10% of the population) we can hardly expect that the mean income of these 5,000 persons $(\bar{x})$ will be exactly equal to the mean of the entire population of 50,000 persons. The difference $(\bar{x} - \mu)$ represents the error. The aim of estimation theory is to devise estimation procedures with the smallest possible errors.

In order to decide which point estimator of a particular parameter is the best one to use, we need to examine their statistical properties and develop some criteria for comparing estimators and also give us the most economical information. Ideally, we want an estimator to generate estimates that can be expected to be close in value to the parameter.

1-4.2 Properties of Point Estimators

Some basic properties of point estimators are:

- *Unbiasedness*: The point estimator, $\hat{\theta}$ is said to be unbiased for the parameter θ if $E(\hat{\theta}) = \theta$. If the estimator is not unbiased, then the difference, $E(\hat{\theta}) - \theta$ is called the *bias* of the estimator $\hat{\theta}$.
- *Minimum variance unbiased estimator*. If we consider all unbiased estimators for θ, the one with smallest variance is called the minimum unbiased estimator $(MVUE)$.
- *Efficient estimator*: Suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimators for the parameter θ. This indicates that the distribution of each estimator is centred at the true value of θ. However, If $\hat{\theta}_1$ has smaller variance it is more likely to produce an estimate close to the true value of θ. It is therefore said to be more efficient then $\hat{\theta}_2$. The relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is defined by $eff\,(\hat{\theta}_1) = \dfrac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$. A logical principle of estimation, when selecting among several estimator unbiased estimators, is to choose the estimator that has minimum variance. The diagram below shows the sampling distributions of three estimators, $\hat{\theta}_1, \hat{\theta}_2$ *and* $\hat{\theta}_3$. The estimators $\hat{\theta}_1$ *and* $\hat{\theta}_2$ are unbiased while $\hat{\theta}_3$ is biased.



- *Consistency*: This property indicates that a said point estimate becomes close to the parameter as the sample size, n become sufficiently large. If $\hat{\theta}$ is an estimator for the parameter θ. Then it is said to be consistent if $\lim_{n\to\infty} E(\hat{\theta}) = \theta$ and $\lim_{n\to\infty} Var(\hat{\theta}) = 0$

- *Sufficiency*: An estimator $\hat{\theta}$ is to be sufficient if it utilizes all the information in a sample relevant to the estimation of θ, that is if all the knowledge about θ that can be gained from the individual sample values and their order can just as well be gained from the value of $\hat{\theta}$ alone.

There are various methods for obtaining points given the random observations $x_1, x_2, \ldots, x_n$. These include the method of moments, method of maximum likelihood, and the least squares method. These methods are fully discussed in MATH 272.

Illustrative Example 4.1:

1. The number of breakdowns per week for a certain mini-computer is a random variable having the Poison distribution with parameter μ. A random sample on the number of breakdowns per week is made up of the observations $y_1, y_2, y_3, \ldots, y_n$.

(a) Suggest an unbiased estimator for μ.

(b) The weekly cost of repairing these breakdowns is $C = 3y + y^2$ Show that $E(C) = 4\mu + \mu^2$ and find an unbiased estimator for $E(C)$ that makes use of the sample observations, $y_1, y_2, y_3, \ldots, y_n$.

Solution:

(a) The sample mean $(\bar{y})$ is shown to be an unbiased estimator for μ as flows:

$$E(\bar{y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i)$$

$$= \frac{1}{n}\left[E(y_1) + E(y_2) + E(y_3) + \ldots + E(y_{n)}\right]$$

$$= \frac{1}{n}(\mu + \mu + \ldots + \mu) = \frac{1}{n}(n\mu) = \mu$$

(b) $E(C) = E(3y + y^2) = 3E(y) + E(y^2)$

$$= 3E(y) + Var(y) + [E(y)]^2$$

$$= 3\mu + \mu + \mu^2 = 4\mu + \mu^2,$$

which is unbiasedly estimated by $\hat{E}(C) = 4\bar{y} + \bar{y}^2$

2. Suppose y_1, y_2, y_3, y_4, y_5 denote a random sample of some population with $E(y_i) = \lambda$ and $Var(y_i) = \lambda^2$, for all $i = 1, 2, 3, 4, 5.$ The following statistics are suggested as estimators for λ.

(i) $\hat{\theta}_1 = y_1$ (ii) $\hat{\theta}_2 = \frac{1}{2}(y_1 + y_3)$

(iii) $\hat{\theta}_3 = \frac{1}{2}(y_2 + 2y_3)$ (iv) $\hat{\theta}_4 = \frac{1}{5}(y_1 + y_2 + y_3 + y_4 + y_5) = \bar{y}$

Determine the most efficient estimator among the unbiased estimators.

Solution:

Given that $E(y_i) = \lambda$ and $Var(y_i) = \lambda^2$, for all $i = 1, 2, 3, 4, 5,$

(a) We first find the unbiased estimators.

(i) $E(\hat{\theta}_1) = E(y_1) = \lambda$

(ii) $E(\hat{\theta}_2) = \frac{1}{2}E(y_1 + y_3) = \frac{1}{2}[E(y_1) + E(y_3)] = \frac{1}{2}(\lambda + \lambda) = \frac{1}{2}(2\lambda) = \lambda$

(iii) $E(\hat{\theta}_3) = \frac{1}{2}E(y_2 + 2y_3) = \frac{1}{2}[E(y_2) + 2E(y_3)] = \frac{1}{2}(\lambda + 2\lambda) = \frac{3}{2}\lambda \neq \lambda$

(iv) $E(\hat{\theta}_4) = \frac{1}{5}E(y_1 + y_2 + y_3 + y_4 + y_5)$

$$= \frac{1}{5}[E(y_1) + E(y_2) + E(y_3) + E(y_4) + E(y_5)] = \frac{1}{5}(5\lambda) = \lambda$$

The unbiased estimators are $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_4$. The estimator $\hat{\theta}_3$ is biased because $E(\hat{\theta}_3) \neq \lambda$. It over estimates λ.

(b) We determine the minimum variance of the unbiased estimators.

(i) $Var(\hat{\theta}_1) = Var(y_1) = \lambda^2$

(ii) $Var(\hat{\theta}_2) = Var\left[\frac{1}{2}(y_2 + y_3)\right] = \frac{1}{4}[Var(y_2) + Var(y_3)] = \frac{1}{4}(\lambda^2 + \lambda^2) = \frac{1}{2}\lambda^2$

(iii) $Var(\hat{\theta}_4) = Var\left[\frac{1}{5}(y_1 + y_2 + y_3 + y_4 + y_5)\right]$

$$= \frac{1}{25}[Var(y_1) + Var(y_2) + \ldots + Var(y_5)] = \frac{1}{25}(5\lambda^2) = \frac{1}{5}\lambda^2$$

The minimum variance among the unbiased estimators is $Var(\hat{\theta}_4)$. Hence $\hat{\theta}_4$, the minimum variance unbiased estimator, is chosen the most efficient estimator.

1-4.3 Self Administered Questions (Exercise 4.1)

1(a) Suppose we have a random sample of size 20 from a population where $E(x) = \mu$ and $Var(x) = \sigma^2$. Let $\bar{x}_1 = \frac{1}{2n}\sum_{i=1}^{2n} x_i$ and $\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i$ be no estimators for μ. Which is the better estimator for μ? Explain your choice.

(b) Let $x_1, x_2, x_3, \ldots, x_n$ denote a random sample from a population with mean, μ haring μ and variance, σ^2. Consider the following estimators of μ

$$\hat{\theta}_1 = \frac{x_1 + x_2 + \ldots + x_7}{7}, \text{ and } \hat{\theta}_2 = \frac{2x - x_6 + x_4}{2}$$

(i) Is either estimator unbiased?

(ii) Which estimator is "best"? In what sense is it best?

2(a) Two meters are used to measure water pressure in a pipeline. One is known to be more accurate than the other but both are subject to random error. Readings by meter A are subject to a standard deviation, σ and readings by meter B have a standard deviation of 1.25σ. Meter A is used to take 6 independent measurements with a mean of $\bar{x}_A$. Meter B is used to take 10 independent measurements with a mean of $\bar{x}_B$. Suppose that both means are unbiased estimators for the true mean water pressure θ.

(i) Which estimator would you prefer if you had to choose between $\bar{x}_A$ and $\bar{x}_B$?

(ii) Find the relationship between the constants α and β if $Z = \alpha\bar{x}_A + \beta\bar{x}_B$ is to be an unbiased estimator θ.

(iii) Find the variance of Z.

(b) Suppose y_1, y_2 and y_3 denote a random sample from the exponential distribution with parameter θ. Consider the following as estimates of θ.

$$\hat{\theta}_1 = y_1 \qquad \hat{\theta}_2 = \frac{1}{2}(y_1 + y_2)$$

$$\hat{\theta}_3 = \frac{1}{3}(y_1 + 2y_3) \qquad \hat{\theta}_4 = \frac{1}{3}(y_1 + y_2 + y_3)$$

(i) Which of the above estimators is/are unbiased for θ?

(ii) What is the most efficient estimator?

2. Let $y_1, y_2, \ldots, y_n$ denote a random sample from the uniform distribution on the interval $(0, \beta)$. Let $\hat{\beta}_1 = \frac{2}{n}\sum_{i=1}^{n} y_i$ and $\hat{\beta}_2 = \left(\frac{n+1}{n}\right) max\left(y_1, y_2, \ldots, y_n\right)$ be estimators for β.

(i) Find the sampling distribution of $\hat{\beta}_2$.

(ii) Show that $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased for β.

(iii) Find the relative efficiency of $\hat{\beta}_1$ with respect to $\hat{\beta}_2$.

(iv) Which of the two estimators is more efficient?

SESSION 2-4: INTERVAL ESTIMATION OF PARAMETERS

2-4.1 Construction of Confidence Intervals

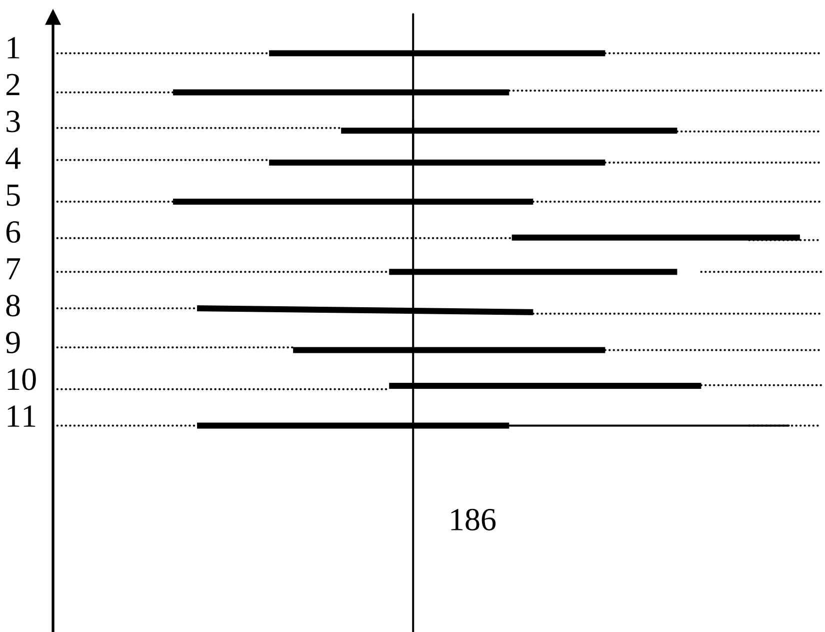
A point estimate, as already noted, provides likely values for an unknown parameter. The associated standard error gives a measure of the average estimation error involved in the use of that particular estimator. The two, together, suggest a range of likely values for the parameter. If the estimate of the population mean, μ, for example, is $\bar{x} = 8.7$ with a standard error, $Se(\bar{x}) = 0.60$, it means that μ is approximated by 8.7, subject to an average error of magnitude 0.60. The statement suggests that the value of the unknown population mean is likely to lie in the interval with end points 8.7 $\pm$ 0.60 (or 8.1, 9.3). With what degree of assurance can we assert that μ lies in this interval? A point estimator and its standard error are not enough to answer such a question. Interval estimation provides a range of values believed to contain the unknown parameter with a degree of confidence. The interval estimating the parameter is called *interval estimator* or *confidence interval*.

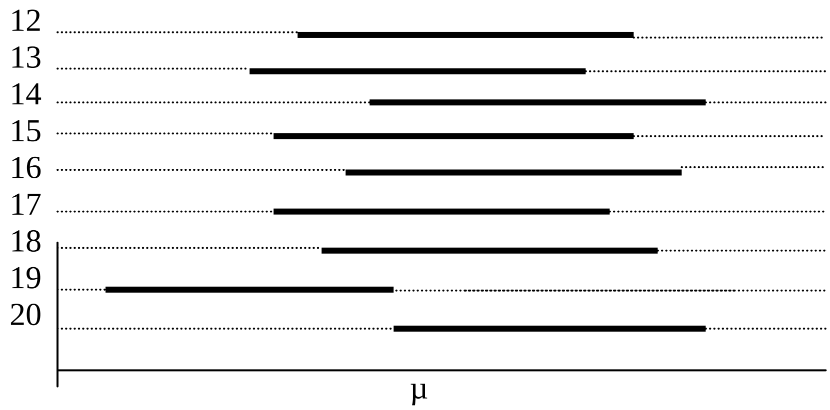
Let $\hat{\theta}$ be an estimator for θ with a known sampling distribution. We find two quantities $L(\hat{\theta})$ *and* $U(\hat{\theta})$ such that

$$P\left(\left|\hat{\theta}-\theta\right| \leq k.Se(\hat{\theta})\right) = 1-\alpha \text{ or } P\left(\hat{\theta}-k.Se(\hat{\theta}) \leq \theta \leq \hat{\theta}+k..Se(\hat{\theta}k.Se(\hat{\theta})\right) = 1-\alpha$$

where $L(\hat{\theta}) = \hat{\theta}-k.Se(\hat{\theta}),\ U(\hat{\theta}) = \hat{\theta}+k.Se(\hat{\theta}),\ k$ is the reliability coefficient or critical value determined by the sampling distribution of $\hat{\theta}$ and α is the probability that the interval will not contain θ. The interval $[L(\hat{\theta}), U(\hat{\theta})]$ is referred to as $(1-\alpha)100\%$ confidence interval for the parameter θ. The limits $L(\hat{\theta})$ *and* $U(\hat{\theta})$ are lower and upper limits respectively while the probability, $(1-\alpha)$ is called the confidence coefficient. The interval estimator $\lfloor L(\hat{\theta}),\ U(\hat{\theta}) \rfloor$ is said to be good if it has shorter length with higher confidence coefficient nearing 1.

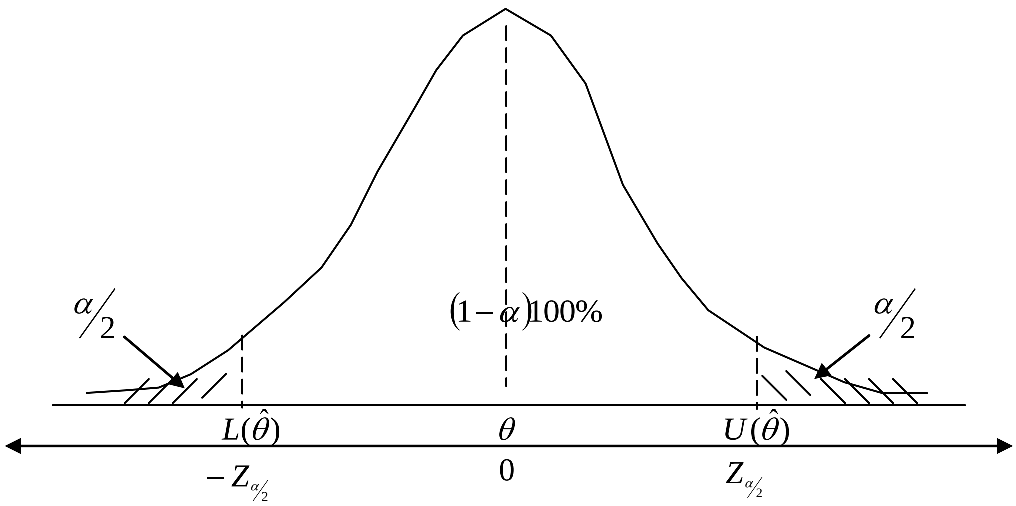
We consider a practical example to illustrate the construction of confidence intervals. Suppose we wish to estimate the mean yearly income of workers in an establishment. We randomly draw, say ${}^{N}C_{n} = 20$ = 20 samples each of size 20 observations and construct a confidence interval for the true mean yearly income, μ for each sample as shown in the diagram below.





The true mean income (μ) which is fixed is represented by the vertical line while the thick horizontal line segments represent the random intervals of the samples. We note that the random intervals vary from μ and that the intervals for samples 6 and 19 do not contain μ while the rest captures it. Hence the required interval estimator for μ is obtained with $\left(\frac{18}{20}\right)(100\%) = 90\%$ degree of confidence.

Suppose the sampling distribution of the estimator $\hat{\theta}$ for θ is normally distributed. Then the critical value, k takes the standard score, $Z_{\alpha/2}$ for large n.



The maximum error of estimation of the parameter θ, denote by E, is defined as:

$$E = Z_{\alpha/2}.Se(\bar{x}), \text{ for large } n \geq 30.$$

2-4.2 Confidence Intervals for Population Mean and Proportion

The $(1-\alpha)100\%$ confidence interval for the population (μ) may exist for two cases: *large sample size* and *small sample size*.

- When n is large $(n \geq 30)$ and σ^2 is known, the $(1-\alpha)100\%$ confidence interval for μ is given by

  $$\bar{x} \pm Z_{\alpha/2}.Se(\bar{x}) = \bar{x} \pm Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}},$$

  where σ is estimated by the sample standard deviation, s if it is unknown. The sample size, n required for estimating μ may be obtained as follows. The maximum error of estimation, $E = Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$, from which we obtain, $n = \frac{Z^2_{\alpha/2}.\sigma^2}{E^2}$.

- When n is small (n < 30) then the sampling distribution of $\bar{x}$ becomes the t-distribution and the confidence interval becomes

  $$\bar{x} \pm t_{\alpha/2(n-1)}.Se(\bar{x}) = \bar{x} \pm t_{\alpha/2(n-1)}.\frac{s}{\sqrt{n}},$$

  where $t_{\alpha/2(n-1)}$ is a critical value obtained from the t-distribution with degrees of freedom $(n-1)$.

