Yashwantrao
Chavan
Maharashtra
Open University

ज्ञानगंगा घरोघरी

CMP 221/504

**Statistical
Techniques**

Statistical Techniques

Writers : Mr. Madhav B. Kulkarni

Yashwantrao Chavan Maharashtra Open University

Vice-Chancellor: Prof. E.Vayunandan

SCHOOL OF COMPUTER SCIENCE : SCHOOL COUNCIL

Shri. Ramchandra Tiwari
Director
School of Computer Science
Y. C. M. Open University, Nashik

Dr. Shyam Ashtekar
Director
School of Health Sciences
Y. C. M. Open University, Nashik

Dr. R. V. Vadnere
Director
School of Continuing Education
Y. C. M. Open University, Nashik

Dr. Prakash Atkare
Controller of Examination
Evaluation Division
Y. C. M. Open University, Nashik

Shri. Manoj Killedar
Director
School of Science & Technology
Y. C. M. Open University, Nashik

Shri. Madhav Palshikar
Lecturer
School of Computer Science
Y. C. M. Open University, Nashik

Shri. Pramod Khandare
Lecturer
School of Computer Science
Y. C. M. Open University, Nashik

Dr. Parvati Rajan
Ex. HOD, Post Graduate Dept. of Computer Science, SNDT Woman's College, Mumbai

Dr. K. V. Kale
Professor, Computer Science and IT Dept. Dr. Babasaheb Ambedkar Marathwada University
Aurangabad

Prof. (Mrs.) Seema Purohit
Head of Department, Dept. of Computer Science & IT, Kirtee College
Kashinath Dhuru Road, Near VSNL Buldg.
Dadar (West), Mumbai - 400 028

Dr. S. S. Sane
Head of Department
Computer Department
K. K. Wagh College of Engineering
Nashik

Prof. S. G. Khurd
Physics Department
Bytco College
Nashik Road, Nashik

Writer

Mr. Madhav B. Kulkarni
Head, Department of Mathematics and Statistics
B. Y. K. College of Commerce
T. A. Kulkarni Vidyanagar
College Road, Nashik
Email - Madhavbk@gmail.com

Editor

Dr. Sudhakar Kunte
Rtd. Professor
Dept. of Statistics
University of Pune
Pune

Coordinator

Mr. Madhav Palshikar
Sr. Lecturer
School of Computer Science
Y.C.M. Open University
Nashik

Production

Shri. Anand Yadav
Manager,
Print Production Centre
Y. C. M. Open University, Nashik - 422 222

(First edition developed under DEC development grant)

First Publication : August 2009 **Reprint :** Sept. 2018 **Publication No. :** 1806
Typesetting & Cover Design : Om Computers, Nashik - 422 007
Printed by : Shri. Prashant Racca, M/s. Racca Printers, Kaveri Sankul, Ashok Stambh, Nashik 422001
Publisher : Dr. Dinesh Bhonde, Registrar, Y. C. M. Open University, Nashik - 422 222
ISBN - 978-81-8055-345-5

Unit 1 : Introduction of Statistics

1.0 Overview

In this introductory unit you will be introduced to the subject of Statistics and you will study some basic terms of used in the study of subject of Statistics. Statistics is many times understood as information. This is correct. But there is something more in it. Statistics can also be used to infer. The information under certain conditions can be used in the process of inference. The information (data) under certain assumptions can be used to draw valid inferences. If you know exactly what is being measured and how it is measured you know lot about it. You will be introduced to measurement scales. In sample surveys on census, investigators collect data on attributes or on variables. The data on attributes are either reported in nominal or ordinal scales. The data on variables are either measured in interval scale of measurement or ratio scale of measurement.

1.1 Learning Objectives

After studying this unit you will be able to –

- Explain "Statistics" and some terms used in it,
- Differentiate between information and data,
- Understand various measurement scales,
- Define and explain Statistics in day to day language,
- Understand use, scope and limitations of Statistics.

1.2 Introduction

In today's world we all have to deal with numbers, proportions, percentages, indexes. Many times we express our opinions with the help of numbers. We study prices of essential commodities such as cereals, pulses, milk, sugar, etc. We are careful about changing price levels of real estates. Many of us study share prices. It is therefore necessary that we understand meaning of the word Statistics and then use it in appropriate context. Statistics as an art of presentation and Statistics as a science of making inference about the population are two different things. For example, study the following statements.

(1) Population of Maharashtra will be 1000 million in near future and that of Mumbai will be 200 million.

(2) It is estimated that total production of wheat will be about 800 million tones this year.

(3) There are about 55,000 primary schools in the state of Maharashtra.

(4) The number of passengers in a local train running in Mumbai is above three times its capacity.

(5) The share index may touch a figure of 10000 in next few months.

We certainly can prepare such list of statements. This is information. It is certainly useful and is of basic importance. This information can be looked upon as a raw material for the subject of Statistics. "How can we use these raw material and for what purpose?" is a challenging question. A statistician as a scientist is expected to make an intelligent use of the information. His aim is to predict or guess the unknown things. Statistician's responsibility consists of developing an appropriate methodology (i) to make sound predictions for the unknown quantities, (ii) to estimate lower and upper bounds on these predictions, (iiI) to prepare and implement a scientific method to draw a sample from the population (iv) to design an experiment and then compare two or more treatments (v) to draft a questionnaire for sample surveys and to educate researchers about statistical analysis, etc. To do all these things it is must that statistical approach rests on a sound scientific footing. Presenting the results of statistical analysis is undoubted an art. Statistics is therefore a science and also an art.

Statistics essentially deals with gathering of information, analyzing it, drawing conclusions and recommending reasonable remedies. Usually the information is incomplete and moreover the quality of information is questionable. Thus there exists an element of uncertainty. But we simply cannot avoid it. The role of a statistician is to suggest ways and help researchers to arrive at rational and correct decisions in the presence of uncertainty. No one can give 100% guarantee that the predicted results will be true but some confidence about it can certainly be stated. Is Statistics a branch of Mathematics? No, certainly not. It is true that a student of Statistics will be required to study basic mathematics (and also higher level of mathematics), but nonetheless statistics as a science differs from mathematics. For example, suppose the price of milk is Rs 22 per liter and you purchase 2 liters of milk. So you will pay Rs 44 (22 x 2 = 44) to a milkman. This is mathematics, not statistics. This is because everything here is certain. We all know that the amount of milk in one litter bag is not exactly 1 litter. There is always some error. A statement that with probability 0.95 the quantity of milk in a one litter bag is within 995 to 1010 ml. Such a statement is Statistics. If the outcome is certain and can be predicted with 100% guarantee then it is not statistics; if the outcome or response can't be predicted with 100% guarantee then it is a subject matter of Statistics. Usually a researcher is interested in deterministic models. However, no such models usually represent the real situation because every observation contains some amount of experimental or observational error, and hence he is forced to use a statistical model, which also include an error term. When you toss a coin and observe it falling down, this event is sure to happen. It is a deterministic experiment. The outcome is certain. But your interest may be to observe whether the coin has shown head or tail. This is another experiment. The outcome of this experiment cannot be predicted beforehand, but some assumptions about it can be made. We assume that the coin has two sides – one of them is labeled as Head (H) and another Tail (T). When tossed we will observe either head or tail, but not both. Also we assume that the coin will never stand on its edge. We can prepare a list of possible outcomes. Such a list (set) is called a sample space and is denoted by symbol Ω (greek letter Omega). The elements of the sample space are called sample points. A sample point is denoted by ω. Thus sample space Ω is the set of all possible sample points. If in the experiment the lady luck (chance factor) plays a role then there is an element of uncertainty and such experiment certainly calls for the subject of statistics. The laws in Physics such as force is equal to the product of mass and acceleration (F = m*a) or Boyle's law ($P_1V_1 = P_2V_2$) or Ohm's laws (V = I/R) are deterministic laws. Even when the experiments for the confirmation of such laws are made in a laboratory, the same results are not observed repeatedly because of the experimental errors. In our daily life we do perform several experiments, knowingly or unknowingly. Suppose you want to know the fuel efficiency of your scooter. What will you do? You can put 1 liter of petrol and check how much it can give service (mileage or Km) in it. You may repeat the experiment say 10 times. Total running divided by total petrol used will be an estimate of average running of your vehicle per liter of petrol. But remember it is an average and therefore you cannot give 100% guarantee that next time it will give you the same average. Some of you may wonder why it should change every time. Think about the variables (factors) that may affect fuel efficiency. Age of a scooter, number of people riding the vehicle, weight of riders, road conditions, petrol + oil combination, type of petrol – "speed" or normal, air pressure in the

wheels, etc. can be some factors that may matter. We will later study random variables, their properties, etc. Here we will begin with basics of descriptive statistics. Let us first study the definition of statistics and in the process understand what statistics is all about.

1.3 Definition of Statistics

Many scientists and researchers have tried to define Statistics. We reproduce below various definitions. Needless to say these definitions do not cover ever-growing world of statistics. In literature the word “Statistics” is used in to different ways: (i) The representative numbers used to describe a data like its mean, variance, standard deviation etc. (ii) A science of making inference about the population quantities from the observed data on a sample.

Definitions:

- “Statistics are the classified facts representing the collections of the people in a State … specially those facts which can be stated in number or in tables of numbers or in any tabular or classified arrangements". (**Webster)**
- “Statistics are numerical statement of facts in any department of enquiry placed in relation to each other". (**Bowley)**
- “By Statistics we mean quantitative data affected to a marked extent by multiplicity of causes." (**Yule and Kendall**)
- “Statistics may be defined as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other." (**Prof. Horace Secrist)**

Remarks:

i. According to Webster's definition only numerical facts can be termed as Statistics. Moreover it restricts the domain of Statistics to the affairs of a State, i.e., to social sciences. This is a very old and narrow definition and is inadequate for modern times, since today Statistics embraces all sciences viz. social, physical and natural.

ii. Bowley's definition is more general than Webster's, since it is related to numerical data in any department of enquiry. Moreover, it also provides for comparative study of the figures as against mere classification and tabulation of Webster's definition.

iii. Yule and Kendall's definition refers to numerical data affected by a multiplicity of causes. This is usually the case in social, economic and business phenomena. For, example, the prices of a particular commodity are affected by a number of factors, viz. supply, demand, imports, exports, money in circulation, competitive products in the market and so on. Similarly, the yield of a crop depends upon the factors like quality of seed, fertility of soil, method of cultivation, irrigation facilities, weather conditions, fertiliser used and so on.

iv. Secrist's definition seems to be the most exhaustive of all the four. Let us describe in details.

- **Aggregate of facts:** Simple or isolated items cannot be termed as Statistics unless they are part of aggregate of facts relating to any particular field of enquiry.
- **Affected by multiplicity of causes:** Numerical figures should be affected by multiplicity of factors. This point is already discussed in remark (iii).
- **Numerically expressed:** Only numerical data constitute Statistics. Thus, the statement like 'the standard of living in Nashik has improved' does not constitute

Statistics. The qualitative or descriptive characteristics do not constitute Statistics unless they are expressed in numbers.