The $(1-\alpha)100\%$ confidence interval for the population proportion (p) is given by

$$\hat{p} \pm Z_{\alpha/2}.Se(\hat{p}) = \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \approx \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $\hat{p} = \frac{x}{n}$ is the point estimator for p which should not be close to 0 or 1 and, $n \geq 30$ and both $n\hat{p}$ and $n\hat{p}(1-\hat{p})$ are ≥ 5. If these conditions are not met the interval estimate becomes unreliable and is not recommended to be used. The sample size needed to estimate p with a specified maximum error of estimation, E and confidence coefficient (1 - α) is obtained as follows:

$$E = Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \text{ from which we have } n = \frac{Z^2_{\alpha/2}\,\hat{p}(1-\hat{p})}{E^2}$$

If there is no prior knowledge of p, or $\hat{p}$, $\hat{p}(1-\hat{p})$ is replaced by its maximum value, 0.25 and the maximum possible error then becomes

$$E = Z_{\alpha/2}\sqrt{\frac{0.25}{n}}, \text{ from which we obtain } n = \frac{0.25\,Z^2_{\alpha/2}}{E^2}$$

<u>Illustrative Example 4.2:</u>

1.(a) A sample survey conducted in a city showed that 180 families spend on the average \$180.45 per week on food with a standard deviation of \$22.60. What can we say with 95% confidence about the maximum error if \$180.45 is weekly food expenditure of families used as an estimate of the actual average in the population? Obtain the interval estimate for this population average and comment on your result.

(b) The financial aid office of a University wishes to estimate the mean cost of recommended textbooks per semester for students. In order for the estimate to be useful, they want it to be within ¢8,500 of the true mean cost.

(i) How large a sample of semesters should be considered in order to be 99% confident of achieving this level of accuracy? Assume a standard deviation of ¢40,000.

(ii) Obtain a 95 interval estimate for the true mean cost if a random sample of size 40 with a mean of ¢80,000 and standard deviation ¢35,000.

(c) Analytical tests were performed to determine the impurity of a substance. A random sample of eight specimens was tested giving the following results.

1.4 1.8 1.5 1.3 2.0 1.7 1.6 2.1

(i) What is the point estimate of the mean impurity level of the substance?

(ii) Obtain a 90% confidence interval for the mean impurity level of the substance.

(iii) If the actual mean impurity level is estimated with a confidence coefficient of 95%, how many specimens must be used?

<u>*Solution:*</u>

(a) Given a confidence coefficient of $1-\alpha =$ 0.95, where $\alpha = 0.05$ and $Z_{\alpha/2} = 1.96$, we compute the maximum error of estimation,

$$E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (1.96)\frac{22.60}{\sqrt{180}} = 3.3$$

The actual population average is estimated within the interval estimate

$\bar{x} \pm E$ = 180.45 $\pm$ 3.30 or (177.15, 183.75).

(b) The error of estimation, $E = 8{,}500$, $\sigma = 40{,}000$ and $(1-\alpha)100\% =$ 99%.

(i) The required sample size is obtained as follows:

$$E = Z_{\alpha/2} . \frac{\sigma}{\sqrt{n}}$$, from which we have,

$$n = \frac{Z_{\alpha/2} . \sigma^2}{E^2} = \frac{(2.58)^2 \ (40{,}000)^2}{(1{,}500)^2} = 147.41 \approx 148,$$

which would be the required number of semesters needed for that level of accuracy.

(ii) A 95% confidence interval for true mean cost μ (in cedis) is

$$\bar{x} \pm z_{\alpha/2} . \frac{\sigma}{\sqrt{n}} = 80{,}000 \pm (1.96)\left(\frac{40{,}000}{\sqrt{40}}\right)$$

$$= 80{,}000 \pm 12{,}396.13$$

Hence $\mu \in [67{,}603.87, 92{,}396.13]$

(c) (i) The point estimate of the mean impurity level of the substance, it is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}(13.4) = 1.675 \approx 1.68$$

(ii) The sample standard deviation,

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]} = \sqrt{\frac{1}{7}\left[23 - \frac{1}{8}(13.4)^2\right]} = 0.282$$

A 90% confidence interval for μ a sample size $n = 8 < 30$ is

$$\bar{x} \ \pm \ t_{\alpha/2(n-1)} . \frac{s}{\sqrt{n}} = 1.675 \ \pm \ t_{0.05(7)} . \frac{0.282}{\sqrt{8}}$$

$$= 1.675 \ \pm \ (1.895).(0.0997)$$

$$= 1.675 \ \pm \ 0.1889$$

The required interval estimate for μ is (1.48, 1.86)

2.(a) To estimate the proportion of unemployed an economist selected 400 persons at random from large group of people in a community and found that 26 persons were unemployed.

(i) Estimate the true proportion of unemployed using a confidence coefficient of 0.90.

(ii) How many persons must be sampled to reduce the error of estimation to 0.012?

Solution:

(a) (i) The point estimate of proportion of unemployed (p) is

$$\hat{p} = \frac{25}{400} = 0.065$$

Then the required 90% confidence interval for p is

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.065 \pm (1.65)\sqrt{\frac{(0.065)\,(0.935)}{400}}$$

$= 0.065 \pm 0.0203$ or $p \in (0.045, 0.085)$

(ii) If the error of estimation is reduced to 0.012, the required sample size,

$$n = \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{E^2}$$

$$= \frac{(1.65)^2(0.065).(0.935)}{(0.012)^2}$$

$= 1{,}149.027 \approx 1{,}150$

2-4.3 Confidence Intervals for Difference between two Population Means and Proportions

Determining an interval estimate for the difference between two population means $(\mu_1 - \mu_2)$ or two population proportions $(p_1 - p_2)$ is a means of comparing two population parameters.

The $(1-\alpha)100\%$ *confidence interval for* $(\mu_1 - \mu_2)$, like for the single parameter μ, two case of sample sizes.

- If the sample sizes are large, we have

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2}.Se(\bar{x}_1 - \bar{x}_2) = (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $n_1 \geq 30$, $n_2 \geq 30$ and σ_1^2 *and* σ_2^2 are estimated using s_1^2 and s_2^2 respectively, if they are unknown. If $n_1 = n_2 = n$, then the error of estimation will be given by

$$E = Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}} = Z_{\alpha/2}\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}},$$ from which we have $n = \dfrac{Z^2_{\alpha/2}(\sigma_1^2 + \sigma_2^2)}{E^2}$

- If the sample sizes are small, the two population variances are assumed to be equal (*that is* $\sigma_1^2 = \sigma_2^2$) and the interval becomes

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}.(n_1 + n_2 - 1).Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $n_1 < 30$, $n_2 < 30$ and $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$, which is the estimate of the common variance σ^2 and the critical value, $t_{\alpha/2(n_1+n_2-2)}$ obtained from the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

The $(1-\alpha)100\%$ confidence interval estimator for $(p_1 - p_2)$ is given by

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}.\sqrt{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)},$$

where n_1 and n_2 are large and $\hat{p}_1 = \dfrac{x_1}{n_1}$, and $\hat{p}_2 = \dfrac{x_2}{n_2}$.

<u>Illustrative Example 4.3:</u>

1.(a) A mid-semester examination in Statistics was given 25 students randomly selected from class *A* and also to another 40 students randomly selected from class *B*. The mean scores obtained from both samples and the standard deviations are as shown below.

Class A	$n_A = 55$,	$\bar{x}_A = 66$	$s_A = 10$
Class B	$n_B = 40$,	$\bar{x}_B = 62$	$s_B = 8$

Construct a 95% confidence interval for the difference in the mean scores.

(b) A random sample of size, $n_1 = 10$ was selected from a population where $\bar{x}_1 = 13.9$ and $s_1 = 2.1$. Another random sample of size, $n_2 = 8$ was selected from a different population with $\bar{x}_2$ = 11.3 and $s_2 = 2.8$. If the difference between the two mean scores is to be estimated,

(i) find an estimate for the common variance of the populations, and

(ii) provide a 95% confidence interval for the difference.

<u>Solution:</u>

(a) The 95% confidence interval for the difference $(\mu_A - \mu_B)$, between the scores where both n_A and $n_B \geq 30$ is

$$\left(\bar{x}_A - \bar{x}_B\right) \pm Z_{\alpha/2}\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

$$= (66 - 62) \pm (1.96)\sqrt{\frac{10^2}{55} + \frac{8^2}{40}}$$

$= 4 \pm 1.96\ (1.85)$

$= 4 \pm 3.626$ or $(\mu_A - \mu_B) \in (0.374,\ 7.626)$

(b) (i) To estimate the difference $(\mu_1 - \mu_2)$ for small sample sizes we assume that the two population variances are equal which is estimated by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

$$= \frac{(10-1)(2.1)^2 + (8-1)(2.8)^2}{(10+8-2)} = 5.91 = (2.43)^2$$

(ii) The 95% confidence interval for $(\mu_1 - \mu_2)$,

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{\alpha/2(n_1+n_2-2)} . s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= (13.9 - 11.3) \pm t_{0.025(16)} . (2.43)\sqrt{\left(\frac{1}{10} + \frac{1}{8}\right)}$$

$= 2.6 \pm (2.120)(2.43)(0.473)$

$= 2.6 \pm 2.44$ or $(\mu_1 - \mu_2) \in (0.16,\ 5.04)$

2. Two independent random samples taken from two different populations produced the following results:

Sample 1	n_1 = 400,	$\hat{p}_1$ = 0.48
Sample 2	n_2 = 300,	$\hat{p}_2$ = 0.36

(i) What is the point estimate of the difference between the two population proportions, $(p_1 - p_2)$?

(ii) Develop a 99% confidence interval for the difference $(p_1 - p_2)$?

<u>*Solution*</u>

(i) The point estimate of $(p_1 - p_2) = 0.48\text{-}0.36 = 0.12$

(ii) The 99% confidence interval for $(p_1 - p_2)$ is

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(0.48 - 0.36) \pm Z_{0.05}\sqrt{\frac{(0.48)(0.25)}{400} + \frac{(0.36)(0.64)}{300}}$$

$$= 0.12 \pm (2.58)(0.03731)$$

$$= 0.12 \pm 0.0963$$

Hence the difference, $(p_1 - p_2) \in$ (0.024, 0.21).

Unit 4

THEORY OF ESTIMATION OF PARAMETERS

Introduction

Statistical inference means drawing conclusions based on data. There are a many contexts in which inference is desirable, and there are many approaches to performing inference. One context for inference is the parametric model, in which data are supposed to come from a certain distribution family, the members of which are distinguished by differing parameter values. The normal distribution family is one example. We shall examine these desirable properties of point estimators: unbiasedness, consistency, minimum variance unbiased estimators, efficiency, consistency and sufficiency.

Learning Objectives

After a careful study of this unit, students should be able to:

- State the probability distribution models and identify the number of parameters involved.
- Define the terminologies involved in point estimation
- Show that a point estimator is unbiased, consistent, sufficient, and efficient as well.
- State the Fisher and Neyman factorization theorem.

Unit content

Session 1-4:

1-4.1

1-4.2

Session 2-4:

2-4.1

2-4.2

2-4.3

SESSION 1-4: PRELIMINARY CONCEPTS

1-4.1Introduction

Estimation is the problem of determining the parameters or numerical characteristics of a population by using sample data. There are two types of estimation: Point Estimation and Interval Estimation. Point Estimation seeks an estimator that, based on the sample data, will give rise to a simple valued-estimate of an unknown parameter value, known as a point estimate. In interval Estimation on determines an interval that is likely to contain the unknown parameter value. Such an interval is called a confidence interval estimate.

Let $X_1, X_2, \ldots, X_n$ be a random sample selected from a population with probability distribution function $f(x,\theta)$ often the random sample $X_1, X_2, \ldots, X_n$ are randomly and independently obtained from the population under study with probability distribution function $f(x,\theta)$. If X is discrete, then $f(x,\theta)$ is known as a probability mass function (*pmf*), otherwise it said to be probability density function (*pdf*). If $X_1, X_2, \ldots, X_n$ are independently and identically distributed (iid) then each X_i will have the probability distribution, $f(x_i,\theta)$ with same mean and variance as the population distribution, $f(x,\theta)$.

The joint probability distribution of $X_1, X_2, \ldots, X_n$ is given by $f(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$

For a particular realization of the random sample we denote $x_1,\ldots,x_n$

1-4.2 TERMINOLOGIES AND NOTATIONS

<u>*Parameters*</u>

Let $f(x,\theta)$ be a probability distribution of the random variable X, where the constant(s), $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ and $k \geq 1$. The constants $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ are called parameters which are estimated by various methods which we shall discuss. For example,

- The Poisson distribution $f(x,\theta) = \dfrac{\theta^x e^{-\theta}}{x!}$, $x = 0, 1, 2, \ldots$

  has one parameter, θ.

- The Binomial distribution, $f(x,n,p) = \dbinom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, 2, \ldots, n$

  has two parameters, n and p.

- The Normal distribution $f(x,\mu,\sigma) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, $-\infty < x < \infty$

  has two parameters , μ and σ.

- The Gamma distribution $f(x,\alpha,\beta) = \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$, $x > 0$

  has two parameters α, β .

Parameter Space

The problem of point estimation is to choose a suitable value for θ from all possible values for θ. The set of all possible values for θ is called the <u>parameter space</u>, usually denoted by Θ. If θ can be positive values then, $\Theta = \{\theta : \theta > 0\}$

Parametric Models

$F = \{f(x;\ \theta):\ \theta \in \Theta\}$, $\theta \in \mathrm{R}^k$ is the parameter space. $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is the set of parameters.

Estimator

Suppose $(X_1, X_2, \ldots, X_n)$ is a random sample taken from a population X with pdf $f(x;\theta)$ where θ is an unknown parameter. An estimator of a parameter θ (or $T(\theta)$) is a function of the observations $(X_1, X_2, \ldots, X_n)$ and is used to estimate θ. So an estimateor is an algebraic formula.

Estimate

An estimate is the value we obtain when the observations from a sample are substituted into the formula for the estimator.

1-4.2.1 Statistic

A statistic $T_n = t(X_1, X_2, \ldots, X_n)$ is a function of the sample, $X_1, X_2, \ldots, X_n$ which is used to estimate the parameter θ. In this case we say, T_n is a point estimator of θ. For example, if $X_1, X_2, \ldots, X_n$ is a random sample drawn from a population distribution with mean μ and variance σ^2 then the statistics: sample mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and sample variance, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are point estimators for μ and σ respectively. The observed value of the statistic, $T_n = t(X_1, X_2, \ldots, X_n)$ is called an estimate of θ.

1-4.2.2 Likelihood Function and Robustness

Let $X_1, X_2, \ldots, X_n$ be a random sample selected independently selected from the population distribution $f(x,\theta)$. Then, the joint distribution of the random sample $X_1, X_2, \ldots, X_n$ and θ, denoted by $L(x_1, x_2, \ldots, x_n, \theta)$ is called the likelihood function.

That is, $L(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$

An estimator is said to be robust if its sampling distribution is not seriously affected by violations of assumptions

Self Assessment 1-4

1. Distinguish between an estimate and an estimator.
2. What is a family of distributions?

Answer tips

SECTION 2-4: PROPERTIES OF POINT ESTIMATORS

2-4.1 Introduction

The estimation of a parameter is done by using an appropriate statistic. A statistic used to estimate the parameter, θ is called *point estimator*, denoted $\hat{\theta}$. The parameters which are frequently estimated are the population mean (μ), proportion (P), variance or standard deviation $(\sigma^2 \text{ or } \sigma)$, the difference between two means $(\mu_1 - \mu_2)$ and proportions, $(\hat{p}_1 - \hat{p}_2)$. The statistics of these parameters are defined as follows:

(i) The sample mean, $\bar{x} = \frac{1}{n}\sum_{i=}^{n} x_i$, which is the point estimator for μ.

(ii) The sample proportion, $\hat{p} = \frac{x}{n}$, the point estimator for P.

(iii) The sample variance, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$, the point estimation for σ^2.

(iv) The sample differences, $(\bar{x}_1 - \bar{x}_2)$ and $(\hat{p}_1 - \hat{p}_2)$ are point estimators for $(\mu_1 - \mu_2)$ *and* $(p_1 - p_2)$ respectively.

These parameters characterize the population under study and are often identified with probability distributions. In order to decide which point estimator of a particular parameter is the best one to use, we need to examine their statistical properties and develop some criteria for comparing estimators and also give us the most economical information. Ideally, we want an estimator to generate estimates that can be expected to be close in value to the parameter. These properties are presented as follows:

2-4.2 Unbiasedness

The point estimator, T_n is said to be unbiased for the parameter θ if $E(T_n) = \theta$, for all $\theta \in \Theta$ Otherwise, we say T_n is biased estimator for θ. The difference, $E(T\ ) - \theta \neq 0$ as already noted is the *biased* of the estimator for T_n. For example, if $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution $f(x, \theta)$ with mean μ and variance σ^2, then the sample mean, $\bar{X}$ and sample variance, S^2 are unbiased estimators for μ and σ^2 respectively.

Example 2-4.1

Let $X_1, X_2, \ldots, X_n$ be random sample from *X*. Assume that

$E(X) = \mu \text{ and } Var(X) = \sigma^2$. Then show that:

(a) $\bar{X}$ is an unbiased estimator for μ

(b) $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Solution:

$$(a)\quad E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\left(n\mu\right) = \mu$$

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

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) = \sigma^2$$

If T_n is an estimator for θ, then the probability *biaz*$(T_n) = E(T_n) - \theta$ is called the bias of T_n. T_n is unbiased for θ iff *biaz*$\,\theta(T_n) = 0\ \forall\ \theta \in \Theta$.

If $\lim_{n\to 0}$ biaz $\theta(\mathrm{T_n}) = 0\ \forall\ \theta \in \Theta$, then we say that $\mathrm{T_n}$ is an *asymptotically unbiased* estimator for θ.

2-4.3 Mean Square Error (*MSE*)

One way of measuring the accuracy of an estimator T_n, is via its MSE. It is defined by

$$MSE(T_n) = E[(T_n - \theta)^2] = Var(T_n) + (E[T_n])^2 - 2\theta E[T_n] = Var(T_n) + (E[T_n] - \theta)^2,$$

where $(E[T_n] - \theta)^2$ is called biased.

Example 2-4.2

Let $X_1, X_2, ..., X_n$ be a random sample from some distribution such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, $i = 1, 2, \ldots, n$. Consider the following two statistics as the possible estimators of μ:

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad T_2 = \frac{1}{n+1}\sum_{i=1}^{n} X_i$$

Determine the mean squared errors of T_1 and T_2 and show that $MSE(T_2) < MSE(T_1)$ for some values of μ while the converse is true for other values of μ.

Solution

(i) $E(T_1) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = E(\bar{X}) = \mu$,

$MSE(T_1) = Var(T_1) = \sigma^2 / n$, since the biased, $E(T_1) - \mu = 0$

$E(T_2) = \frac{1}{n+1} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n+1}\sum_{i=1}^{n} E(X_i) = \left(\frac{n}{n+1}\right)\mu$

$Var(T_2) = \frac{1}{(n+1)^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{(n+1)^2}$

$MSE(T_2) = Var(T_2) + [\mu - E(T_2)]^2$

$= \frac{n\sigma^2}{(n+1)^2} + \left[\mu - \frac{n\mu}{(n+1)}\right]^2 = \frac{n\sigma^2 + \mu^2}{(n+1)^2}$

(ii) Let's say $n = 10$ and $\sigma^2 = 100$. Then

$MSE(T_1) = \frac{100}{10} = 10$; $MSE(T_2) = \frac{1000 + \mu^2}{121}$

For $MSE(T_1) = MSE(T_2)$ we have $\mu = \sqrt{210}$.

If $\mu < \sqrt{210}, MSE(T_2) < MSE(T_1)$ but if $\mu > \sqrt{210}$, then $MSE(T_1) < MSE(T_2)$.

2-4.4 Efficiency

Suppose that T_n' and T_n are unbiased estimators for the parameter θ. This indicates that the distribution of each estimator is centred at the true value of θ. However, If T_n has smaller variance it is more likely to produce an estimate close to the true value of θ. It is therefore said to be more efficient than T_n'. The relative efficiency of T_n with respect to T_n' is defined by $eff(T_n) = \frac{Var(T_n')}{Var(T_n)}$

A logical principal of estimation, when selecting among several estimator unbiased estimators, is to choose the estimator that has minimum variance.

If we consider all unbiased estimators for θ, the one with smallest variance is called the minimum unbiased estimator $(MVUE)$.

2-4.5 Consistency

Usually, a point estimate becomes closer to its parameter as the sample n become sufficiently large. That, is, we should have a better estimator for θ when it is based on, say, 50 observations than on 10 observations. Such a property is what is known as consistency.

Definitions

(i) Let $X_1, X_2, \ldots, X_n$ be a random sample and $T_n = t(X_1, X_2, \ldots, X_n)$ be an estimator for the parameter θ. Then T_n is said to be consistent if, $\lim_{n\to\infty} E(T_n) = \theta$ and $\lim_{n\to\infty} Var(T_n) = 0$.

(ii) An estimator T_n (precisely a sequence of estimators) is said to be (weekly) consistent if it converges to θ in probability. That is, if for every $\varepsilon > 0$, $\lim_{n\to\infty} P\left\{|T_n - \theta| > \varepsilon\right\} = 0$, $\forall\, \theta \in \Theta$

(iii) The estimator T_n for the parameter θ is also said to be consistent in MSE if $\lim_{n\to\infty} MSE(T_n) = 0$, $\lim_{n\to\infty} E\left[\left(T_n - \theta\right)^2\right] \to 0$, for every $\theta \in \Theta$.

Note that if an estimator is MSE consistent, it is weakly consistent. This follows from the Chebyshev's inequality: $P\left\{|T_n - \theta| > \varepsilon\right\} \le \dfrac{E[(T_n - \theta)^2]}{\varepsilon^2} = \dfrac{MSE(T_n)}{\varepsilon^2}$.

So, if $MSE(T_n) \to 0$ for $n \to \infty$ does $P\left\{|T_n - \theta| > \varepsilon\right\} \to 0$.

The above definitions refer to the asymptotic behaviour of T_n in which $\hat{\theta}$ may have undesirable properties for small n but becomes reasonable estimator in certain applications as $n \to \infty$.

2-4.6 Sufficient Statistic

The information contained in a random sample of size n are summarized by statistics such as $\bar{X}$ and s^2 to estimate μ and σ^2 respectively. We then ask the question: "Has this process of summarizing or reducing the data to the two statistics, $\bar{X}$ and S^2, retained all the information about μ and σ^2 in the original set of n observations? Or, has some information about these parameters been lost or obscured through the process of reducing the data?" The property of sufficiency provides methods for finding statistics that in a sense summarize all the information in a sample about a target parameter. Such statistics are said to be sufficient. Indeed, sufficient statistics often can be used to develop estimators that have the minimum variance among all unbiased estimators.

Definition

In making inference about an unknown parameter θ, the statistician makes a reduction of the data by using a statistic. Intuitively, the statistic should be such that "no information about θ is lost". $T_n = t(X_1, X_2, \ldots, X_n)$

Definition:

A statistic $T_n = t(X_1, X_2, \ldots, X_n)$ is called sufficient for a parameter θ if the conditional distribution of $X_1, \ldots, X_n$ given $T_n = c$, does not depend on θ, for all values of c.

$$f(x_1, x_2, \ldots, x_n | T = c) = \frac{f(x_1, x_2, \ldots, x_n, \theta)}{h(c, \theta)} = g(x_1, x_2, \ldots, x_n)$$

Example 2-4.3

for a random sample $X_1, \ldots, X_n$: drawn from $X \sim B(1, \theta)$. Show that $T_n = \sum X_i$ is a sufficient statistic

Proof

$$P(X_1 = x_1, \ldots, X_n = x_n | \sum_{i=1}^{n} X_i = c)$$

$$= \begin{cases} \dfrac{P(X_1 = x_1, \ldots, X_n)}{P(\sum_{i=1}^{n} X_i = c)} & if \sum x_i = c \\ 0 & otherwise \end{cases}$$

$$= \begin{cases} \dfrac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{c}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}} & if \ \sum x_i = c \\ 0 & otherwise \end{cases}$$

$$= \begin{cases} \dfrac{\theta^{c}(1-\theta)^{n-c}}{\binom{n}{c}\theta^{c}(1-\theta)^{n-c}} & if \ \sum x_i = c \\ 0 & otherwise \end{cases}$$

$$= \begin{cases} \dfrac{1}{\binom{n}{c}} & if \ \ x_1, \ldots, x_n = 0 \text{ or } 1 \text{ with } \sum x_i = c \\ 0 & otherwise \end{cases}$$

Which is independent of θ for all c.

Example 2-4.4

Show that $T_n = \sum_{i=1}^{n} X_i$ is a sufficient statistic for a random sample $X_1, \ldots, X_n$ from $X \sim Poisson(\theta), \ \ \theta > 0$.

Proof

$$P(X_1 = x_1, \ldots, X_n = x_n | \sum_{i=1}^{n} X_i = c)$$

$$= \begin{cases} \dfrac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(\sum_{i=1}^{n} X_i = c)} & if \ \sum x_i = c \\ 0 & otherwise \end{cases}$$

$$\Rightarrow P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} f(x;\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i} = \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i\,!}$$

$$\Rightarrow \ P(\sum_{i=1}^{n} X_i = c) = \frac{e^{-n\theta}(n\theta)^c}{c\,!}$$

$$= \begin{cases} \dfrac{\dfrac{e^{-n\theta}\theta^{\sum x_i}}{x_1\,!\ldots x_n\,!}}{\dfrac{e^{-n\theta}(n\theta)^c}{c\,!}} & if \ \sum_{i=1}^{n} x_i = c \\ 0 & otherwise \end{cases}$$

$$= \begin{cases} \dfrac{c\,!}{x_1\,!\ldots x_n\,!n^c} & if \ \sum x_i = c \\ 0 & otherwise \end{cases}$$

Which is independent of θ, $\forall c$.

2-4.7 Factorization theorem of Fisher and Neyman.

The statistic $T_n = t(X_1, X_2, \ldots, X_n)$ is sufficient for θ if and only if

$$f(x_1;\theta)f(x_2;\theta)\ldots f(x_n;\theta) = g(t(x_1,\ldots,x_n);\theta).h(x_1,\ldots,x_n)$$
$$or \ L(\theta;\underset{\sim}{x}) = g(t(x_1,\ldots,x_n);\theta).h(x_1,\ldots,x_n)$$

Where g is a non-negative function depending on θ and an $x_1, \ldots, x_n$ only through $t(x_1, \ldots, x_n)$ and h is a non-negative function, not depending on θ.

Example 2-4.5

Determine the sufficient statistic for a random sample $X_1, \ldots, X_n$ from $X \sim B(1,\theta)$, $0 < \theta < 1$

$$f(x;\theta) = \begin{cases} \theta^x (1-\theta)^{1-x} & if \ x = 0,1 \\ 0 & otherwise \end{cases}$$

Solution

$$\prod_{i=1}^{n} f(x;\theta) = g(t(x_1,\ldots,x_n);\theta)h(x_1,\ldots,x_n)$$

$$\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$= g(\sum x_i;\theta)h(x_1,\ldots,x_n)$$

$$g(\sum x_i;\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$L(x_1,\ldots,x_n) = \begin{cases} 1 & if \ x_1,\ldots,x_n = 0 \text{ or } 1 \\ 0 & otherwise \end{cases}$$

Example 2-4.6

A random sample $X_1, X_2, \ldots, X_n$ from $N(\mu, \sigma^2)$, then we have

$$f(x_1, x_2, \ldots, x_n, \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu\right)\right\}$$

which is a function of $\sum x_i$ and $\sum x_i^2$ which are the sufficient statistics.

Self Assessment 2-4

1. Show that $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})$ is an asymptotically unbiased estimator for σ^2, if

   E(x) = μ and Var(x) = σ^2.

2. Find the sufficient statistic for θ, a random sample from $X \sim N(\theta, \sigma^2)$ with σ^2 known.
3. What is a sufficient statistics?

Answer tips

1. Refer to session 2-4.2

Illustrative Examples

<u>Example 2-4.7</u>

Let $Y_{1,} Y_2, \ldots, Y_n$ denote a random sample from the uniform distribution on the interval $(0, \theta)$. Two unbiased estimators defined for θ are $T_1 = 2\frac{1}{n}\sum_{i=1}^{n} Y_i$ and $T_2 = \frac{n+1}{n} Y_n$ where $Y_n = \max(Y_{1,} Y_2, \ldots, Y_n)$. Find the efficiency T_1 of relative to T_2 .

Solution:

Given the random sample $Y_{1,} Y_2, \ldots, Y_n$ from the probability distribution,

$$f(y,\theta)=\begin{cases}\frac{1}{\theta}, 0<y<\theta\\ 0, elsewhere\end{cases}$$

$$E(Y)=\frac{\theta}{2} \text{ and } Var(Y)=\frac{\theta^2}{12}$$

We show that T_1 is unbiased estimator and find $Var(T_1)$:

(i) $$E\left(T_1\right)=E\left(2\frac{1}{n}\sum_{i=1}^{n}Y_i\right)=2\left(\frac{1}{n}\sum_{i=1}^{n}E(Y_i)\right)=2\left(\frac{n\theta}{2n}\right)=\theta$$

(ii) $$Var\left(T_1\right)=Var\left(2\frac{1}{n}\sum_{i=1}^{n}Y_i\right)=4Var\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)$$

$$=\frac{4}{n^2}\sum_{i=1}^{n}Var(Y_i)=\frac{4}{n^2}\left(\frac{n\theta^2}{12}\right)=\frac{\theta^2}{3n}$$

To show that T_2 is unbiased estimator and find $Var(T_2)$, we find $E(Y_n)$ and $Var(Y_n)$ by first determining the probability distribution of Y_n;

(i) $$f_{Y_n}(y)=n[E(Y)]^{n-1}f(y)$$

$$=n\left(y/\theta\right)^{n-1}\left(\frac{1}{\theta}\right)$$

$$=\begin{cases}\frac{ny^{n-1}}{\theta^n}, 0\le y\le\theta\\ 0 \quad , elsewhere\end{cases}$$

$$P(Y_n\le y)=F(y)=y/\theta, 0\le y\le\theta$$

(ii) $$E(Y_n)=\int_o^\theta y\frac{ny^{n-1}}{\theta^n}dy=\int_0^\theta\frac{ny^n}{\theta^n}dy=\frac{n}{(n+1)\theta^n}\, y^{n+1}\Big|_0^\theta=\frac{n\theta}{(n+1)}$$

$$\text{Hence, } E\left(T_2\right)=\left(\frac{n+1}{n}\right)E(Y_n)=\left(\frac{n+1}{n}\right)\frac{n\theta}{(n+1)}=\theta$$

(iii) $$E\left(Y_n^2\right)=\int_0^\theta y^2\frac{ny^{n-1}}{\theta^n}dy=\int_0^\theta\frac{ny^{n+1}}{\theta^n}dy=\frac{n}{n+2}\theta^2$$

$$Var(Y_n)=E\left(Y_n^2\right)-\left[E\left(Y_n\right)\right]^2$$

$$=\left(\frac{n}{n+2}\right)\theta^2-\left(\frac{n}{n+1}\right)^2\theta^2=\left(\frac{n}{n+2}-\frac{n^2}{(n+1)^2}\right)\theta^2=\frac{n\theta^2}{(n+1)^2(n+2)}$$

Hence, $Var\left(T_2\right)=\left(\frac{n+1}{n}\right)Var\left(Y_n\right)=\left(\frac{n+1}{n}\right)^2\frac{n\theta^2}{(n+1)^2(n+2)}=\frac{\theta^2}{n(n+2)}$

The relative efficiency of T_1 with respect to T_2

$$eff\left(T_1\right)=\frac{Var(T_2)}{Var(T_1)}=\frac{\frac{\theta^2}{n(n+2)}}{\frac{\theta^2}{3n}}=\frac{3}{n+2}$$

<u>Example 2-4.8</u>

Let $X_1, X_2, \ldots, X_n$ is a random sample selected from the Poisson probability distribution $f(x,\theta)$. We show that the statistic, $T_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ is unbiased, consistent.

$$f(x,\theta)=\frac{\theta^x e^{-\theta}}{x!},\ x=0,1,2,\ldots \text{ where } E(X)=\theta=Var(X)$$

(i) To show that $T_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ is unbiased:

$$E(T_n)=E\left(\frac{1}{n}\sum_{i=}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(X_i\right)=\frac{1}{n}\left[E(X_1)+E(X_2)+\ldots+E(X_n)\right]=\left(n\theta\right)=\theta$$

(ii) To show that $T_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ is consistent:

$$Var(T_n)=Var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}Var(X_i)$$

$$=\frac{1}{n^2}\left[Var(X_1)+Var(X_2)+\ldots+VarE(X_n)\right]=\frac{1}{n^2}\left(n\theta\right)=\frac{\theta}{n}$$

$$\lim_{n\to\infty}E(T_n)=\lim_{n\to\infty}\theta=\theta,\text{and }\lim_{n\to\infty}Var(T_n)=\lim_{n\to\infty}\frac{\theta}{n}=0$$

<u>Example 2-4.9</u>

(a) The number of breakdowns per week for a certain mini-computer is a random variable having the Poison distribution with parameter μ. A random sample on the number of breakdowns per week is made up of the observations $y_1, y_2, y_3, \ldots, y_n$.

(i) Suggest an unbiased estimator for μ.

(ii) The weekly cost of repairing these breakdowns is $C = 3y + y^2$ Show that $E(C) = 4\mu + \mu^2$ and find an unbiased estimator for $E(C)$ that makes use of the sample observations, $y_1, y_2, y_3, \ldots, y_n$.

<u>Solution:</u>

(i) The sample mean $(\bar{Y})$ is shown to be an unbiased estimator for μ as flows:

$$E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(y_i)$$

$$= \frac{1}{n}\left[E(Y_1) + E(Y_2) + E(Y_3) + \ldots + E(Y_n)\right]$$

$$= \frac{1}{n}(\mu + \mu + \ldots + \mu) = \frac{1}{n}(n\mu) = \mu$$

(ii) $E(C) = E(3Y + Y^2)$

$$= 3E(Y) + E(Y^2)$$

$$= 3E(Y) + Var(Y) + [E(Y)]^2$$

$$= 3\mu + \mu + \mu^2 = 4\mu + \mu^2$$

which is unbiasedly estimated by $\hat{E}(C) = 4\bar{y} + \bar{y}^2$

(b) Suppose Y_1, Y_2, Y_3, Y_4, Y_5 denote a random sample of some population with $E(Y_i) = \lambda$ and $Var(Y_i) = \lambda^2$, for all $i = 1, 2, 3, 4, 5$. The following statistics are suggested as estimators for λ.

(i) $T_1 = y_1$ (ii) $T_2 = \frac{1}{2}(Y_1 + Y_3)$

(iii) $T_3 = \frac{1}{2}(Y_2 + 2Y_3)$ (iv) $T_4 = \frac{1}{5}(Y_1 + Y_2 + Y_3 + Y_4 + Y_5) = \bar{Y}$

Determine the most efficient estimator among the unbiased estimators.

<u>Solution:</u>

Given that $E(Y_i) = \lambda$ and $Var(Y_i) = \lambda^2$, for all $i = 1, 2, 3, 4, 5$,

(a) We first find the unbiased estimators.

(i) $E(T_1) = E(Y_1) = \lambda$

(ii) $E(T_2) = \frac{1}{2}E(Y_1 + Y_3) = \frac{1}{2}\left[E(Y_1) + E(Y_3)\right] = \frac{1}{2}(\lambda + \lambda) = \frac{1}{2}(2\lambda) = \lambda$

(iii) $E(T_3) = \frac{1}{2}E(y_2 + 2y_3) = \frac{1}{2}\left[E(y_2) + 2E(y_3)\right] = \frac{1}{2}(\lambda + 2\lambda) = \frac{3}{2}\lambda \neq \lambda$

(iv) $E(T_4) = \frac{1}{5}E(Y_1 + Y_2 + Y_3 + Y_4 + Y_5)$

$$= \frac{1}{5}\left[E(Y_1) + E(Y_2) + E(Y_3) + E(Y_4) + E(Y_5)\right] = \frac{1}{5}\left(5\lambda\right) = \lambda$$

The unbiased estimators are $T_1,\ T_2\ and\ T_4$. The estimator T_3 is biased because $E(T_3)\ \neq \lambda$. It over estimates λ.

(b) We determine the minimum variance of the unbiased estimators.

(i) $Var(T_1) = Var(Y_1) = \lambda^2$

(ii) $Var(T_2) = Var\left[\frac{1}{2}(Y_2 + Y_3)\right]$

$$= \frac{1}{4}\left[Var(Y_2) + Var(Y_3)\right] = \frac{1}{4}\left(\lambda^2 + \lambda^2\right) = \frac{1}{2}\lambda^2$$

(iii) $Var(T_4) = Var\left[\frac{1}{5}\left(Y_1 + Y_2 + Y_3 + Y_4 + Y_5\right)\right]$

$$= \frac{1}{25}\left[Var(Y_1) + Var(Y_2) + \ldots + Var(Y_5)\right]$$

$$= \frac{1}{25}\left(5\lambda^2\right) = \frac{1}{5}\lambda^2$$

The minimum variance among the unbiased estimators is $Var(T_4)$. Hence T_4, the minimum variance unbiased estimator, is chosen the most efficient estimator.