- **Enumerated or Estimated according to Reasonable Standard of accuracy:** The numerical data pertaining to any field of enquiry can be obtained by completely enumerating the underlying population. In such a case data will be exact and accurate. However, if complete enumeration of the underlying population is not possible or not practicable, then the population quantities are estimated by using the principles of Statistics (Sampling and Estimation theory). In such case, the estimated values may not be exactly same as the actual values. But certain standards of accuracy are maintained for drawing meaningful conclusions.
- **Collected in Systematic manner:** The data must be collected in a systematic manner. We must define the characteristics under study and also the population. You can either adopt census survey or sample survey. Trained investigators must conduct the enquiry.
- **Collected for a Pre-determined Purpose:** It is of utmost importance to define in clear and concrete terms the objectives or the purpose of the enquiry and the data should be collected keeping in view these objectives. We should not waste our time in collecting the information, which is irrelevant for our enquiry. Also check that no essential data are omitted.
- **Comparable:** You can compare two or more data sets only when these data are arising from comparable populations. They may be compared with respect to some unit, time period or place.

From the above discussion, it can be concluded that, “All Statistics are numerical statements of facts but all numerical statements of facts are not Statistics".

Statistics as a plural noun means masses of figures. As a singular noun it means the science of statistics. We routinely ask for the data of various types for one or more reasons. What is “Data”? The word data means information in numerical form generated by observations or experimentation. In the statistical language the data are of two types:

(A) Categorical data (data on qualitative variables)

(B) Measurement data (data on quantitative variables)

The data type invariably calls for statistical techniques appropriate to it. It is therefore necessary that we first study various types of data. Also measurement data can further be subdivided as **discrete** or **continuous.** In discrete data we can have only isolated real values. These values may or may not be integers. For example, number of defective items in a batch of 1000 items, white blood cell count (WBC), litter size, proportion of “damaged” cells in a sample of 100 cells, proportion of male children in a family having n children, number of seeds per pod, are example of discrete data. In case of continuous data all the values in an interval [a, b] on a real line are possible. Blood pressure of a patient, height and weight of an individual, residual life time of a cancer patient, the time required to recover from a disease, skin fold thickness, head circumference, reliability of a component, reliability of a system are examples of a continuous data. One can argue that all these continuous data sets are also discrete because they are measured only up to their smallest unit of measurement. However, this discreteness in the data is because of our inability to measure any thing continuously. The discreteness here is not a property of the variable under consideration. Hence, we treat these data as continuous data.

1.4 Scales of Measurement

Qualitative data are measured either on a nominal or an ordinal scale. Quantitative data are measured on an interval or a ratio scale.

Nominal Scale: Here the observations are classified into categories that are different in character and cannot be measured or ordered. For a patient, heart disease may be present (+) or absent (-). Or he may be a cancer patient (+) or may not be a cancer patient (-). His hair color may be black (+) or may be other than black (-), the patient is either male (1) or female (0).

Ordinal Scale Here all the observations refer to a common character. *These observations can be grouped into a limited number of categories, which can be ordered* in an ascending series. Thus herein categories can be ranked. For example, the health status of an individual may be malnourished or normal. A cancer patient may be in stage 3 (+++) or in stage 2 (++) or in stage 1 (+). Instead of + sign we may use codes such as 1, 2, 3. It is true that ordinal data is 'semi quantitative', but ++ is not same as (2* +). Never confuse codes and measurements. The interval between successive categories will usually be different.

Both nominal and ordinal data give rise to **counts** of the frequency of occurrence of the various categories within the groups. These counts are therefore bound to be whole numbers such as 0, 1, 2… n. We may convert them into rates, proportions or percentages. Therefore they may appear in decimal numbers, but these counts need to be analyzed in different manner.

Interval Scale: Here the observations are made on a **scale** so that in addition to ordinal level of measurement distance between any two successive numbers is fixed and equal. For example, body temperature (in ^{0}C). A zero degree temperature does not mean no temperature.

Ratio Scale: Here in addition to interval level of measurement the variable under consideration has **true zero** point as its origin. For example body weight (in Kg), body mass volume (in m^3), etc.

Note: During a sample survey or census investigators collect data. It may happen that for certain variables for one or more cases either the data are not available, or missing. If in your studies you have such missing data points then indicate such data points clearly in your data sheet by inserting "appropriate" codes. In case of Excel data sheet keep blank space in place of missing data, in

MINITAB use asterisk (*). In Statgraphics leave a blank in the corresponding position in the worksheet. **Know the software you may use in statistical analysis.**

1.5 Scope and Importance of Statistics

In the ancient times Statistics was regarded only as the science of Statecraft and was used to collect information relating to population, wealth, crimes and military strength etc. for devising military and fiscal policies. Later on the scope of Statistics widened to social and economic phenomena. Moreover, with the developments in the Statistical techniques during the last some decades, Statistics is viewed not only as a mere device for collecting numerical data but as a technique for handling, analysing and drawing inferences from them.

Statistics has its applications in social sciences and pure sciences. Statistical techniques such as statistical quality control, statistical process control are extensively used. In pharmaceutical industries two or more drugs are compared for testing their effectiveness and clinical trials are designed. Geneticists and agricultural scientists work jointly with statisticians to compare a new variety of a crop (wheat / soyabean / cotton). Economists use lot of statistical techniques to forecast future demands for essential commodities, to predict economic growth rate, to estimate poverty in India. Biologists make use of statistical techniques in micro-array data analysis (human genome projects), actuaries use statistical approaches to decide premium rates and so on. With the availability of high-speed computers, Statistics as a science and technique has assumed unprecedented importance. Statistical thinking is becoming more and more indispensable. The present era is of information. Information is in the form of data. An elementary knowledge of statistical methods has become a part of the general education. H. G.