Example 2-4.10

(a) Let suppose $X \sim B(n,\ \theta)$ and $Y = \frac{X+1}{n+2}$ be a propose estimator for θ. Then

$$E(Y) = \frac{1}{(n+2)}E(X+1) = \frac{1}{(n+2)}\left\{E(X)+1\right\} = \frac{1}{n+2}\left(\left\{n\theta+1\right\}\right)$$

$$Var(Y) = \frac{1}{(n+2)^2}Var(X+1) = \frac{1}{(n+2)^2}\left\{Var(X)\right\} = \frac{1}{(n+2)^2}n\theta(1-\theta)$$

$$\lim_{n\to\infty} E(Y) = \lim_{n\to\infty}\left\{\frac{n\theta+1}{n+2}\right\}$$

$$= \lim_{n\to\infty}\left\{\frac{\theta+1/n}{1+2/n}\right\}$$

$= \theta$ (y is asymptotically unbiased)

$$\lim_{n\to\infty} Var(Y) = \lim_{n\to\infty}\frac{1}{(n+2)^2}\left\{n\theta(1-\theta)\right\} = 0$$

Hence y is consistent for θ.

(b) Let $X_1, X_2, ..., X_n$ be a random sample from the Gamma distribution;

$$f(x,\alpha,\theta) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta}, x > 0 \ ,$$

where the value of the parameter α is known. Show that $\sum_{i=1}^{n} x_i$ is a sufficient statistic for the parameter θ.

Solution:

$$L(x_1, x_2, ..., x_n; \theta) = f(x_1,\theta) --- f(x_n,\theta)$$

$$= \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x_1^{\alpha-1} e^{-x_1/\theta} .... \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x_n^{\alpha-1} e^{-x_n/\theta}$$

$$= \frac{1}{(\Gamma(\alpha))^n \theta^{n\alpha}} \prod_{i=1}^{n} x^{\alpha-1} e^{-\sum_{i=1}^{n} x_i/\theta}$$

$$= \frac{1}{\theta^{n\alpha}} e^{-\sum_{i=1}^{n} x_i/\theta} \cdot \frac{\prod_{i=1}^{n} x^{\alpha-1}}{(\Gamma(\alpha))^n}$$

$$= h\left(\sum_{i=1}^{n} x_i, \theta\right) g(x_1, x_2, ..., x_n)$$

Hence by theorem, $\sum_{i=1}^{n} X_i$ is sufficient statistic for θ.

(b)Let $X_1, X_2, ..., X_n$ be a random sample from a Poisson distribution with the parameter λ. Show that the estimator for λ, $T = \sum_{i=1}^{n} X_i$ sufficient for λ

Solution:

The Poisson Probability Distribution,

$$f(x,\lambda)=\frac{\lambda^{x}e^{-\lambda}}{x!},\ x=0,1,----$$

The joint probability distribution function of $x_1, x_2, ..., x_n$ is

$$f\left(x_1,x_2,...,x_n,\lambda\right)=\prod_{i=1}^{n} f\left(x_i,\lambda\right)=\prod_{i=1}^{n}\left(\frac{\lambda^{x_i}e^{-\lambda}}{x_i\,!}\right)=\frac{\lambda^{\sum_{i=1}^{n}x_i}e^{-n\lambda}}{\prod_{i=1}^{n}x_i\,!}$$

$T_n=\sum_{i=1}^{n}X_i$ has also Poisson distribution with pdf $h(c,\lambda)=\dfrac{(n\lambda)^{c}e^{-nc}}{c!}$, $c=0,\ 1,\ 2,\ ...$

The conditional pdf of $x_1, x_2, ..., x_n$ given $T=t$ is given as

$$g\left(x_1,x_2,...,x_n\,/\,t,\lambda\right)=\frac{f\left(x_1,x_2,...,x_n\right),\lambda}{h(t,\lambda)}=\frac{\left(\dfrac{\lambda^{\sum_{i=1}^{n}x_i}e^{-n\lambda}}{\prod_{i=1}^{n}x_i\,!}\right)}{\left(\dfrac{(n\lambda)^{t}e^{-nt}}{t!}\right)}=\frac{t!}{n^{t}\prod_{i=1}^{n}x_i\,!},$$

where $t=\sum_{i=1}^{n}x_i$ which does not depend on λ and therefore it is sufficient for λ.

Learning Track Activities

Unit Summary

Key terms/ New Words in Unit

1. [State key term/New Words: explain what it means]
2. [State key term/New Words: explain what it means]
3. [State key term/New Words: explain what it means]
4. [State key term/New Words: explain what it means]

(choose an appropriate activity and give detail instructions)

- **Review Question:** [insert details of review question]
- **Discussion Question:** [insert details of discussion question]
- **Web Activity:** [insert details of web Activity or online v-classroom]
- **Reading:** [insert details of literature for reading]
- **Interactive CD:** [insert details of interactive CD]

Unit Assignments 4

1. **(a)** Show that statistic $H = \left(\frac{n+1}{n}\right) y_{\max}$ is an unbiased estimator for θ, where $y_{\max} = \max(y_{1,} y_2, ..., y_n)$.

(b) State and prove the Cramer-Rao lower bound theorem.

(c) Let s^2 denote the variance of random sample of size $n > 1$ from a distribution $N(\mu, \sigma), 0 \le \theta \le \infty$

(i) Find $E\left[\frac{ns^2}{n-1}\right]$

(ii) What is the efficiency of the statistics $\left(\frac{n}{n-1}\right) s^2$?

2. **(a)** Let $x_1, x_2, ..., x_n$ be a random sample from function,

$$f(x, \theta) = \frac{\lambda^x e^{-\theta}}{x!}, x = 0, 1, . . .$$

Show that the statistic $H = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ is efficient of θ.

(b) Let $x_1, x_2, ..., x_n$ be a random sample from the Bernoulli distribution;

$$f(x, p) = \begin{cases} p^x (1-p)^{1-x}; \ x = 0, 1; \ 0 \le p \le 1 \\ \quad 0 \quad , elsewhere \end{cases}$$

(c) Prove that the largest order statistic $y_{\max}$ is a sufficient statistic for the parameter of a uniform distribution $f(y,\theta)=\frac{1}{\theta}, 0 \le y \le \theta$.

3. **(a)** Show that $H=\overline{y}$ is sufficient statistic for θ in the exponential pdf

$$f(y,\theta)=\frac{1}{\theta}e^{-y/\theta}, 0 < y < \infty$$

(b) Let $x_1, x_2, ..., x_n$ be a random sample selected from the distribution,

$f(x,\theta)=\theta(1-\theta)^{x-1}, x=1,2,\ldots$ Show that $T=\sum_{i=1}^{n} x_i$ is a sufficient for θ.

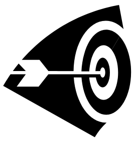
(c) Let $x_1, x_2, \ldots, x_n$ be a random sample from the Gamma distribution, $f(x,\alpha_0,\beta)$, where α_0 is a known constant. Show that $T=\sum_{i=1}^{n} x_i$ is sufficient for β.

Unit 5

METHODS OF ESTIMATION

Introduction

We shall discuss some useful methods of deriving estimators, most importantly the Maximum likelihood estimation and method of moment



Learning Objectives

After reading this unit you should be able to identify situations and model them using :

- Apply the methods of least squares and moments and use them to obtain estimates of parameters.
- Determine the maximum likelihood point and likelihood interval estimators of parameters.
- State the Cramer-Rao lower bound (CRLB) and apply it to obtain minimum variance unbiased estimator.
- Use the pivotal quantity method and that of the approximation method to obtain interval estimator of a parameter.

Unit content

2-5.4[add title of sub-session 2-5.4]

SESSION 1-5: METHODS OF POINT ESTIMATION

There are several methods used to estimate parameters of probability distributions. We shall discuss the following methods; method of moments, maximum likelihood method and the Least Squares method.

1-5.1 Method of Moments

Suppose a random variable X is known, to follow a distribution f(x| θ), where θ is unknown parameter , then the problem is to find an estimate of θ . In this method, we set the first population (distribution) moment equal to the first sample moment, then solve for θ . If the distribution has more than one unknown parameter θ , we set the second moment equal to the second sample moment and so on until we have a set of simultaneous equations, equal in number to the number of unknown parameters.

The procedure is as follows:

(i) Let $X_1, X_2, \ldots, X_n$ be random sample selected from the probability distribution $f(x,\theta)$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta$, $k \geq 1$ is/are parameter(s).

(ii) Obtain the first k moments about the origin $\mu_j(\theta)$ and it corresponding sample moments, m_j, defined by

$$\mu_j(\theta) = E(x^j) = \int_{R_x} x^j f(x,\theta)dx, \ \textit{if x is continuous}$$
$$= \sum_{\forall x} x^j f(x,\theta), \ \textit{if x is discrete}$$

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,j}, \text{ where } j = 1, 2, \ldots, k$$

(iii) Set up the equation(s),

$\mu_j = m_j$, where $j = 1, 2, \ldots, k$ and obtain the unique solution(s), $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$.

DEFINITION

r^th population moment $\Rightarrow \mu_r' = E(x^r)$

r^th sample moment $\Rightarrow m_r' = \dfrac{\sum x_i^r}{n}$

Example 1-5.1

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a negative exponential distribution. $f(x|\lambda) = \lambda e^{-\lambda x}$; $\lambda > 0$; $0 < x < \infty$.

   Find an estimator for λ by the method of moments.

Solution

$\mu_1' = m_1'$ . . . . (1)

First population moment $\mu_1' = E(X) = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$

First sample moment $m_1' = \sum \frac{X_i}{n} = \overline{X}$

From (1)

$\Rightarrow \frac{1}{\lambda} = \bar{x} \Rightarrow \hat{\lambda}_n = \frac{1}{\overline{X}} = \frac{n}{\sum X_i}$

2. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a uniform distribution on $[0, \theta]$, i.e

   $f(x|\theta) = \frac{1}{\theta}$; $0 \le x \le \theta$

   Find an estimate for θ by the method of moments.

Solution

$\mu_1' = E(x) = \int_0^\theta \frac{x}{\theta} dx = \frac{\theta}{2}$

$m_1' = \frac{\sum X_i}{n} = \overline{X}$

Thus equating first moments;

$\mu_1' = m_1'$; i. e $\frac{\theta}{2} = \overline{X} \Rightarrow \hat{\theta} = 2\overline{X}$

3. For example, if $X_1, X_2, \ldots, X_n$ denote a random and independent sample observations from the probability, $f(x,\theta) = (\theta+1)x^\theta$, $0 < x < 1$, we find the method of moments estimator for the parameter θ as follows:

   $$\mu_1 = E(x) = \int_0^1 x f(x,\theta) dx = \int_0^1 x(\theta+1)x^\theta dx = \int_0^1 (\theta+1)x^{\theta+1} dx = \frac{(\theta+1)}{(\theta+2)}$$

   $$m_1 = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$$

Now equating μ_1 to m_1 and solving for we have θ:

$$\frac{(\theta+1)}{(\theta+2)} = \bar{x}$$

$$(\theta+1) = \bar{x}(\theta+2) = \theta\bar{x} + 2\bar{x}$$

$$\theta(1-\bar{x}) = 2\bar{x} - 1$$

$$\hat{\theta}_n = \frac{2\bar{X} - 1}{(1-\bar{X})}$$

Given the sample observations: 0.25, 0.30, 0.40, 0.60, 0.25, we obtain estimate of $\hat{\theta}$ as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\left(1.80\right) = 0.36$$

$$\hat{\theta} = \frac{2\bar{x}-1}{(1-\bar{x})} = \frac{2(0.36)-1}{1-0.36} = -0.4375$$

1-5.2 Maximum Likelihood Method

This method often leads to estimators possessing desirable properties, particularly large sample properties. The idea is to use a value in the parameter space that corresponds to the largest "likelihood" for the observed data as an estimate of the unknown parameter. We explain the method as follows:

(i) Let $X_1, X_2, \ldots, X_n$ be an iid random sample from a probability distribution function $f(x,\theta)$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \Omega$.

(ii) Find the likelihood function $L(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$

(iii) The estimator $\hat{\theta}(x_1, x_2, \ldots, x_n)$ for θ is that value of θ which maximizes the likelihood function $L(x,\theta) = L(x_1, x_2, \ldots, x_n, \theta)$. The estimator $\hat{\theta}$ is called the maximum likelihood estimator (MLE). That is, MLE $\hat{\theta}$ maximizes $L(x,\theta)$ for each outcome $x_1, x_2, \ldots, x_n$.

(iv) To determine MLE, $\hat{\theta}$ the for which $L(x_1, x_2, \ldots, x_n, \theta)$ is maximized, we rather maximize $InL(x,\theta) = InL(x_1, x_2, \ldots, x_n, \theta)$ for computational convenience:

$$\frac{\partial InL(x,\theta)}{\partial\theta} = \frac{\partial InL(x_1, x_2, \ldots, x_n, \theta)}{\partial\theta} = \ell(x_1, x_2, \ldots, x_n, \theta) = 0$$

and solve $\hat{\theta}(x_1, x_2, \ldots, x_n)$.

DEFINITION

A value $\hat{\theta}_n$, say $\hat{\theta}_n = t(x_1,...,x_n)$ which maximizes the likelihood function, $L(\theta, x_1,...,x_n)$ over $\forall\theta \in \Theta$ is called a maximum likelihood estimate for θ. Hence $\forall\theta \in \Theta$

$L(t(x_1,...,x_n), \underset{\sim}{x}) \geq L(\theta; x_1,...,x_n) \to \hat{\theta}_n$ is the ML – estimate, so, the random variable $T_n = (x_1, ..., x_n)$ is called a maximum likelihood estimator for θ.

REMARK

a) ML – estimator is not necessarily unique
b) ML – estimator is not necessarily unbiased. Asymptotically normal
c) Since L is a product, it is more convenient to maximize In(L).

Define the likelihood for a random sample $X_1,...,X_n$ from a negative exponential distribution

$\theta > 0$

$$f(x;\theta) = \theta e^{-\theta x}; (x > 0)$$
$$L(\theta; \underset{\sim}{x}) = \theta^n e^{-\theta \sum x_i}; (\forall x_i > 0)$$

DEFINITION

The function $l(\theta; \underset{\sim}{x}) = l(\theta; x_1,...,x_n) = \ln L(\theta; x_1,...,x_n)$ is called the <u>Log Likelihood</u> function of $X_1, ..., X_n$. That is taking the natural log of the likelihood function.

DEFINITION

The function $s(\theta; \underset{\sim}{x}) = s(\theta; x_1,...,x_n) = \dfrac{\partial}{\partial\theta} l(\theta; \underset{\sim}{x})$ is called the <u>Score function</u> of $X_1, ..., X_n$.

That is the first partial derivative with respect to the log-likelihood function.

DEFINITION

The function $I(\theta; \underset{\sim}{x}) = I(\theta; x_1,...,x_n) = -\dfrac{\partial}{\partial\theta} s(\theta; x) = -\dfrac{\partial^2}{\partial\theta^2} l(\theta; \underset{\sim}{x})$ is called the <u>information function</u> of $X_1, ..., X_n$.

So, in many cases, $\hat{\theta}_n$ can be found by solving the <u>maximum likelihood equation</u>: $s(\theta; \underset{\sim}{x}) = 0$ and by checking that $I(\hat{\theta}_n) > 0$.

Example 1-5.2

(a) For example, we find the maximum likelihood estimator for the Poisson parameter θ given the random sample, $X_1, X_2, \ldots, X_n$ as:

$$f(x,\theta) = \frac{\theta^x e^{-\theta}}{x!},\ x = 0, 1, 2, \ldots$$

$$L(x,\theta) = L(x_1, x_2, \ldots, x_n, \theta)$$

$$= \prod_{i=1}^{n} f(x_i,\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i\,!} = \frac{\theta^{\sum x_i} e^{-n\theta}}{\prod_{i=1}^{n} x_i\,!}$$

$$InL(x,\theta) = \left(\sum_{i=1}^{n} x_i\right) In\theta - n\theta - \sum_{i=1}^{n} Inx!$$

$$\frac{\partial InL(x,\theta)}{\partial\theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - n = 0$$

$\sum_{i=1}^{n} x_i = n\theta$, from which we have $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$

(b) Find the ML – estimator for θ.

$X_1, \ldots, X_n$: random sample form X

X: Bernoulli, parameter θ, θ]0,1[

$$f(x,\theta) = \theta^x (1-\theta)^{1-x},\ x = 0, 1$$

Then we define the log-likelihood function;

$$l(\theta; \underset{\sim}{x}) = \sum x_i \ln\theta + (n - \sum x_i)\ln(1-\theta)$$

$$s(\theta; \underset{\sim}{x}) = \frac{\sum x_i}{\theta} - \frac{(n - \sum x_i)}{1-\theta}$$

$$I(\theta; \underset{\sim}{x}) = \frac{\sum x_i}{\theta^2} - \frac{n - \sum x_i}{(1-\theta)^2}$$

So solve $s(\theta; \underset{\sim}{x}) = 0$ and check $I(\theta; \underset{\sim}{x}) > 0$,

$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the ML- estimate for θ, then

$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ is the ML- estimate for θ.

(c) Find the ML estimator for θ.

$X_1, \ldots, X_n$: random sample from X

X: exponential with parameter $\theta \in]0,\infty[$ or $\theta > 0$ or $0 < \theta < \infty$

Solution

$f(x,\theta) = \theta e^{-\theta x},\ x > 0, \theta > 0$

$l(\theta; \underset{\sim}{x}) = n \ln \theta - \theta \sum x_i$

$s(\theta; \underset{\sim}{x}) = \frac{n}{\sigma^2} - \sum x_i$

$I(\theta; \underset{\sim}{x}) = \frac{n}{\sigma^2} > 0$

It follows that $\hat{\theta}_n = \frac{n}{\sum x_i}$ is the ML – estimate for θ and that

$T_n = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}$ is the ML – estimator for θ.

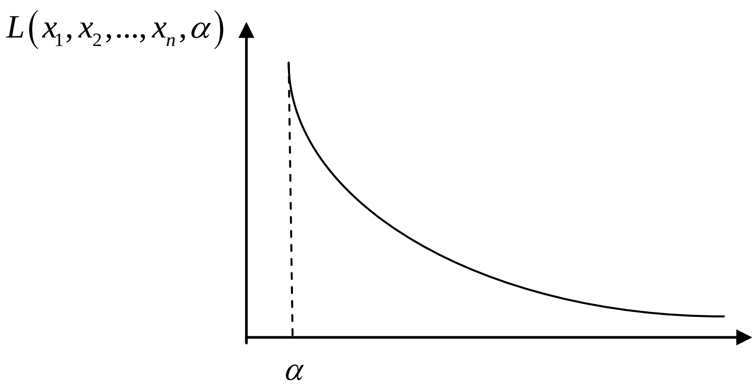
Remark: The procedure for obtaining MLE as outlined above does not always work. This is illustrated by the example: Suppose that the random variable X is uniformly distributed over the interval $[0, \alpha]$, where α is an unknown parameter. The pdf of X is

$$f(x, \alpha) = \begin{cases} \frac{1}{\alpha}, 0 \le x \le \alpha \\ 0, elsewhere \end{cases}$$

If $X_1, X_2, ..., X_n$ is a random sample from X, then the likelihood function

$L(x_1, x_2, ..., x_n, \alpha) = \left(\frac{1}{\alpha}\right)^n, 0 \le x_i \le \alpha,$ for all i, 0, *elsewhere.*

Thus, for $L(x_1, x_2, ..., x_n, \alpha) > 0, x_i \le \alpha$ or $\alpha \ge \max(x_1, x_2, ..., x_n)$. To find the value of α that maximizes $InL(x_1, x_2, ..., x_n, \alpha) = \ell(x_1, x_2, ..., x_n, \alpha)$ we obtain the graph as below



That is $\hat{\alpha} = \max(x_1, x_2, ..., x_n)$.

1-5.3 Theorem: Cramer Rao Lower Bound (CRLB) Inequality

DEFINITION

Under regularity conditions:

i. $E\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)=0$

ii. $E\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2=-E\left(\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right)$

The quantity:

$i(\theta)=E\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right]$ is called Fisher information number

CRAMER-RAO INEQUALITY or CRAMER-RAO LOWER BOUND (CRLB)

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x;\theta),\theta\in\Theta\subset R$. Let T_n be any unbiased for θ. Then under regularity conditions, $Var(T_n)\geq\frac{1}{ni(\theta)};\forall\theta\in\theta$. Where $i(\theta)$ is the fisher information number.

NB:

The following are the consequences of the CRLB inequality:

(i) An unbiased estimator T is said to efficient if it attains the CRLB. That is, $Var(T_n)=CRLB=\left[ni(\theta)\right]^{-1}$

(ii) For any unbiased estimator T, its efficiency is defined by $eff(T)=\frac{CRLB}{Var(T)}$

(iii) Often $T_n=t(X_1,X_2,\ldots,X_n)$ is efficient at $n\to\infty$. Specifically, T_n is said to be asymptotically efficient.

Example

a) Find the Cramer-Rao lower bound for $X\sim poisson(\theta).\theta>0$
b) Find the Cramer-Rao lower bound for $X\sim Bernoulli(\theta);0<\theta<1$

1-5.4 Definition: Relative Likelihood Functions (RLF)

Let the likelihood function be defined as $L(x,\theta)=kP(E,\theta)$, the probability of observing the event E, and θ_1 and θ_2 be two possible values of θ. Then the relative likelihood of θ_1 with respect to θ_2 is given by

$$\frac{L_1(x,\theta_1)}{L_2(x,\theta_2)}=\frac{kp(E,\theta_1)}{kp(E,\theta_2)}=\frac{probability\ of\ \ data\ \ for\ \ \theta=\theta_1}{probability\ of\ \ data\ \ for\ \ \theta=\theta_2}$$

Thus, RLF of θ is defined by $R(\theta)=\dfrac{L(x,\theta)}{L(x,\hat{\theta})}$, where $\hat{\theta}$ is the estimate of θ.

If θ_1 is a particular value of θ then $R(\theta_1)=\dfrac{L_1(x,\theta_1)}{L_2(x,\theta_2)}$, where $0\le R(\theta_1)\le 1$. If $R(\theta_1)$ is small such that $R(\theta_1)<0.1(10\%)$, θ_1 is said not to be a plausible value of the parameter, θ. That is $R(\theta_1)=0.1$ means there is a value of θ_1 which is 10% times more probable than θ_1. If however $R(\theta_1)$ is large such that $R(\theta_1)\ge 0.5(50\%)$, θ_1 is said to be fairly plausible value of θ because it gives the data at least 50% of the maximum probability which is possible under the model.

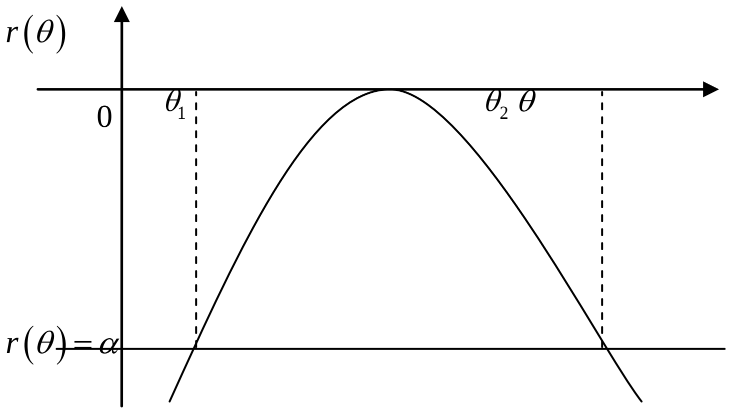
The RLF ranks all possible parameter values according to their plausibility in the light of the data. These values are contained in a likelihood region or interval. $R(\theta)=\alpha$ is called the $100\alpha\%$ likelihood interval or region for θ.

Let $r(\theta)=\ln R(\theta)=\ln L(x,\theta)-\ln L(x,\hat{\theta})=\ell(x,\theta)-\ell(x,\hat{\theta})$. Then $-\infty<r(\theta)<0$.

That is, when,

- $R(\theta)=0.1 \Rightarrow r(\theta)=-2.303-$ *plausible*
- $R(\theta)=0.5 \Rightarrow r(\theta)=-0.693-$very plausible
- $R(\theta)=0.01 \Rightarrow r(\theta)=-4.605-$ not plausible

To determine the likelihood interval for particular values of θ , we plot a graph of $r(\theta)$ against and the horizontal line $r(\theta)=\alpha$, obtaining the interval (θ_1,θ_2).



That is $r(\theta)=-0.693\big(i.eR(\theta)=0.5\big)$. Hence (θ_1,θ_2) is the $100\alpha\%=50\%$ likelihood interval for θ.

As an illustrative example: The lifetime of an electronic component has an exponential distribution; $f(x,\theta)=\frac{1}{\theta}e^{-x/\theta}, x\geq 0$ where θ is the expected time to the component. If 10 of such components is are tested independently and their lifetime measured to the nearest day as given as follows: 70, 11, 66, 5, 20, 4, 35, 40, 29, and 8. Determine the values of θthat are plausible in the light of the given data.

(i) The likelihood function,

$$L(x_1,x_2,...,x_n,\theta)=\prod_{i=1}^{n}\left(\frac{1}{\theta}e^{-\frac{x}{\theta}}\right)=\frac{1}{\theta^n}e^{-\frac{1}{\theta}\sum_{i=1}^{n}x_i}$$

(ii) The maximum likelihood estimator,

$$\ln L(x_1,x_2,...,x_n,\theta)=-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i$$

$$\frac{\partial}{\partial}\ln L(x_1,x_2,...,x_n,\theta)=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i$$

$$-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i=0 \quad \Leftrightarrow \hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}$$

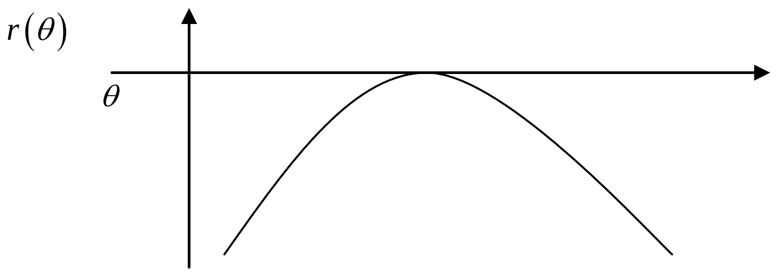
(iii) From the given data we estimate, $\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{10}(288)=28.8$

(iv) The relative likelihood function (RLF),

$$r(\theta)=\ln L(x,\theta)-\ln\left(x,\hat{\theta}\right)=\left(-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i\right)-\left(-n\ln\hat{\theta}-\frac{1}{\hat{\theta}}\sum_{i=1}^{n}x_i\right)$$

$$=-10\ln\theta-\frac{288}{\theta}+10\ln 28.8+10=-10\ln\theta-\frac{288}{\theta}-43.6$$

Plotting $r(\theta)$ against various values of θ, we obtain a curve of the form shown below.



From which we can estimate 10% likelihood interval for θ

1-5.5 The Least Squares Method

In this method we will consider a simple linear regression model, $y = \alpha + \beta x + \varepsilon,$ which involves two parameters α and $\beta.$ The dependent variable y is called the response, x the independent variable is also called predictor and $\varepsilon,$ the random error term. Now, given the paired sample observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$ where the $y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, 2, \ldots, n,$ the parameters α and β are estimated based upon the following assumptions:

(i) $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2,$ a constant, for all x_i

(ii) The probability distribution of ε_i is $\varepsilon_i \sim N(0, \sigma^2)$

(iii) The random error ε_i and ε_j are independent. That is, $C\text{ov}(\varepsilon_i, \varepsilon_j) = 0,\ \textit{for } i \neq j$

(iv) The probability distribution of y_i is $y_i \sim N\left(\alpha + \beta x_i,\ \sigma^2\right)$

The least squares method find the estimates $\hat{\alpha}$ and $\hat{\beta}$ of α and β respectively such the sum of the squares of the errors (SSE) is minimized. The sum of squares of the errors about estimated regression line is given by

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2$$

Differentiating SSE with respect to $\hat{\beta}_0$ and β_1 we obtain

$$\frac{\partial SSE}{\partial \alpha} = -2\sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right) = 0$$
$$\frac{\partial SSE}{\partial \beta} = -2\sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right) x_i = 0$$

Setting the partial derivatives to zero and using the rules of summation, we obtain the equations.

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

These equations are called the normal equations. Solving we obtain $\hat{\alpha}$ and $\hat{\beta}$, being the least squares estimators for α and β respectively as given below:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}\frac{1}{n}\sum_{i=1}^{n} x_i = \overline{y} - \hat{\beta}\overline{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{SS_{xy}}{SS_{xx}}$$

where $SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$ and $SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$

Hence, the least squares estimated linear regression model is

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i, \; i = 1, 2, \ldots, n,$$

For example, we use the data points given below to estimate the parameters β_0, and β_1.

X	95	82	90	81	99	100	93	95	93	87
Y	214	152	156	129	254	266	210	204	213	150

where $\sum x_i = 915; \sum y_i = 1,948; \sum x_i^2 = 841,03;$ $\sum y_i^2 = 379,054$ and $\sum x_i y_i = 180798$

(i) The sample means: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{915}{10} = 91.5;$ $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1948}{10} = 194.8$

(ii) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} == \frac{180798 - \frac{1}{10}(915)(1948)}{84103 - \frac{1}{10}(915)^2} = \frac{2556}{3805} = 6.7175$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 194.8 - 6.7175(91.5) = -419.5$$

Hence, $\hat{y}_i = -419.5 + 67175 x_i$

(iii) $$\sigma^2 = s^2 = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \frac{1}{8}\left(2093.43\right) = 261.68$$

Illustrative Examples

Example 1-5.4

(a) Suppose a random sample of size n from random variable X with Gamma distribution,

$$f\left(x,\alpha,\beta\right)=\frac{1}{\Gamma\left(\alpha\right)\beta^{\alpha}}x^{\alpha-1}e^{-x/\beta},x>0,\alpha>0,\beta>0$$

$$E\left(X\right)=\mu_1^{'}=\alpha\beta;\ E\left(X^2\right)=\mu_2^{'}=\alpha\left(\alpha+1\right)\beta^2$$

The first two sample moments;

$$m_1=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X};\ m_2=\frac{1}{n}\sum_{i=1}^{n}X_i^2$$

Hence we have the equations,

$$\alpha\beta=\bar{x} \qquad ....(1)$$

$$\alpha\left(\alpha+1\right)\beta^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2 \qquad ....(2)$$

From which we obtain the following:

$$\left(\alpha+1\right)\bar{X}^2=\frac{\alpha}{n}\sum_{i=1}^{n}X_i^2,\ where\ \ \beta=\frac{\bar{X}}{\alpha}$$

$$\hat{\alpha}=\frac{n\bar{x}^2}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=\frac{n\bar{x}^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\ \text{and}\ \hat{\beta}=\frac{\bar{x}}{\alpha}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n\bar{x}}$$

(b) A lifetime random variable t has a Gamma distribution $Gm(40,\theta)$. If a random sample of 10 lifetimes resulted in a mean of 75.3, we calculate the maximum likelihood estimate of θ based on these observations as follows:

(i) We write out the Gamma distribution,

$$f(t,\theta)=\frac{1}{\Gamma(40)\theta^{40}}t^{39}e^{-\frac{t}{\theta}},\ t>0$$

(ii) The likelihood function,

$$L(t_1,t_2,..,t_n,\theta)=\prod_{i=1}^{n}\left(f(t_i,\theta)\right)$$

$$=\frac{1}{\left(\Gamma(40)\right)^n\theta^{40n}}\prod_{i=1}^{n}t_i^{39}e^{-\frac{1}{\theta}\sum_{i=1}^{n}t_i},\ t>0$$

(iii) To find the maximum likelihood estimator (MLE), $\hat{\theta}$:

$$InL(t,\ \theta)=39\sum_{i=1}^{n}Int_i-\frac{1}{\theta}\sum_{i=1}^{n}t_i-40nIn\theta-nIn\Gamma(40)$$

$$\frac{\partial InL(x,\theta)}{\partial\theta} = \frac{1}{\theta^2}\sum_{i=1}^{n} t_i - \frac{40}{\theta} = 0, \text{from which we obtain}$$

$$\hat{\theta} = \frac{1}{40n}\sum_{i=1}^{n} t_i$$

(iv) We compute the estimate of the MLE for θ using the given sample data:

$$\hat{\theta} = \frac{1}{40n}\sum_{i=1}^{n} t_i = \frac{1}{40(10)}(10)(75.3) = 1.8825$$

(v) The relative likelihood function (RLF),

$$r(\theta) = InL(t,\theta) - InL(t,\hat{\theta})$$

$$= \left(39\sum_{i=1}^{n} Int_i - \frac{1}{\theta}\sum_{i=1}^{n} t_i - 40nIn\theta - nIn\Gamma(40)\right) - \left(39\sum_{i=1}^{n} Int_i - \frac{1}{\hat{\theta}}\sum_{i=1}^{n} t_i - 40nIn\hat{\theta} - nIn\Gamma(40)\right)$$

$$= 400(1 + In1.8825) - 400In\theta - \frac{753}{\theta}$$

Example 1-5.5

Let the time to failure of an electrical generator has life length α whose pdf is the Gamma distribution,

$$f(x,\lambda,\gamma) = \frac{\lambda^{\gamma} x^{\gamma-1} e^{-\lambda x}}{\Gamma(\gamma)}, x \geq 0$$

$$L(x_1, x_2, ..., x_n, \lambda, \gamma) = \frac{\lambda^{n\gamma}\left(\prod_{i=1}^{n} x_i\right)^{\gamma-1} e^{-\lambda\sum_{i=1}^{n} x_i}}{\left[\Gamma(\gamma)\right]^n}$$

$$\ln L(x_1, x_2, ..., x_n, \lambda, \gamma) = n\gamma\ln\lambda + (\gamma-1)\sum_{i=1}^{n}\ln x_i - \lambda\sum_{i=1}^{n} x_i - n\ln\Gamma(\gamma)$$

$$\frac{\partial}{\partial\lambda}\ln L(x_1, x_2, ..., x_n, \lambda, \gamma) = \frac{n\gamma}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\frac{\partial}{\partial\gamma}\ln L(x_1, x_2, ..., x_n, \lambda, \gamma) = n\ln\lambda + \sum_{i=1}^{n}\ln x_i - \frac{\Gamma'(\gamma)}{\Gamma(\gamma)} = 0$$

from which we have $\hat{\lambda} = \frac{\gamma}{\bar{x}}$. $\ln\gamma = \frac{\Gamma'(\gamma)}{\Gamma(\gamma)} = \ln\bar{x} - \frac{1}{n}\sum_{i=1}^{n}\ln x_i$

SECTION 2-5: INTERVAL ESTIMATION METHODS

2-5.1 Introduction

Sometimes, interval estimators are more preferred because point estimators provide no indication of inherent precision despite their desirable properties presented in section 2-2. The unknown parameters may accurately be estimated to lie within an interval with certain degree of confidence. Point estimators do not give such degree of confidence

Definition

(a) *Confidence Interval*: Let $X_1, X_2, ..., X_n$ be a random sample from the pdf, $f(x,\theta)$. Suppose that $L_1(x_1, x_2, ..., x_n; \theta)$ and $L_2(x_1, x_2, ..., x_n; \theta)$ are two statistic such that $L_1 \le L_2$. The random interval (L_1, L_2) is called $(1-\alpha)100\%$ for θ. L_1 and L_2 can be selected such that $p(L_1 \le \theta \le L_2) = 1-\alpha$ where L_1 and L_2 are called the lower and upper confidence limits respectively.

CONFIDENCE REGIONS

The probability that the random set contains θ is given by:

$P_\theta(\theta \in D(X_1, ..., X_n)) = 1-\alpha,\ \forall \theta \in \Theta \subset \Re$. Where $D(X_1, \ldots, X_n)$ is a subset of θ.

Such a region $D(X_1, \ldots, X_n)$ is called a $100(1-\alpha)\%$ confidence region for θ.

CONFIDENCE INTERVALS

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x;\theta),\ \theta \in \Theta \subset R$. Let $\alpha \in]0,1[$

If $L_n = l(X_1, \ldots, X_n)$ and $R_n = r(X_1, \ldots, X_n)$ are two statistics satisfying

$P_\theta(L_n \le \theta \le R_n) = 1-\alpha\ ,\forall \theta \in \Theta$, then, the random interval $[L_n,R_n]$ is called a $100(1-\alpha)\%$ interval estimator for θ. For observations $x_1, \ldots, x_n$, the corresponding interval estimate for θ $[\, l(x_1, \ldots, x_n), r(x_1, \ldots, x_n)]$ is called a $100(1-\alpha)\%$ confidence interval (C.I) for θ.

An important practical method for finding a confidence interval is the so called Pivotal method. The method has 2 steps:

1. Find a pivotal quantity: i.e $t(X_1, ..., X_n; \theta)$
2. Try to invert ("pivot") the inequality, $a \le t(X_1, ..., X_n; \theta) \le b$ as an inequality of the form $l(x_1, ..., x_n) \le \theta \le r(x_1, ..., x_n)$, $\forall$ values of $x_1, ..., x_n$

Pivotal-Quantity: Let $X_1, X_2, ..., X_n$ be a random sample from the pdf $f(x,\theta)$ and $T = t(X_1, ..., X_n, \theta)$ a function of $X_1, X_2, ..., X_n$ and θ. The statistic T is called pivotal-quantity if its distribution does not depend on θ.