Wells rightly remarked that “Statistical thinking will one day be as necessary for effective citizenship as the ability to read and write". According to Bowley, “A knowledge of Statistics is like a knowledge of foreign language or of algebra; it may prove of use at any time under any circumstances".

1.6 Limitations of Statistics

Statistics can be used or misused by its users. We briefly state below the limitations of Statistics. It is necessary for the user of statistics to know these limitations so that he can interpret the results with a wit.

Statistics is not suited to the study of qualitative phenomenon. Statistics, being a science dealing with a set of numerical data, is applicable to the study of only those subjects of enquiry, which are capable of quantitative measurements. For example, poverty, culture, race, religion, sex, etc. cannot be expressed numerically. Therefore, they are not capable of direct statistical analysis. However, these non-measurable concepts can be measured using statistics. Qualitative characteristics can be labelled by codes and then only its appropriate analysis can be done. Statistics does not study individuals.

Statistics deals with an aggregate of objects and does not give any specific recognition to the individual item in the series.

Statistical laws are not exact. Unlike law of Physics and Natural sciences, Statistical laws are only approximations. Using statistical analysis we can suggest appropriate probabilistic models.

An intelligent person can misuse statistical techniques. But then aeroplanes were also used as weapons by terrorists. This is therefore not a surprise. **“Statistical methods are the most dangerous tools in the hands of the inexpert. Statistics is one of those sciences whose adepts must exercise the self-restraint of an artist"**. As King says, “**Statistics are like clay of which one can make a god or devil as one pleases**”.

1.7 Summary

Statistics as information and statistics as inference are two different things. Statistics can be defined in many ways. In earlier days it was looked upon as a science of data gathering, analyzing and reporting the results. Statistical analysis is a science and presenting results is an art. In today’s world it is being used in social and pure sciences. The statistical analysis invariably deals with data. Data are measured in appropriate scale and accordingly reported. In Statistics we deal with non-deterministic models. Statisticians can use statistics and statistical techniques effectively but user must also know its limitations.

1.8 Self Test

1. Read the following paragraph and answer multiple choice questions

“A researcher interested in estimating expenditure pattern on various commodities conducted a sample survey in various parts of a city. He collected data on age, sex, caste, monthly income, and monthly expenditure on food, housing, travel, entertainment and family size. He also referred data reported in journals published by National Sample Survey Organization.”

(i) Measurement on age is in ---

a. nominal scale b. ordinal scale c. ratio scale d. interval scale

(ii) Measurement on family size is in ---

a. nominal scale b. ordinal scale c. ratio scale d. interval scale

(iii) Data collected by the researcher is ---

a. primary b. secondary c. groped d. primary and secondary

(iv) The researcher can use statistics as a science to study ---

a. individual **b.** group or individuals

c. population from which sample is drawn **d.** none of a, b and c

(v) Data on sex and caste is

a. both nominal **b.** nominal on sex but ordinal on caste

c. both ordinal **d.** nominal on caste but ordinal on sex

2. Identify scale of measurement for following variables:

a) I.Q of a student

b) Blood glucose level

c) sex of a newborn baby.

d) Nutritional status of an infant.

e) ABO blood group system.

f) Sitting height (in cm)

g) Anxiety score

h) Stress level

i) Magnesium level of a patient before and after the operation.

3. Match the pairs:

Variable	Scale of Measurement
1. Monthly Expenditure on petrol	Ratio level
2. State of domicile	Ratio level
3. Quality of a product	Ratio level
4. grade in the final examination.	Ratio level
5. Sex of a person	Interval level
6. Status of a patient	Interval level
7.Profession of the respondent	Interval level
8. # seeds set in a pod	Interval level
9. Impurity level in petrol.	Nominal scale
10. Stress level of a manager	Ordinal scale
11. Temperature in a room	Nominal scale
	Ordinal scale
	Ordinal scale

4. Define and explain the term “Statistics”.

5. Discuss use of Statistics in planning /economics.

6. State limitations of Statistics

7. List real life experiments wherein the outcomes are not certain.

8. In a socio-economic survey you are asked to collect data on age, sex, caste, income, expenditure, land holdings, real-estate ownership, and family size. Identify the variables and explain how you will measure them.

Unit 2 : Classification and Tabulation

2.1 Overview

In this unit you will be introduced to the summarizing techniques essential in data analysis. To begin with concepts of data – ungrouped and grouped data are discussed. A brief review of methods of preparing one way, two way and three way tables is discussed. Also students will learn the different methods of preparing frequency tables and its usefulness. At the end cumulative frequencies and ogive curves are introduced.

2.2 Learning Objectives

After studying this unit, you will be able to

- Distinguish between ungrouped (raw) and grouped data, primary and secondary data,
- Construct tables with one, two and three factors of classification,
- Classify the data into various types of frequency distribution tables,
- Draw cumulative frequency (ogive) curves.