2-5.1.1 Construction of Confidence Intervals

Pivotal-Quantity Method

Let $T = t(x,\theta)$ be a pivotal quantity which is monotonic function of θ for each fixed X. Then we find t_1 and t_2 such that $P\left(t_1 \le T(x,\theta) \le t_2\right) = 1-\alpha$, which can be inverted to the form: $p\left(L_1 \le \theta \le L_2\right) = 1-\alpha$, where L_1 and L_2 are functions of pof the random sample, $x_1, x_2, ..., x_n$.

Suppose that the life span, X of n items has the normal distribution, $N\left(\mu,\sigma^2\right)$ where σ^2 is known. Consider the pivotal quantity

$$T = t\left(x_1, x_2, ..., x_n, \mu\right) = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

That is, t has the distribution $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}, -\infty < t < \infty$, which is independent of μ.

Let $\phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$. That is,

$$P\left(t_1 \le T\left(x_1, x_2, ..., x_n\right) \le t_2\right) = P\left(t_1 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le t_2\right) = 1-\alpha$$

$$= P\left(\bar{x} - t_2.\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} - t_1.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$= \int_{t_1}^{t_2} f(t)\,dt = \phi(t_2) - \phi(t_1) = 1-\alpha$$

The length of random confidence interval,

$$L = \frac{\sigma}{\sqrt{n}}\left(t_2 - t_1\right)$$

To determine t_1 and t_2 we minimize the random confidence interval, $L = \frac{\sigma}{\sqrt{n}}\left(t_2 - t_1\right)$ subject to $\phi(t_2) - \phi(t_1) = 1-\alpha$ as follows:

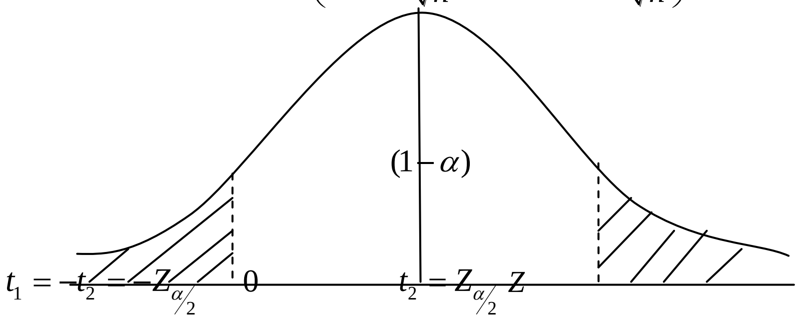
$$\frac{dL}{dt_1} = \frac{\sigma}{\sqrt{n}}\left(\frac{dt_2}{dt_1} - 1\right) = 0$$

$$\phi(t_2)\frac{dt_2}{dt_1} - \phi(t_1) = 0$$

$$\Rightarrow \frac{dt_2}{dt_1} = \frac{\phi(t_1)}{\phi(t_2)}, \text{from which we have:}$$

$$\frac{\phi(t_1)}{\phi(t_2)} - 1 = 0 \iff \phi(t_2) = \phi(t_1)$$

Hence $t_1 = t_2$, which is admissible since $\int_{t_1}^{t_2} f(t)\,dt = 0 \neq 1-\alpha$ and so we take we take $t_1 = -t_2$ and we have, $P\left(\bar{x} - t_2 . \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t_2 . \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$



Example 2-5.2
Confidence Intervals for parameters of a normal distribution
Let $X_1, \ldots, X_n$ be a random sample of X with $X \sim N(\mu, \sigma^2)$

a) Construct a C.I for μ if σ^2 is known.

Solution

An estimator for μ is the ML-estimator $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ by the central limit theorem

$$\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

Hence for any a:

$$P\left(-a \le \frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \le a \right) = 1-\alpha = \Phi(a)-\Phi(-a)$$

now

$$P\left(\bar{x} - a\sqrt{\frac{\sigma^2}{n}} \le \mu \le \bar{x} + a\sqrt{\frac{\sigma^2}{n}} \right) = 1-\alpha = \Phi(a)-\Phi(-a)$$

show that $\alpha = \Phi^{-1}(1-\frac{\alpha}{2}) \cong Z_{1-\frac{\alpha}{2}}$

Then we have:

$$P\left(\bar{x} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \le \mu \le \bar{x} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \right) = 1-\alpha$$

Conclusion:

If $x_1, \ldots, x_n$ are the observed values of a sample from $X \sim N(\mu, \sigma^2)$ with σ^2 known, then a $100(1-\alpha)\%$ C.I for μ is $\left(\bar{x} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \right)$.

Example 2

Confidence interval for μ if σ^2 is unknown.

We replace σ^2 by the unbiased estimator $\frac{ns^2}{n-1}$. We know $\frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n-1}}} \sim t(n-1)$

As before, we obtain: $P\left(\bar{X} - t_{n-1;1-\frac{\alpha}{2}}\sqrt{\frac{s^2}{n-1}} \le \mu \le \bar{X} + t_{n-1;1-\frac{\alpha}{2}}\sqrt{\frac{s^2}{n-1}} \right) = 1-\alpha$

Conclusion:

If $x_1, \ldots, x_n$ are the observed values of a sample from $X \sim N(\mu, \sigma^2)$ with σ^2 unknown, then a $100(1-\alpha)\%$ C.I for μ is $\left(\bar{X} - t_{n-1;1-\frac{\alpha}{2}}\sqrt{\frac{s^2}{n-1}}, \bar{X} + t_{n-1;1-\frac{\alpha}{2}}\sqrt{\frac{s^2}{n-1}} \right)$

Estimation of sample size: Suppose we require that the estimate should not deviate from the parameter by a certain maximum error bound, e of probability to be at least $(1-\alpha)$, then we estimate the sample size (n) by the following:

$$P\left(|\bar{x}-\mu| < e\right) \ge 1-\alpha$$

$$P\left(\bar{x}-e < \mu < \bar{x}+e\right) \ge 1-\alpha$$

$$e \le Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} \iff n \ge \left(\frac{Z_{\alpha/2}.\sigma}{e} \right)^2$$

The assumptions underlying the preceding interval and sample size estimation are the distribution of the pivotal quantity is normally distributed with known variance, σ^2. In the case where σ^2 unknown and n is small the pivotal quantity for confidence interval for μ is obtained as follows:

The required pivotal-quantity is given by:

$$T(x,\theta) = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{\alpha(n-1)}$$

To obtain the distribution of T we make use of the following facts:

(i) $\quad Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

(ii) $\quad V = \dfrac{\sum_{I=1}^{N}(x_i-\bar{x})^2}{\sigma^2} = \dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{\alpha(n-1)}$

(iii) $\quad Z$ and V are independently random.

Theorem 4-2.1 (Student $t-$ distribution):

Suppose that the random variables Z and V are independent and have distributions $N(0, 1)$ and $\chi^2_{\alpha(n-1)}$ respectively. Define $T = \dfrac{Z}{\sqrt{\frac{v}{k}}}$, where $k = n-1$, then the pdf of T called *student $t-$distribution* with $k = (n-1)$ degrees of freedom is given by

$$f(t) = \begin{cases} \dfrac{\left|\overline{\left(\frac{k+1}{2}\right)}\right.}{\left|\overline{\left(\frac{k}{2}\right)}\right.\sqrt{\pi k}}\left(1+\dfrac{t^2}{k}\right)^{-\frac{(k+1)}{2}} & ,-\infty < t < \infty \\ 0 & , elsewhere \end{cases}$$

where $f(z) = \dfrac{1}{2\pi} e^{-\frac{1}{2}z^2}, -\infty < z < \infty$

$$g(u) = \frac{1}{\left|\overline{\left(\frac{v}{2}\right)}\right. 2^{v/2}} u^{\frac{v}{2}-1} e^{-\frac{u}{2}}, 0 < u < \infty, v = n-1$$

For the confidence interval for σ^2 in $N(\mu,\sigma^2)$, the pivotal-quantity is,

$$T = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{\alpha(n-1)} \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 .$$

That is, T has a chi-square distribution with $v = n-1$ degrees of freedom. The pdf of t

$$f(t) = \begin{cases} \dfrac{1}{\Gamma\left(\frac{v}{2}\right)2^{v/2}} t^{\frac{v}{2}-1} e^{-\frac{t}{2}}, -\infty < t < \infty \\ 0 \qquad\qquad , \; elsewhere \end{cases}$$

Putting $\alpha = v/2$, $\beta = 2$, then

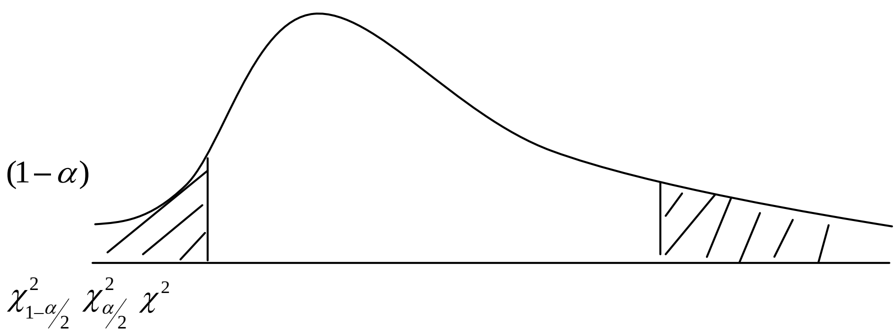
$$f(t) = f(t,\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} t^{\alpha-1} e^{-\frac{t}{\beta}}, t > 0, \beta > 0$$

which is a Gamma distribution. The pivotal-quantity $(1-\alpha)100\%$ confidence interval for σ^2,

$$P(a < T < b) = P\left(a < \frac{(n-1)s^2}{\sigma^2} < b\right) = 1-\alpha$$

$$= P\left(\frac{(n-1)s^2}{b} < \sigma^2 < \frac{(n-1)s^2}{a}\right) = 1-\alpha$$

$$= P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$



For example a 95% confidence interval for σ^2 for $n = 30$ and $s^2 = 128$ is given as

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = \left(\frac{29\times128}{45.72}, \frac{29\times128}{16.05}\right) = (81.20, 231.30)$$

The confidence interval for the ratio of two variances from $N\left(\mu,\sigma^2\right)$: Let s_1^2 and s_2^2 be sample variances from tow independent normal populations based on sizes n_1 and n_2. The pivotal quantity $F = .\dfrac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \dfrac{\sigma_2^2}{\sigma_1^2}.\dfrac{s_1^2}{s_2^2}$ is a random variable which has $F-$distribution with $v_1 = \left(n_1 - 1\right)$ and $v_2 = \left(n_2 - 1\right)$ degrees of freedom.

Suppose $u \sim \chi^2_{\alpha(v_1)}$ and $v \sim \chi^2_{\alpha(v_2)}$, then $F = \dfrac{u/v_1}{v/v_2} = \dfrac{v_1.s_1^2}{v_1.\sigma_1^2} \Big/ \dfrac{v_2.s_2^2}{\sigma_2^2.v_2} = \dfrac{s_1^2.\sigma_2^2}{\sigma_1^2.s_2^2}$

is said to have an $F-$distribution with pdf

$$g\left(f, v_1, v_2\right) = \begin{cases} \dfrac{\left\lceil\left(v_1+v_2\right)}{2} v_1^{v_1/2} v_2^{v_2/2} f^{\left(\frac{v_1}{2}-1\right)} \left(v_1 + v_2 f\right)^{-\left(\frac{\left(v_1+v_2\right)}{2}\right)}, & f > 0 \\ 0 & , \ elsewhere \end{cases}$$

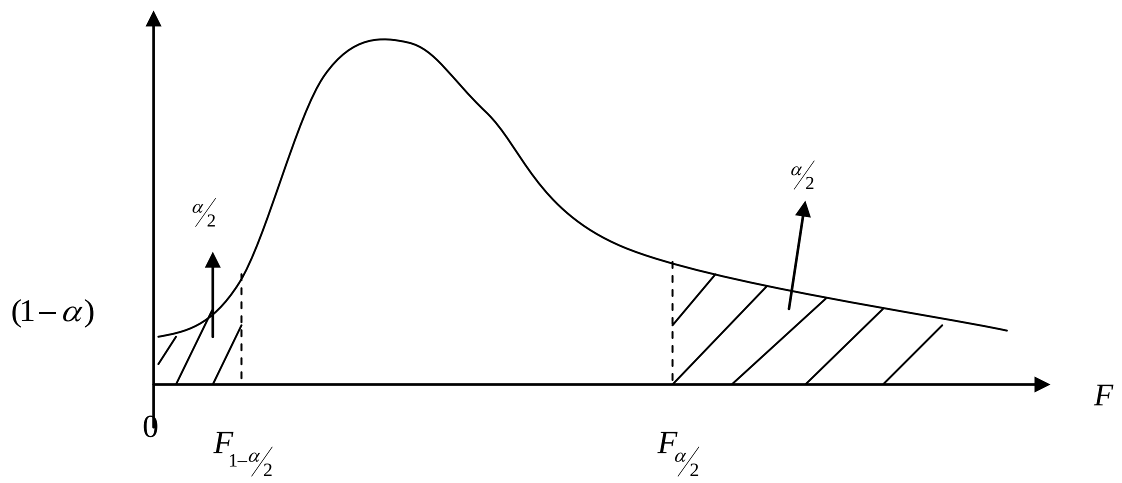
where $E(f) = \dfrac{v_2}{v_2 - 2}$, $v_2 > 2$; $Var(f) = \dfrac{v_2^2\left(2v_2 + 2v_1 - 4\right)}{v_1\left(v_2-2\right)^2\left(v_2-4\right)}$, $v_2 > 4$

The $F-$distribution is positively skewed for any values of v_1 and v_2 but becomes less skewed as v_1 and v_2 take on larger values.

$$P\left(F \le F_{(1-\alpha,v_1,v_2)}\right) = \int_0^{F_{(1-\alpha,v_1,v_2)}} g\left(f_1, v_1, v_2\right) df = 1-\alpha$$

Using the pivotal-quantity,

$$p\left(F_{\left(1-\alpha/2,v_1,v_2\right)} \le F \le F_{\left(\alpha/2,v_1,v_2\right)}\right) = 1-\alpha \text{ where } F_{\left(1-\alpha/2,v_1,v_2\right)} = \frac{1}{F_{\left(\alpha/2,v_1,v_2\right)}}$$



For example, given the data: sample $n_1 = 8,\ \bar{x}_1 = 3.1,\ s_1 = 0.25;\ n_2 = 8,\ \bar{x}_2 = 2.7,\ s_2 = 0.7,$ we find a 95% confidence interval for $\dfrac{\sigma_1^2}{\sigma_2^2}$:

$$P\left(F_{0.925,(9,7)} \le \frac{\sigma_2^2}{\sigma_1^2}.\frac{s_1^2}{s_2^2} \le F_{0.025,(9,7)} \right) = 0.95$$

$$P\left(\frac{s_1^2}{s_2^2}.\frac{1}{F_{0.025,(7,9)}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}.F_{0.025,(9,7)} \right) = 0.95$$

$$P\left(\frac{(0.25)^2}{(0.70)^2}.\frac{1}{4.83} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{(0.25)^2}{(0.70)^2}.(4.2) \right) = 0.95$$

Hence the required 95% confidence interval is $(0.0264, 0.5357)$

2-5.1.2 Approximate Method for Large Samples

The confidence interval can also be approximated for large samples where the maximum likelihood estimate (MLE), $\hat{\theta}$ has the distribution $N\left(\theta,\ \sigma_{\hat{\theta}}^2\right)$, where

$$\sigma_{\hat{\theta}}^2 = CRLB = \left(nE\left[\left(\frac{\partial}{\partial\theta} \ln f(x,\theta) \right)^2 \right] \right)^{-1}$$

That is for large n, $P\left(-Z_{\alpha/2} \le \dfrac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \le Z_{\alpha/2} \right) = 1-\alpha$

(a) An approximate $(1-\alpha)100\%$ confidence interval for Poisson parameter λ : We assume that the random variable X has the Poisson distribution,

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

We find a $(1-\alpha)100\%$ confidence interval for λ as follows:

(i) We find the CRLB for λ :

$$\ln f(x,\lambda) = -\lambda + x\ln\lambda - \ln x!$$

$$\frac{\partial}{\partial\lambda}\ln f(x,\lambda) = -1 + \frac{x}{\lambda} = \frac{x-\lambda}{\lambda}$$

$$E\left[\left(\frac{\partial}{\partial\lambda}\ln f(x,\lambda) \right)^2 \right] = E\left[\left(\frac{x-\lambda}{\lambda} \right)^2 \right] = \frac{1}{\lambda^2}E\left[(x-\lambda)^2 \right] = \frac{1}{\lambda}$$

$$\sigma_{\lambda}^{2} = CRLB = \frac{1}{n\left(1/\lambda\right)} = \frac{\lambda}{n}$$

(ii) A $(1-\alpha)100\%$ confidence interval for λ:

$$P\left(-Z_{\alpha/2} \le \frac{\bar{x}-\lambda}{\sqrt{\lambda/n}} \le Z_{\alpha/2}\right) = 1-\alpha$$

$$\left|\frac{\bar{x}-\lambda}{\sqrt{\lambda/n}}\right| \le Z_{\alpha/2} \iff \left(\frac{\bar{x}-\lambda}{\sqrt{\lambda/n}}\right)^{2} \le Z_{\alpha/2}^{2}$$

$$(\bar{x}-\lambda)^{2} \le Z_{\alpha/2}^{2}.\lambda/n$$

$$\lambda^{2} - 2\bar{x}\lambda + \bar{x}^{2} - Z_{\alpha/2}^{2}.\lambda/n \le 0$$

$$\lambda^{2} - 2\lambda\left(\bar{x} + Z_{\alpha/2}^{2}/2n\right) + \bar{x}^{2} \le 0$$

$$\lambda = \frac{2\left(\bar{x} + Z_{\alpha/2}^{2}/2n\right) \pm \sqrt{4\left(\bar{x} + Z_{\alpha/2}^{2}/2n\right)^{2} - 4\bar{x}^{2}}}{2}$$

$$= \bar{x} + \frac{Z_{\alpha/2}^{2}}{2n} \pm \sqrt{\left(\bar{x} + Z_{\alpha/2}^{2}/2n\right)^{2} - \bar{x}^{2}}$$

$$= \bar{x} + \frac{Z_{\alpha/2}^{2}}{2n} \pm \sqrt{\bar{x}^{2} + 2\bar{x}.\frac{Z_{\alpha/2}^{2}}{2n} + \frac{Z_{\alpha/2}^{4}}{4n^{2}} - \bar{x}^{2}}$$

$$= \bar{x} + \frac{Z_{\alpha/2}^{2}}{2n} \pm \sqrt{\frac{Z_{\alpha/2}^{2}}{n}\left(\bar{x} + \frac{Z_{\alpha/2}^{2}}{4n}\right)}$$

$$= \bar{x} + \frac{Z_{\alpha/2}^{2}}{2n} \pm Z_{\alpha/2}\sqrt{\frac{1}{n}\left(\bar{x} + \frac{Z_{\alpha/2}^{2}}{4n}\right)}$$