2.3 Introduction

In every walk of life we always ask for data. Remember that there must be some reliable agencies that collect these data. The data must be collected taking into consideration principles of statistics, probability and sample surveys. Such agencies or institutes need to be established and promoted in the interest of society and for national development. In India various state governments and also the Central government have done a wonderful work in developing the system of collecting, documenting the data and its monitoring is regularly done. Some government organizations which does this data collection job regularly are: National Sample Survey Organization (NSSO), Central Statistical Organization (CSO), Directorate of Economics and Statistics, District Statistical Offices. Indian Statistical Institute at Kolkata, founded by Prof. P C. Mahalnobis, is internationally recognized centre for research and training in Statistics. Various types of data are regularly published by these agencies. It may not be always feasible, nor even desirable, to ask the researcher on his own to collect the requisite data If the researcher collects the data on his own then for him such data is labeled as **primary data.** Another researcher or other users can use the data collected by one agency. For another user these data will be labeled as **secondary data.** In every type of enquiry one must first explore the possibility of looking into secondary data. If these data are not sufficient then only one can plan for primary data. In any sample or census surveys it is necessary to train the investigators first and then only they should collect appropriate data. These data are called as **raw data.** The raw data collected must be arranged in a neat, systematic form. Representation of data in a good understandable format is an art. It must be attractive and salient features of the data must figure out promptly. There are three ways to representation a data: Textual Representation, Tabular representation and Diagrammatic representation.

2.4 Representation of Data

In this unit we will study only the Textual and the Tabular Representation of the data. The Diagrammatic Representation will be studied in the next unit.

2.4.1 Textual Representation

In textual representation the data are described in a paragraph(s). If the linguistic style is not user friendly or poor, user cannot read whatever is expected. In such a poorly constructed text the necessary Information is either missing or cannot be located easily.

2.4.2 Illustration of Textual Representation

"Exactly a fifth of the number of students in a University of strength 20,000 are ladies, 33 of every 40 students are Maharashtrians, 13 out of every 16 gents are Maharashtrian gents. 40% of non-Maharashtrian gents and 60%Maharashtrian gents have offered Arts subjects. 40% of ladies from Maharashtra and equal percentage of non-Maharashtrian ladies have offered Science subjects."

If you read the paragraphs similar to above you will have to make special efforts to understand the meaning of each sentence. Surely this must be avoided.

2.4.3 Tabular Representation

The tabular representation is most suitable in many situations. It has the merits of having a diagrammatic presentation effect and is easy to understand. In tabulation all the information is visible at a glance. We will study a typical statistical table and its parts.

2.4.4 Statistical Table

The data collected can be presented in one-way, two-way or three-way table. Whatever may be the type of table, an ideal table must have following parts.

Title,

Headings of rows (stub) or subheadings,

Headings of columns or captions,

Footnote for explanations,

Sources of information,

The numerical data are entered in the body of the table.

Sketch of Statistical table

Title:……

Stub

	Headings of columns or captions
Headings of The Rows or Sub-headings	Body Of The Table

Foot note : ……

Source : ……

Note: the body of the table always represents the frequency of values of the characteristic observed in the data.

2.4.5 Types of Tables

If the data consists of values of a single characteristic of the observed units, then you prepare a one-way table. If you have two such characteristics, you need a two-way table. Do not prepare tables of dimensions higher than three because they are very complicated to understand. We shall study the three types of Statistical tables, viz., One-way table, Two-way table and Three-way table.

One-way table

If only one characteristic is observed, the table representing such data is called one-way table. Here the first column consists of the values of the characteristic under consideration and the second column consists of the frequencies of these values in the data. The one-way tables below are prepared separately using only one characteristic at a time out of the data on three characteristics of students belonging to either Science or Arts faculty, viz. faculty, state and sex.

Table 2.1 : University students by faculty

Faculty	***Number of Students***
Science	8,600
Arts	11,400
Total	20,000

Table 2.2 : University students by state

State	***Number of Students***
Maharashtra	16,500
Other than Maharashtra	3,500
Total	20,000

Table 2.3 : University students by sex

Sex	***Number of students***
Gents	16,000
Ladies	4,000
Total	20,000

Two-way Table

If the data on each unit are available on two characteristics then make use of 'two-way' table. One of the characteristics is represented row-wise and other column-wise. Here the first column consists of the values of one characteristic under consideration and the first row consists of the values of the other characteristic under consideration. The cells represent the frequencies of pairs of corresponding values of the two characteristics in the data. Such tables also help us understand the relationship between the two characteristics under consideration. Three two-way tables are given below. In these tables two out of three characteristics viz. faculty and state, faculty and sex, state and sex are considered one by one.

Table 2.4: University students by faculty and state

State → *Faculty ↓*	*Maharashtra*	***Other than Maharashtra***	***Total***
Science	6,600	2,000	8,600
Arts	9,900	1,500	11,400
Total	16,500	3,500	20,000

Table 2.5: University students by faculty and sex

Sex → *Faculty ↓*	***Gents***	***Ladies***	***Total***
Science	7,000	1,600	8,600
Arts	9,000	2,400	11,400
Total	16,000	4,000	20,000

Table 2.6: University students by sex and faculty

Sex → *State ↓*	***Gents***	***Ladies***	***Total***
Maharashtra	13,000	3,500	16,500
Other than Maharashtra	3,000	500	3,500
Total	16,000	4,000	20,000

Three-way table

When the information is to be reported on three characteristics of the units, you will require a 'three-way' table. It is a little complicated table, but there are many occasions where such tables are reported. **Here each column representing one value of the second characteristic is subdivided into sub-columns each representing one value of the third characteristic**. As an illustration a three-way table is constructed below. The three characteristics being reported are faculty, state and sex. Here sex is considered to be less important characteristic and is appearing as sub-class of state.

Table 2.7: University students by faculty, state and sex

State →	*Maharashtra*			*Other than Maharashtra*		
Sex → *Faculty ↓*	*Gents*	*Ladies*	*Total*	*Gents*	*Ladies*	*Total*
Science	5,200	1,400	6,600	1,800	200	2,000
Arts	7,800	2,100	9,900	1,200	300	1,500
Total	13,000	3,500	16,500	3,000	500	3,500

If the number of characteristics observed on the unit is more than three, it is better to prepare separate statistical tables for different characteristics taking into account the relationships among them. It facilitates to simplify the representation and understanding the data.

2.4.6 Requirements of a Good Statistical Table

1. A table should possess a brief but a self-explanatory title which can answer "what, when, where" abut the data.
2. Headings of rows (stub) and columns (caption) should be clearly stated.
3. Units should be mentioned whenever necessary.
4. Avoid short forms as far as possible.
5. Classes and subclasses should be clearly separated by lines of different colors or thick and thin lines.
6. Whole of the table should be visible at glance.
7. Footnote must be included for explanations of signs / abbreviations.
8. Source note must include reference of the source of data, if available.
9. For easy reference a table should bear a table number.

2.4.7 Advantages of Tabular Presentations

1. It is convenient and self-sufficient form of presenting the statistical information,
2. It summarizes the information and displays important features of it,
3. Unnecessary repetitions (that may appear in texts) are avoided,
4. Comparison between localities, age-groups, etc. can be made easily,
5. Errors and omissions in the information can be easily detected,
6. Reference to any details of the data is facilitated.

Note: Frequency tables of the data are also called the frequency distribution.

2.5 Classification of Data

Suppose our interest is to study a characteristic or a set of characteristics of any type. For example, it can be height of adult healthy individual, expenditure on a mobile phone, use of internet facility (in hours/day), etc. The very first step is to state a problem and then second step is to collect relevant data from the part of the population. The part of the population on which the data are collected is called as a sample from that population. We all are aware that valid conclusions cannot be drawn if we have a very few values in the data. Statisticians therefore advocate a large sample size and insist on scientific methods to select the sample on which to collect the data from the population under consideration. Generally for large samples raw data is difficult to understand and comprehend. If we are given a large sample and values on the set of variables we fail to understand the significance of these raw data. Moreover if the number of observations in the data is large, we observe that values of some of the observations are the same or are very close. We can find how many times each of the different values is occurring. If a value x_i occurs f_i times, we say that f_i is the frequency of x_i. This idea of condensing the data can be further extended. This is a process of preparing a frequency distribution. During your school days, you might have studied how to prepare a frequency distribution. However we will revisit these details in brief. In general the procedure for preparing a frequency table is as follows:

a. If the study variable is discrete, and it takes only a few values, then first make a table with first column consisting of the different values of the characteristics. Next read the observed values sequentially and for each value place a tally mark **"/"** in table against appropriate value.

b. If the study variable is discrete and its takes large number of isolated values in an interval then first make a few classes corresponding to consecutive values. Next read the values one by one carefully and for each value place a tally mark **"/"** in table against appropriate interval called class.

c. if the variable is continuous, then first divide the range in suitable groups (subintervals) which are not overlapping. and then Next read the values one by one and for each value place a tally mark **"/"** in table against appropriate subinterval called class.

To facilitate easy counting of tally marks, tally marks are arranged in blocks of five, every fifth tally mark being drawn across the preceding four. This was perhaps necessary few decades ago when computers were not easily available. With availability of computers (using statistical packages or EXCEL) the computation work of statisticians is reduced considerably

You may ask the question, "How and how many of the classes should be formed? We do not have any specific rule but guidelines can be provided. First of all remember that the total number of groups should be adequately small. If the number of groups is large then the table would not be easy to understand. If the number is too small approximation to the data values will be too large. Also take care that each observation should go to one and only one group and not even a single observation is omitted. The number of observations in each class is called the

frequency of that class. A basic assumption that is made while performing any calculations is that all the observations in a class interval have values actually equal to the mid-value of the class-interval. Many times the secondary data is available in the grouped form only. We give below some illustrative examples. Study these illustrations and acquaint yourself with the steps used in preparing frequency table of various types.

Illustration (1): The number of misprints per page (X) is a discrete random variable. The possible values of X are 0, 1, 2, … The data on observed number of misprints on 100 pages of a book are reported below. .

Table 2.8: Misprints in a book of 100 pages

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

The frequency distribution of the values is prepared taking single valued classes as 0, 1, 2, 3, 4 and 5 as 5 is the maximum number of misprints on a page.

Table 2.9: Frequency distribution of Number of misprints

Description(Class: x) (# misprints)	Tally marks	Frequency (f)
0	||||	4
1	𝍸 𝍸 |||	13
2	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |	36
3	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||	32
4	𝍸 𝍸 |	11
5	||||	4

Illustration (2): Data on weight (in kg) of 100 adult healthy individuals is recorded during a nutrition survey. In Table 2.10, these weights are reported (in nearest integer). We will prepare a frequency distribution of these data. Note that weight is a continuous random variable. In the Table 2.10, the values of weights (in nearest integer) are reported. You must know how the data are reported in the tables by the investigators. It helps in preparing appropriate type of frequency distribution.

Table 2.10: Weights of adults (in kg)

66	60	57	55	60	63	54	67	51	60	66	50	62	55	56	63	62	53	53	62
64	69	64	59	63	69	69	55	64	63	62	56	51	63	59	50	62	61	61	55
59	48	61	61	59	60	63	57	61	69	59	59	56	60	68	58	58	66	61	65
56	66	59	48	58	67	58	60	60	64	63	58	58	61	68	59	62	65	52	65
65	48	53	55	63	52	61	65	61	64	57	59	70	59	56	57	65	60	66	64

Here the classes proposed are 47 - 49, 50 - 52, 53 – 55… If you want to remove discontinuity in the classes you may prepare classes such as 46.5 - 49.5, 49-5 – 52.5, and so on. We will study these details later in this chapter.