Neglecting terms in $Z_{\alpha/2}^{2}$ we have,

$$\hat{\lambda} = \bar{x} \pm Z_{\alpha/2}\sqrt{\frac{\bar{x}}{n}}$$

(b) An approximate $(1-\alpha)100\%$ confidence interval for the binomial parameter p:

Let the random variable x has the Binomial distribution,

$$x \sim B(n,p) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0,1,2,\ldots,n$$

(i) We find $CRLB = \sigma_p^2 = \dfrac{p(1-p)}{n}$

(ii) $$P\left(-Z_{\alpha/2} \le \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \le Z_{\alpha/2} \right) = P\left(\left| \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \right| \le Z_{\alpha/2} \right) = 1-\alpha$$

$$\left| \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \right| \le Z_{\alpha/2} \quad \Leftrightarrow \quad \left(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \right)^2 \le Z_{\alpha/2}^2$$

$$(\hat{p}-p)^2 \le Z_{\alpha/2}^2 . \frac{p(1-p)}{n}$$

$$\hat{p}^2 - 2\hat{p}p + p^2 \le Z_{\alpha/2}^2 \frac{p}{n} - Z_{\alpha/2}^2 \frac{p^2}{n}$$

$$p^2 + Z_{\alpha/2}^2 \frac{p^2}{n} - 2\hat{p}p - Z_{\alpha/2}^2 \frac{p}{n} + \hat{p}^2 \le 0$$

Solving for p we have

$$p^2 \left(1 + \frac{Z_{\alpha/2}^2}{n} \right) - \left(2\hat{p} + \frac{Z_{\alpha/2}^2}{n} \right) p + \hat{p}^2 \le 0$$

$$p = \frac{2\hat{p} + \frac{Z_{\alpha/2}^2}{n} \pm \sqrt{\left(2\hat{p} + \frac{Z_{\alpha/2}^2}{n} \right) - 4\left(1 + \frac{Z_{\alpha/2}^2}{n} \right) \hat{p}^2}}{2\left(1 + \frac{Z_{\alpha/2}^2}{n} \right)}$$

$$= \frac{2\hat{p} + \frac{Z_{\alpha/2}^2}{n} \pm \sqrt{4\hat{p}^2 + 4\hat{p}.\frac{Z_{\alpha/2}^2}{n} + \frac{Z_{\alpha/2}^4}{n^2} - 4\hat{p}^2 - 4Z_{\alpha/2}^2\frac{\hat{p}}{n}}}{2\left(1 + \frac{Z_{\alpha/2}^2}{n}\right)}$$

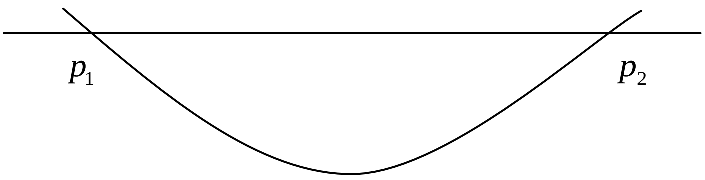
$$\frac{2\hat{p} + \frac{Z_{\alpha/2}^2}{n} \pm \sqrt{4Z_{\alpha/2}^2\frac{\hat{p}}{n}\left(1-\hat{p}\right) + \frac{Z_{\alpha/2}^4}{n^2}}}{2\left(1 + \frac{Z_{\alpha/2}^2}{n}\right)}$$

Ignoring terms in $\frac{Z_{\alpha/2}^2}{n}$ and higher we have.

$$p = \hat{p} \pm Z_{\alpha/2}.\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

Hence the approximate $\left(1-\alpha\right)100\%$ confidence interval for p is

$p \in \left[p_1, p_2\right]$, where $p_1 = \hat{p} - Z_{\alpha/2}.\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$ and $p_2 = \hat{p} + Z_{\alpha/2}.\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$



Learning Track Activities

Unit Summary

1. [Summary based on learning objective 1]
2. [Summary based on learning objective 2]
3. [Summary based on learning objective 3]
4. [Summary based on learning objective 4]

Key terms/ New Words in Unit

1. [State key term/New Words: explain what it means]
2. [State key term/New Words: explain what it means]
3. [State key term/New Words: explain what it means]
4. [State key term/New Words: explain what it means]

(choose an appropriate activity and give detail instructions)

- **Review Question:** [insert details of review question]
- **Discussion Question:** [insert details of discussion question]
- **Web Activity:** [insert details of web Activity or online v-classroom]
- **Reading:** [insert details of literature for reading]
- **Interactive CD:** [insert details of interactive CD]

Unit Assignments 5

1. **(a)** Let $X_1, X_2, \ldots, X_n$ be a random sample from the probability density function,

$$f(x,\alpha) = \begin{cases} \frac{2}{\alpha^2}(\alpha - x), & 0 \leq x \leq \alpha \\ 0 & , \ elsewhere \end{cases}$$ where α is an unknown positive constant.

(i) Find the method of moments estimator for α .

(ii) Use your results in (i) and (ii) to show whether the estimator in (ii) is unbiased or not.

(iii) Evaluate the estimator in (ii) given the sample observations, 12.4, 5.9, 10.1, 7.3 and 15.5.

(iv) Find CRLB

(b) Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from the uniform distribution on the interval $(0, \beta)$. Let $\hat{\beta}_1 = \frac{2}{n}\sum_{i=1}^{n} y_i$ and $\hat{\beta}_2 = \left(\frac{n+1}{n}\right) Max\left(Y_1, Y_2, \ldots, Y_n\right)$ be estimators for β.

(i) Find the sampling distribution of $\hat{\beta}_2$.

(ii) Show that $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased for β.

(iii) Find the relative efficiency of $\hat{\beta}_1$ with respect to $\hat{\beta}_2$.

(iv) Which of the two estimators is more efficient?

2. **(a)** Suppose we have a random sample of size 20 from a population where $E(x) = \mu$ and $Var(x) = \sigma^2$. Let

$$\bar{x}_1 = \frac{1}{2n}\sum_{i=1}^{2n} x_i \text{ and } \bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

be no estimators for μ. Which is the better estimator for μ? Explain your choice.

(b) Let $x_1, x_2, x_3, \ldots, x_n$ denote a random sample from a population with mean, μ haring μ and variance, σ^2. Consider the following estimators of μ

$$\hat{\theta}_1 = \frac{x_1 + x_2 + \ldots + x_7}{7}, \text{ and } \hat{\theta}_2 = \frac{2x - x_6 + x_4}{2}$$

(iii) Is either estimator unbiased?

(iv) Which estimator is "best"? In what sense is it best?

3. **(a)** A random variable has probability density function,

$$f\left(x, \beta\right) = \left(\beta + 1\right) x^{\beta}, \ 0 < x < 1, \ \beta > 1$$

Obtain the MLE of β, based on the random sample $X_1, X_2, ..., X_n$ and comment on your result. Evaluate the estimate if he sample values are: 0.30, 0.80, 0.27, 0.35, 0.62, 0.55.

(b) An experiment produces the data $X_1, X_2, ..., X_n$ for random variable X with pdf $f\left(x, \theta\right) = \frac{c}{\theta^2} x e^{-2x/\theta}, x > 0$

Determine the likelihood function for the parameter given the random sample $x_1, x_2, ..., x_n$. Derive the expression for the MLE. If 6 observations are: 5.1, 6.2, 1.3, 6.3, 5.6, 3.1. Derive the expression for $r(\theta) = \ln R(\theta)$

4. **(a)** A relationship existing between x and y is given by the linear regression model, $y_i = 5 + \beta x_i + \varepsilon_i$, $i = 1, 2, . . . , n.$ What is the least squares estimator for β?

(b) Let x and y be related by the linear relation $y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i$, where $i = 1, 2, . . ., 6$ and $\sum_{i=1}^{n} x_i = 21$, $\sum_{i=1}^{n} x_i^2 = 91$, $\sum_{i=1}^{n} x_i y_i = 1,027$, $\sum_{i=1}^{n} y_i = 225$, and $\sum_{i=1}^{n} y_i^2 = 11,879$.

(i) Find the estimates of α_0 and α_1.

(ii) Find an estimated an value of y for $x = 5$.

(c) The drying time, y (in hours) of a vanish is related to a certain amount of chemical, x (in grams) that is added. The following are random sample of the data:

x	1	2	3	4	5	6	7
y	9.0	7.5	6.0	5.0	5.5	6.0	7.0

(i) Estimate the regression model suggested in (i) by estimating the parameters of the model.

(ii) Obtain a point for the mean drying time of a vanish, $E(y)$ if 7.5 grams of the chemical is added.

5 **(a)** Two meters are used to measure water pressure in a pipeline. One is known to be more accurate than the other but both are subject to random error. Readings by meter A are subject to a standard deviation, σ and readings by meter B have a standard deviation of 1.25σ. Meter A is used to take 6 independent measurements with a mean of $\bar{x}_A$. Meter B is used to take 10 independent measurements with a mean of $\overline{x}_B$. Suppose that both means are unbiased estimators for the true mean water pressure θ.

(iv) which estimator would you prefer if you had to choose between $\overline{x}_A$ *and* $\overline{x}_B$?

(v) Find the relationship between the constants α and β if $Z = \alpha \overline{x}_A + \beta \overline{x}_B$ is to be an unbiased estimator θ.

(vi) Find the variance of Z.

(b) Suppose Y_1, Y_2 and Y_3 denote a random sample from the exponential distribution with parameter θ. Consider the following as estimators of θ.

$$T_1 = y_1 \qquad T_2 = \frac{1}{2}(Y_1 + Y_2)$$

$$T_3 = \frac{1}{3}(Y_1 + 2Y_3) \qquad T_4 = \frac{1}{3}(Y_1 + Y_2 + Y_3)$$

(iii) Which of the above estimators is/are unbiased for θ?

(iv) What is the most efficient estimator?

(c) Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from $f(y,\theta) = \dfrac{2y}{\theta^2}$, $0 < y < \theta$ and $\hat{\theta} = \dfrac{3}{2}\overline{y}$ an estimator for θ. Find the efficiency of the estimator.

(c) A random sample, X_1, X_2, X_3 *and* X_4 from the probability distribution,

$$f(x,\theta) = \frac{\theta^{2x} e^{-\theta^2}}{x!}, \; x = 0, 1, 2, \ldots \;, \theta > 0$$

produced the data 17, 10, 32 and 5.

(i) Write down the likelihood function,

(ii) Find the maximum likelihood estimate of θ.

Unit **6**

THEORY OF HYPOTHESIS TESTING

Introduction

Very often, researchers are confronted with some problems which call upon them to take decisions on the basis of available data instead of finding estimates of the parameters characterizing the population(s) being studied. This decision-making procedure is called hypothesistesting and isone of the most useful aspects of statistical inference since many types of such problems, tests, or experiments in various fields of real world can be formulated as hypothesis-testing problems. Furthermore, as we will see, there is a very close connection between hypothesis testing and confidence intervals. In unit 2, we discussed the various concepts of parameter estimation. In this unit, we build on the develop methods of testing or verifying a hypothesized distribution involving parameter for a population on the basis of a sample taken from the population presented in MATH 271. In our development, statistical hypotheses concern functional forms of the assumed distributions. These distributions may be specified completely with pre-specified values for their parameters or they may be specified with parameters yet to be estimated from the sample. Topics to be discussed in this unit include computation of Type I and Type II errors and determination of critical regions of hypothesis testing.

Learning Objectives

After careful study of this unit, students should be able to appreciate some basic principles of construction of hypothesis testing and apply them. Specifically, they should be able to:

1. Understand the concepts of Type I and Type II errors committed in hypothesis testing.
2. Compute the Type I and II errors and power of a test, and make sample size selection decisions for tests on means and proportions
3. Explain and use the relationship between confidence intervals and hypothesis tests
4. Apply the likelihood ratio tests for the determination of the most uniformly critical regions.

SECTION 1-3: ERRORS IN HYPOTHESIS TESTING

1-3.1 Introduction

A hypothesis is a statistical assertion about the probability distribution function, $f(x, \theta)$ of population. If a statistical hypothesis completely specifies the underlying distribution or value of population parameter θ, it is referred to as a *simple hypothesis*; otherwise it is called *composite hypothesis*. A *hypothesis testing* is a statistical procedure that uses a random sample data to determine whether a statement about a population should be rejected or not. In testing for the validity of a hypothesis we usually propose two types of hypotheses namely; the *null hypothesis*, denoted H_0, which is the tentative statement assumed to be true, and the *alternative hypothesis*, denoted H_1, which contradicts with the null hypothesis. It is accepted only when sufficient evidence exists to establish its truth.

As in parameter estimation, errors or risks are inherent in deciding whether testing the null hypothesis, H_0 against the alternative hypothesis, H_1 on the basis of random sample information, should be accepted or rejected. There are basically types of two errors likely to be made. These are the *Types I* and *Type II errors* which come about as a result of the decisions which are taken. These are illustrated in the diagram below.

Decision	*Actual Situation*	
	H_0 *is true*	H_0 *is false* (H_1 *is true)*
Accept H_0	Correct decision $(1-\alpha)$	Type II error (β)
Reject H_0	Type I error (α)	Correct decision $(1-\beta)$

Definition 1-3.1 (Type I and II Errors):

(a) Type I error is the error or mistake committed of rejecting the null hypothesis, H_0

when it is actually true. The probability of committing this error is denoted by α, which is also referred to *level of significance* and indicates the size of the critical region.

(b) Type II error is error or mistake committed of failing to reject (accept) the null hypothesis, H_0 when it is actually false. The probability of committing this error is denoted by β, which can only be determined by specifying the alternative hypothesis.

(c) The probability of rejecting a false null hypothesis, denoted by $(1-\beta)$, is referred to as the *power* of a hypothesis test. It is often used to gauge the test's effectiveness in recognizing that a null hypothesis is false. A common guideline is to plan an experiment so that the resulting power is at least 0.80, so that the hypothesis test is very effective in rejecting a false null hypothesis.

In order for a good decision to be taken for a given hypothesis testing, it must be designed so as to minimize the errors α and β. Ideally, both α and β to should be equal to zero, which is not possible in reality. Mathematically, α, β, and the sample size n are related and the usual practice in research and industry is to select the values of α and n, so the value of β is determined. Usually, a value of β greater than 0.20 is often considered too high for a hypothesis test to provide significant results. Depending on the seriousness of a Type I error we try to use the largest α that can be tolerated. For Type I errors with more serious consequences, we select smaller values of α and then choose n as large as is reasonable, based on considerations of time, cost, and other relevant factors.

The following practical considerations may be relevant:

- For any fixed α, an increase in the sample size n will cause a decrease in β. That

  is, a larger sample will lessen the chance that you make the error of not rejecting the null hypothesis when it is actually false.
- For any fixed sample size n, a decrease in α will cause an increase in β. Conversely, an increase in α will cause a decrease in β.
- The size of the critical region, and consequently the probability of a type I error, α can always be reduced by appropriate selection of the critical values. To decrease both α and β, increase the sample size.
- When the null hypothesis is false, β increases as the true value of the parameter approaches the value hypothesized in the null hypothesis. The value of β decreases as the difference between the true mean and the hypothesized value increases.

Generally, the analyst controls the type I error probability when he or she selects the critical values. Thus, it is usually easy for the analyst to set the type I error probability at (or near) any desired value. Since the analyst can directly control the probability of wrongly rejecting H_0, we always think of rejection of the null hypothesis H_0 as *a strongconclusion*. On the other hand, the probability of type II error β is not a constant, but depends on the true value of the parameter. It also depends on the sample size that we have selected. Because the type II error probability β is a function of both the sample size and the extent to which the null hypothesis H_0 is false, it is customary to think of the decision to accept H_0 as a *weak conclusion***,** unless we know that β is acceptably small. Therefore, rather than saying we “accept H_0”, we prefer the terminology “fail to reject H_0”.

Failing to reject H_0 implies that we have not found sufficient evidence to reject H_0, that is, to make a strong statement. Failing to reject H_0 does not necessarily mean that there is a high probability that H_0 is true. It may simply mean that more data are required to reach a strong conclusion. This can have important implications for the formulation of hypotheses. Another important concept that we will make use of is the power of a statistical test.

1-3.2 Computation of Type I and Type II Errors

We illustrate the computation of Type I and Type II errors as well as the sample size with the following examples.

Example 1-3.1:

Find the Type I and II errors for testing $H_0:\ p=\frac{1}{5}$ against $H_1: p=\frac{4}{5}$, given 10 sample observations and rejecting H_0 if $x \ge 3$, where $x \sim B(n,\ p)$.

Solution:

Given that $x \sim B(10,\ p)$ for the test $H_0:\ p=\frac{1}{5}$ against $H_1: p=\frac{4}{5}$.

(i) $\alpha = P(\text{Type I error}) = P\left(\text{reject } H_0 \,|\, H_0 \text{ is true}\right)$

$$\alpha = P\left(x \ge 3 \middle| H_0 : p = \frac{1}{5}\right)$$

$$= 1 - P\left(x \le 2 \middle| H_0 : p = \frac{1}{5}\right)$$

$$= 1 - \left\{ P\left(x = 0 \middle| H_0 : p = \frac{1}{5}\right) + P\left(x = 1 \middle| H_0 : p = \frac{1}{5}\right) + P\left(x = 2 \middle| H_0 : p = \frac{1}{5}\right) \right\}$$

$$= 1 - \left\{ \binom{10}{0}\left(\frac{1}{5}\right)^0\left(\frac{4}{5}\right)^{10} + \binom{10}{1}\left(\frac{1}{5}\right)^1\left(\frac{4}{5}\right)^{9} + \binom{10}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^{8} \right\}$$

$$= 1 - \left\{ \left(\frac{4}{5}\right)^{10} + 10\left(\frac{1}{5}\right)^1\left(\frac{4}{5}\right)^{9} + 45\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^{8} \right\}$$

$$= 1 - 4.04\left(\frac{4}{5}\right)^8 = 0.32220$$

(ii) $\beta = P(\text{Type II error}) = P\left(\text{accept } H_0 \,|\, H_0 \text{ is false}\right)$

$$\beta = P\left(x \le 2 \middle| H_1 : p = \frac{4}{5}\right)$$

$$= \left\{P\left(x = 0 \middle| H_0 : p = \frac{4}{5}\right) + P\left(x = 1 \middle| H_0 : p = \frac{4}{5}\right) + P\left(x = 2 \middle| H_0 : p = \frac{4}{5}\right)\right\}$$

$$= \left\{\binom{10}{0}\left(\frac{4}{5}\right)^0\left(\frac{1}{5}\right)^{10} + \binom{10}{1}\left(\frac{4}{5}\right)^1\left(\frac{1}{5}\right)^9 + \binom{10}{2}\left(\frac{4}{5}\right)^2\left(\frac{1}{5}\right)^8\right\}$$

$$= \left\{\left(\frac{1}{5}\right)^{10} + 10\left(\frac{4}{5}\right)^1\left(\frac{1}{5}\right)^9 + 45\left(\frac{4}{5}\right)^2\left(\frac{1}{5}\right)^8\right\}$$

$$= 30.44\left(\frac{1}{5}\right)^8 = 0.00007793$$

<u>Example 1-3.2:</u>

Consider a random sample of size 20 from the normal distribution $N(\mu, \sigma^2)$, where that $\bar{x} = 11$ and $s^2 = 16$. If $H_0 : \mu = 12$ is tested against $H_1 : \mu < 12$ at significance level $\alpha = 0.01$:

(a) Compute the probability of committing Type II error and the power of the test if $H_1 : \mu = 10.5$.