Table 2.11: Frequency distribution of weights of 100 individuals

Weight in kg.	Tally marks	Frequency
47 – 49	|||	3
50 – 52	𝍸	6
53 – 55	𝍸 𝍸	9
56 – 58	𝍸 𝍸 𝍸	15
59 – 61	𝍸 𝍸 𝍸 𝍸 𝍸 ||	27
62 – 64	𝍸 𝍸 𝍸 𝍸	20
65 – 67	𝍸 𝍸 |||	13
68 – 70	𝍸 ||	7
Total		100

Some times, the frequency distribution is prepared stating actual values belonging to the classes. For example, the frequency distribution of weights can be prepared as follows:

Weight in kg	Observed values	Frequency
47 – 49	48 48 48	3
50 – 52	50 50 51 51 52 52	6
53 – 55	53 53 53 54 55 55 55 55 55	9
56 – 58	56 56 56 56 56 57 57 57 57 58 58 58 58 58 58	15
59 – 61	59 59 59 59 59 59 59 59 59 59 60 60 60 60 60 60 60 60 61 61 61 61 61 61 61 61 61	27
62 – 64	62 62 62 62 62 62 63 63 63 63 63 63 63 63 64 64 64 64 64 64	20
65 – 67	65 65 65 65 65 65 66 66 66 66 66 67 67	13
68 – 70	68 68 69 69 69 69 70	7

This method of preparing frequency distribution is more cumbersome than that based on tally marks. Here the observed values are rewritten in the table in place of tally marks. Above is the frequency distribution of a continuous variable. For a discrete random variable such a frequency distribution can also be used. But such a presentation is used less frequently. Sometimes instead of frequency relative frequency is reported in frequency distribution table. Relative frequency of a class is expressed in percentage is (100*frequency of that class)/Size of the data.

2.5.1 Inclusive Method of Frequency Distribution

If the observations equal to lower value of a class and the observation equal to upper value of a class are included in the same class, then it is an inclusive method of frequency distribution. That is, if both the values designating a class are included in the same class then we have inclusive method of frequency distribution.

In the above distribution, the values equal to 47 and 49, 50 and 52, etc are included in the same (respective) class therefore it is inclusive method of frequency distribution. If inclusive method is used to prepare frequency distribution, there will be a gap between upper values of a class and lower value of the next class.

2.5.2 Exclusive Method of Frequency Distribution

If in a class, lower value of a class is included but the upper value of class is excluded from the same class (and included in the next class) then it is called exclusive method of frequency distribution. For example, 47 is included in the class 47 – 50 but 50 is excluded from it and is included in the class 50 – 53

Class limit: The two numbers that specify the class are called class limits. Lower value of the class is called Lower Class Limit (LCL) and upper value is called Upper Class Limit (UCL).

Remark: A continuous variable can take infinite values in the interval [a, b] and we do not expect any discontinuity in successive classes. A gap between LCL of a class and UCL of next class is trivial for a discrete variable. Therefore, when we construct a frequency distribution of inclusive type for a continuous type variable, a care must be taken to interpret the class intervals. For example, the class 47 - 49 in the above frequency distribution contains all the values between 46.5 and 49.5. Here we are assuming that in the data values are reported up to the nearest unit. In other words, the class 47 - 49 actually stands for 46.5 - 49.5. The limits 46.5 and 49.5 are known as **class boundaries** of the class 47 - 49.

Rule for finding class boundaries: Let ***d*** **denote** the difference between UCL and LCL of successive classes. Then subtract **d**/2 from LCL of each class and add **d**/2 to UCL. You will get upper class boundary. The class boundaries for frequency distribution of weights of adults are constructed. Here d = 1, subtract **d**/2 = 0.5 from every LCL to get lower class boundary (LCB) and add 0.5 to every UCL to get upper class boundary (UCB). Study Table 2.12 reported below.

Table 2.12: Frequency distribution of weights of 100 adults

Weight in kg.	Class boundaries	Frequency
47 – 49	46.5 – 49.5	3
50 – 52	49.5 – 52.5	6
53 – 55	52.5 – 55.5	9
56 – 58	55.5 – 58.5	15
59 – 61	58.5 – 61.5	27
62 – 64	61.5 – 64.5	20
65 – 67	64.5 – 67.5	13
68 – 70	67.6 – 70.5	7

Class Width: The difference between UCB and LCB of a class is defined as class-width of a class and is usually denoted by **h**. For ease of calculations it is usually recommended to have classes of same width, but this may not be always possible. For example, in case of income distribution class width cannot be same. In the frequency distribution of weights of adults, the class width is 3 (49.5 - 46.5 = 3) or 50 - 47= 3. Note that the class width is not given by UCL and LCL Thus the width is not 49 - 47 = 2,

Mid-value (Class mark): It is the central value of a class. It is the average of lower class limit (LCL) and upper class limit (UCL) or lower class boundary (LCB) and upper class boundary (UCB). That is,

$$\text{Midvalue} = (\text{LCL} + \text{UCL})/2 = (\text{LCB} + \text{UCB})/2$$

Generally class marks are denoted by small alphabet (of variable). Some authors do report frequencies in terms of relative frequencies.

Relative frequency: Relative frequency of a class is given by **f / N** where **f** denotes frequency of a class and **N** denotes total frequency. The relative frequencies, when multiplied by 100 are called percent relative frequencies.

Frequency density: Number of units (f) in a class per unit length of class is known as frequency density. It is the ratio of frequency of a class to width of a class. The frequency densities are useful while drawing histogram when the data are reported in classes of unequal width.