(b) What sample size is needed for the power of the test to be 0.90 for the alternative value $\mu = 10.5$?

<u>Solution</u>:

Given the sample data $12 = 20,\ \bar{x} = 11,\ s^2 = 16$ drawn from $N(\mu, \sigma^2)$, and the test $H_0 : \mu = 12$ is tested against $H_1 : \mu < 12$ at significance level $\alpha = 0.01$:

(i) We first determine the critical region (CR)

$$\alpha = P\left(\bar{x} \in CR \middle| H_0 : \mu = \mu_0 = 12\right) = 0.01$$

$$= P\left(\bar{x} \le k \middle| H_0 : \mu = 12\right) = 0.01, \text{where } k \text{ is the critical region.}$$

$$P\left(Z \le \frac{k - \mu_0}{\sigma/\sqrt{n}} \middle| H_0 : \mu = 12\right) = 0.01$$

$$P\left(Z \le \frac{k - 12}{4/\sqrt{20}} \middle| H_0 : \mu = 12\right) = 0.01$$

$$\phi\left(Z \le \frac{k - 12}{0.8944}\right) = 0.01$$

$$\left(Z \le \frac{k-12}{0.8944}\right) = -\phi^{-1}\left(0.01\right)$$

$$Z \le \frac{k-12}{0.8944} = -2.33\left(0.01\right)$$

$$\Rightarrow k-12 \le -2.33\left(0.8944\right)$$

$$\Rightarrow k \le 12-2.33\left(0.8944\right) \approx 9.92$$

Hence the critical region is $CR: \overline{x} \le 9.92$

(ii) $\beta = P\left(Type\ II\ error\right) = P\left(accept\ H_0 \left| H_0\ is\ false\right.\right)$

$$\beta = P\left(\overline{x} > 9.92 \left| H_1 : \mu = 10.5\right.\right)$$

$$= 1 - P\left(\overline{x} \le 9.92 \left| H_1 : \mu = 10.5\right.\right)$$

$$= 1 - \phi\left(\frac{9.92-10.5}{0.8944}\right)$$

$$= 1 - \phi\left(-0.65\right) = 1 - 0.2578 = 0.7422$$

Hence the power $= 1 - \beta = 0.2578$

(iii) Given a power of $= 1 - \beta = 0.90$ it means than $\beta = 0.10$ and the required sample size is given by $n = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{(\mu_1 - \mu_0)^2} = \frac{(2.33+1.28)^2(16)}{(10.5-12)^2} \approx 93$

SECTION 2-3: BEST CRITICAL REGIONS

2-3.1 Generalized Likelihood Ratio (GLR)

Ideally we would like both α and β be small but simultaneous reduction is usually impossible. To compromise this we fix α at an acceptable low level and maximize β. The generalized likelihood ratio (GLR) is a working criterion for actually suggesting test procedures.

Definition 2-3.1:

Consider the hypothesis test $H_0: \theta = \theta_0$ against $H_1: \theta < \theta_0$ at a specific level of significance, α. Let the two parameter spaces, ω and Ω be defined as follows:

$\omega = \{\theta : \theta = \theta_0\}$, the set of unknown parameter values admissible under H_0 $\Omega = \{\theta : 0 < \theta < \theta_0\}$, the set of all possible values of all unknown parameters. That is, ω is a particular subset of Ω.. Let $x_1, x_2, ---, x_n$ be a random sample from the probability density function, $f(x,\theta)$. The generalized likelihood ratio

$$\Lambda(\lambda) = \frac{\underset{\omega}{Max} L(X, \theta_1, \theta_2, ---, \theta_k)}{\underset{\Omega}{Max} L(X, \theta_1, \theta_2, ---, \theta_k)} = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{L(\theta_0)}{L(\theta_1)}$$

where $0 \le \lambda \le 1$. If λ is close to 1, it suggests that the data are very compatible with H_0 and hence H_0 is accepted. If λ is close to 0 the data would not be very compatible with the parameter values in ω and it would make sense to reject H_0. This is the rationale behind the Generalized Likelihood Ration principle as stated in the next definition.

A generalized likelihood ratio test (GLRT) is one that rejects H_0 whenever $0 < \lambda < \lambda^*$, where ere λ^* is chosen so that $P(0 < \Lambda < \lambda^* \mid H_0) = \alpha$ where Λ is a random variable under H_0 and λ^* is the critical value in the critical region which is determined from the equation,

$\int_0^{\lambda^*} g_\lambda(\lambda \mid H_0)\, d\lambda = \alpha$, where $g_\lambda(\lambda \mid H_0)$ is the probability density function of Λ which is monotonic function.

2-3.2Neyman-Pearson Lemma

The GLR criterion test is a modification of an even more basic result in the theory of Hypothesis Testing, called the Neyman-Pearson Lemma. Unlike the GLR criterion, which carries with it no guarantee of optimality, a test based on the Neyman-Pearson Lemma is "best" in the sense that no other procedure having he same Type I error probability has a lower Type II error probability. However, Neyman-Pearson Lemma results is too restrictive for much practical significance.

In intuitive terms we describe the result as follows: Suppose a random sample of size n, $x_1, x_2, ---, x_n$ is drawn from pdf $f(x,\theta)$, where θ is unknown but is restricted to one of only two possible values θ_0 or θ_1 in testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$. Consider the set, C, which is a collection of n tuples of size α, such that $\Pr\left[(x_1, x_2, ---, x_n) \in C \mid H_0\right] = \alpha$, where C is said to be the critical region for a test of $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$ (where $\theta_1 > \theta_0$). If the observed n-tuples $(x_1, x_2, ---, x_n)$ falls in the preselected C, we reject H_0. What if H_1 is true? Intuitively, the "best" critical region would be the one having the highest probability of containing $(x_1, x_2, ---, x_n)$

<u>Definition 2-3.2:</u>

C^* is said to be the best critical region of size α if

$\Pr\left[(x_1, x_2, ---, x_n) \in C^* \mid H_0\right] = \alpha$, and

$\Pr\left[(x_1, x_2, ---, x_n) \in C^* \mid H_1\right] \geq \Pr\left[(x_1, x_2, ---, x_n) \in C \mid H_1\right]$ for every $C \neq C^*$

The Neyman-Pearson Lemma provides us with a working criterion for actually finding C^*. The theorem of Neyman-Pearson Lemma is based on the likelihood ratio, $\dfrac{L(\theta_0)}{L(\theta_1)}$.

<u>Theorem 2-3.1 (The Neyman-Pearson Lemma):</u>

Let $(x_1, x_2, ---, x_n)$ be a random sample from the pdf $f(x,\theta)$ where θ is either θ_0 or θ_1. Let $L(\theta_i; x_1, x_2, ---, x_n)$ denote the likelihood function of the observed sample when $\theta = \theta_i, i = 0 \text{ or } 1$. Let k be a positive number. The best critical region C^* for testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$ is the set of n-tuples, $(x_1, x_2, ---, x_n)$ satisfying the following three conditions:

$$\frac{L(\theta_0; x_1, x_2, ---, x_n)}{L(\theta_1; x_1, x_2, ---, x_n)} \leq k \text{ for each n-tuple in } C^*$$

$$\frac{L(\theta_0; x_1, x_2, ---, x_n)}{L(\theta_1; x_1, x_2, ---, x_n)} > k \text{ for each n-tuple not in } C^*$$

$$\Pr\left[\left(x_1, x_2, ---, x_n\right) \in C^* \mid \theta = \theta_0\right] = \alpha$$

2-3.3Illustrative Example

Example 2-3.1:

Suppose y represents a single observation from the pdf given by

$$f(y,\theta) = \begin{cases} \theta y^{\theta-1}, 0 < y < 1 \\ 0, elsewhere \end{cases}$$

Find the most powerful test with significance level $\alpha = 0.05$ to test $H_0 : \theta = 1$ against $H_1 : \theta = 2$.

Solution

The likelihood ratio,

$$\frac{L(y,\theta_0)}{L(y,\theta_1)} = \frac{f(y,\theta_0)}{f(y,\theta_1)} = \frac{f(y,1)}{f(y,2)} = \frac{1}{2y}$$

Hence by the Neyman-Pearson Lemma,

$$\frac{L(y,\theta_0)}{L(y,\theta_1)} \le k$$

$$\Rightarrow \frac{1}{2y} \le k$$

$\Rightarrow y \ge \frac{1}{2k} = k'$ where $k' = \frac{1}{2k}$ and is determined by

$$\Pr\left(y \ge k' \mid H_0 : \theta = 1\right) = \alpha = 0.05$$

$$\int_{k'}^{1} f(y,1)\,dy = 0.05$$

$$\int_{k'}^{1} 1.dy = 0.05$$

$$y\Big|_{k'}^{1} = 1 - k' = 0.05$$

$$\Rightarrow k' = 1 - 0.05 = 0.95$$

This indicates the size of the critical region (CR): $\{y : y \ge 0.95\}$

Thus among all test for H_0 against H_1 based on the smallest size of one and α fixed at 0.05, this test has the smallest Type II error. That is, it is the most powerful test with significance level $\alpha = 0.05$.

Example 2-3.2:

Let $(x_1, x_2, ---, x_n)$ be a random sample of size n from a distribution with pdf,

$$f(x,\theta) = \begin{cases} \theta e^{\theta x}, x \geq 0 \\ 0, elsewhere \end{cases}$$

For testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_1$ where θ_0 is specified, what is the best critical region (or a uniformly most powerful test)?

Solution:

Let $\theta_1 > \theta_0$. To test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, the Neyman-Pearson criterion gives the best critical region based on a sample size n so that

$$\frac{L(\theta_0, x_1, x_2, ---, x_n)}{L(\theta_1, x_1, x_2, ---, x_n)} = \frac{\theta_0^n e^{\theta_0 \sum x_i}}{\theta_1^n e^{\theta_1 \sum x_i}} = \left(\frac{\theta_0^n}{\theta_1^n}\right)^n e^{(\theta_1 - \theta_0)\sum x_i}$$

That is $\left(\frac{\theta_0^n}{\theta_1^n}\right)^n e^{(\theta_1 - \theta_0)\sum x_i} \leq k$

$$n \ln\left(\frac{\theta_0^n}{\theta_1^n}\right) + (\theta_1 - \theta_0)\sum x_i \leq \ln k$$

$$(\theta_1 - \theta_0)\sum x_i \leq \ln k - n\ln\left(\frac{\theta_0^n}{\theta_1^n}\right)$$

$$\sum x_i \leq \frac{\ln k - n\ln\left(\theta_0^n / \theta_1^n\right)}{(\theta_1 - \theta_0)}$$

$$\bar{x} = \frac{1}{n}\left\{\frac{\ln k - n\ln\left(\theta_0^n / \theta_1^n\right)}{(\theta_1 - \theta_0)}\right\} = k'$$

That is the best critical region is defined by $\bar{x} \leq k$. This region forms a uniformly most powerful region for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

Example 2-3.3:

Suppose $\left(x_1, x_2, ---, x_n\right)$ is a random sample from the $N\left(\theta,1\right)$ distribution, where θ is

either 3 or 2 and we want to test $H_0 : \theta = 3$ against $H_1 : \theta = 2$. Setting $\alpha = 0.05$, the likelihood ratio is

$$\frac{L\left(\theta_0; x_1, x_2, ---, x_{10}\right)}{L\left(\theta_1; x_1, x_2, ---, x_{10}\right)} = \frac{\left(1/2\pi\right)^{10} e^{-1/2\sum\left(x_i-3\right)^2}}{\left(1/2\pi\right)^{10} e^{-1/2\sum\left(x_i-2\right)^2}}$$

$$= e^{-1/2\left\{\sum\left(x_i-3\right)^2 - \sum\left(x_i-2\right)^2\right\}} = e^{\sum x_i - 25}$$

By the Neyman-Pearson Lemma, the 10-tuples that should be included in the critical region C^* are those for which $e^{\sum x_i - 25} \le k$ where k is to be chosen so that

$$\Pr\left(e^{\sum x_i - 25} \le k \mid H_0 : \theta = 3\right) = 0.05$$

$$e^{\sum x_i - 25} \le k$$

$$e^{\sum x_i - 25} \le k$$
$$\sum x_i - 25 \le \ln k$$
$$\sum x_i \le \ln k + 25$$
$$\bar{x} \le \frac{1}{n}\left\{\ln k + 25\right\} = k'$$

That is $\Pr\left(e^{\sum x_i - 25} \le k \mid H_0 : \theta = 3\right) = \Pr\left(\bar{x} \le k' \mid \theta = 3\right) = 0.05$

But $\bar{x} \sim N\left(\theta, 1/10\right)$ so that

$$\Pr\left(\bar{x} \le k' \mid \theta = 3\right) = \Pr\left(\frac{\bar{x} - 3}{\sqrt{1/10}} \le k'' \mid \theta = 3\right) = 0.05$$

That is $\Pr\left(Z \le k''\right) = 0.05$

$$\Rightarrow k'' = \phi^{-1}\left(0.05\right) = -1.645$$

form which we obtain the critical region as follows:

$$\frac{\bar{x} - 3}{\sqrt{1/10}} \le -1.645$$

$$\bar{x} \le 3 - \frac{1}{\sqrt{10}}\left(1.645\right) = 2.48$$

That is we reject $H_0 : \theta = 3$ if $\bar{x} < 3$

Example 2-3.4:

Let X possess a Poisson distribution with mean μ, $f(x,\mu) = \dfrac{e^{-\mu}\mu^x}{x!}, x = 0,1,2,---$

Suppose we want to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternate hypothesis $H_1 : \mu = \mu_1$, where $\mu_1 < \mu_0$. Find the best critical region for this test.

Solution:

We use the likelihood Ratio Test to find the best critical region. The method suggested by the Neyman-Pearson Theorem.

$$\frac{L(x_1, x_2, ---, x_n, \mu_1)}{L(x_1, x_2, ---, x_n, \mu_0)} = \frac{\prod_{i=1}^{n}\left(e^{-\mu_1}\mu_1 / x!\right)}{\prod_{i=1}^{n}\left(e^{-\mu_0}\mu_0 / x!\right)} = e^{n(\mu_0 - \mu_1)}\left(\frac{\mu_1}{\mu_0}\right)^{\sum x_i}$$

That is. $e^{n(\mu_0 - \mu_1)}\left(\dfrac{\mu_1}{\mu_0}\right)^{\sum x_i} \geq k$

$$n(\mu_0 - \mu_1) + (\ln \mu_1 - \ln \mu_0)\sum x_i \geq \ln k$$

$$(\ln \mu_1 - \ln \mu_0)\sum x_i \geq \ln k - n(\mu_0 - \mu_1)$$

$$\sum x_i \leq \frac{\ln k - n(\mu_0 - \mu_1)}{(\ln \mu_1 - \ln \mu_0)}, \text{ since } \mu_0 - \mu_1 < 0$$

$$\bar{x} \leq \frac{1}{n}\left\{\frac{\ln k - n(\mu_0 - \mu_1)}{(\ln \mu_1 - \ln \mu_0)}\right\} = k'$$

$$\bar{x} \leq k'$$

2-3.4Self-Assessment Questions (Exercise 2.2)

1. **(a)** Define the following as used in Hypothesis Testing.

(i) *Type I and Type II errors*

(ii) *Power of a Test*

(iii) *Best Critical Region*

(b) State and explain the significance of the *Neyman-Pearson Lemma.*

(c) The null hypothesis, $H_0: \mu = \mu_0$ is tested against the alternative $H_1: \mu \neq \mu_1$ at a significance level of α. Derive the formula for the determination of the sample size *n.*

2. **(a)** Let P represent the proportion of defective items in a manufacturing process. To test $H_0: p = 0.25$ against $H_1: p \geq 0.25$, a random sample of 5 observations is taken from the process. If the number of defective items is 4 or more, the null hypothesis is rejected. What is the probability of rejecting H_0 if $p = 0.20$?

(b) Let $\bar{x}$ be the mean of a random sample from a normal distribution with variance, 9. The hypothesis $H_0: \mu = 100$ is rejected in favour of $H_1: \mu > 100$ if $\bar{x} > C$. If the level of significance is 0.05, find the minimum sample size necessary to achieve a power of 0.50 when $\mu = 101$.

(c) A random sample of size 36 items is selected from a population with standard deviation, 12. What is the critical region for the test $H_0: \mu \geq 120$ against $H_1: \mu < 120$ at $\alpha = 0.10$?

3. Let $y_1, y_2, ---, y_n$ be a random sample from a normal distribution with unknown mean μ and known variance $\sigma^2 = 5$. We wish to test $H_0: \mu = 7$ against $H_1: \mu > 7$.

(a) Find the uniformly most powerful test with significance level $\alpha = 0.05$.

(b) For the test in (i), find the power against each of the following alternatives: $\mu = 7.5, 8.0, 8.5, 9.0$

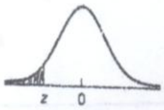
(c) Sketch a graph of the power function.

4. Let y have the Poisson distribution,

$$f(y, \theta) = \frac{\theta^y e^{-\theta}}{y!}, \textit{ where } \ y = 0, 1, 2, \ldots, \textit{ and } \theta > 0.$$

What is the likelihood ratio critical region for testing $H_0: \theta = 1$ against $H_1: \theta \neq 1$?

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	0.0006	0.0007	0.0007
−3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010
−3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
−2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	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	0.0446
−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.

LECTURE NOTES

FOR

MATH 153

(Statistical Methods I)