Following table shows classes, class boundaries, frequencies, class marks, relative frequencies, percent relative frequencies and frequency densities for the frequency distribution of weights of adults.

Table 2.13: Frequency distribution of weights of 100 adults

Weight in kg.	*Class boundaries*	*Class Mark (x)*	*Frequency f*	*Relative Frequency f/N*	*% relative frequency (f/N)*100*	*Frequency Density (f/h)*
47 – 49	46.5 – 49.5	48	3	0.03	3	1.00
50 – 52	49.5 – 52.5	51	6	0.06	6	2.00
53 – 55	52.5 – 55.5	54	9	0.09	9	3 .00
56 – 58	55.5 – 58.5	57	15	0.15	15	5.00
59 – 61	58.5 – 61.5	60	27	0.27	27	9.00
62 – 64	61.5 – 64.5	63	20	0.20	20	6.67
65 – 67	64.5 – 67.5	66	13	0.13	13	4.33
68 – 70	67.6 – 70.5	69	7	0.07	7	2.33
Total			N =100			

Preparing a frequency distribution is a process of summarisation of data. Such a summarisation certainly leads to some "loss of information". For example: (1) the particular order in which the raw data was observed is lost. (2) Knowing the frequency of a class we know only how many observations in the data belong to that class but we do not know their individual values. Quite often this lost information has no significant relevance to the conclusions. Since Statistics is a science of averages the particular individual value is of no relevance. The conclusions drawn are for the sample or the population as a whole and not for any individuals.

Remark: We saw that the method of grouping the data is very useful. It has its own advantages. For example, we can quickly see that how various values are distributed. In general, the frequency distribution is represented by $\{x_i, f_i\}$, i = 1, 2, …, k; x_i may be the single valued classes or the class marks in case of interval type frequency distribution and f_i is the frequency of that class.

2.6 Cumulative Frequency Distribution

When raw data is condensed in the form of frequency distribution, what sort of information we generate? The frequency distribution informs us how many values of the characteristics being studied belong to a particular class. But this may not suffice. You may be interested in knowing number of observations below or above a particular value. Can we answer such questions reasonably? Yes, we can. We therefore introduce terms such as cumulative frequency and cumulative frequency distribution. Refer the data (Table 2.13) on frequency distribution of weights of adults. We can answer questions such as, "How many adults have their weights between 49.5 and 64.5?" The exact number is available by adding or cumulating frequencies of the classes representing this range of weights. Another way of obtaining such cumulative frequencies is to include one more column in the frequency table of cumulative frequencies showing the number of items (units/ individuals) with values less than or equal to the upper class boundaries. Such cumulative frequencies are called **less than cumulative frequencies** denoted by **CF↓** or **CFLT.** These frequencies are obtained by adding the frequencies from above. The cumulative frequency of the last class is the total of all frequencies. Similarly, the number of items with values more than a given value, generally the lower class boundaries, are called **more than cumulative frequencies** denoted by **CF↑** or **CFMT**. These frequencies are obtained by adding the frequencies from below. The set of cumulative frequencies together with relevant class boundaries is called cumulative frequency distribution.

Table 2.14: Cumulative frequency distribution of weights of 100 adults

Weight in kg.	*Class boundaries*	***Frequency f***	***CFLT (CF↓)***	***CFM (CF↑)***
47 – 49	46.5 – 49.5	3	3	100
50 – 52	49.5 – 52.5	6	9	97
53 – 55	52.5 – 55.5	9	18	91
56 – 58	55.5 – 58.5	15	33	82
59 – 61	58.5 – 61.5	27	60	67
62 – 64	61.5 – 64.5	20	80	40
65 – 67	64.5 – 67.5	13	93	20
68 – 70	67.5 – 70.5	7	100	7

Some times, instead of giving frequency distribution, the cumulative frequency distribution of either less than type or more than type is given. They are represented as follows:

Table 2.15: Cumulative Frequency Distribution

Less than cumulative frequency distribution		More than cumulative frequency distribution	
UCB	CFLT (CF↓)	LCB	CFMT (CF↑)
Below 49.5	3	Above 46.5	100
Below 52.5	9	Above 49.5	97
Below 55.5	18	Above 52.5	91
Below 58.5	33	Above 55.5	82
Below 61.5	60	Above 58.5	67
Below 64.5	80	Above 61.5	40
Below 67.5	93	Above 64.5	20
Below 70.5	100	Above 67.5	7

Given cumulative frequency distribution, frequency distribution can be formed and vice versa.

2.7 Cumulative Frequency Curve (Ogive Curves)

We can use cumulative frequency distribution to draw cumulative frequency curve, also called as ogive curve. There are two types of cumulative frequencies and hence two types of cumulative frequency curves. There are many situations that require information on number of observations below or above a certain specific value. For example, an enquiry can be made about what is the percentage of students who have scored marks less than 35 or more than 90. Answers to such questions can be given using graphs. We therefore study concept of cumulative frequency and cumulative frequency curve. To draw a less than type ogive curve, plot CFLT type on Y-axis against UCB on X-axis. Now join the points by a smooth line. We get less than cumulative frequency curve. This curve is S-shaped curve. If we plot CFMT on Y-axis against LCB on X-axis, we get more than type curve. Study the Figure 2.1 below. It is possible to draw either type of curve separately or both the curves can be drawn on the same graph paper. These curves will intersect (meet) at a point. This point is important. (If you draw a perpendicular from this point on X-axis, you will get median value – the value below which and above which exactly 50% observations lie. We will study median as a measure of central tendency in next unit). These curves are of use. We will discuss more on it in next unit.

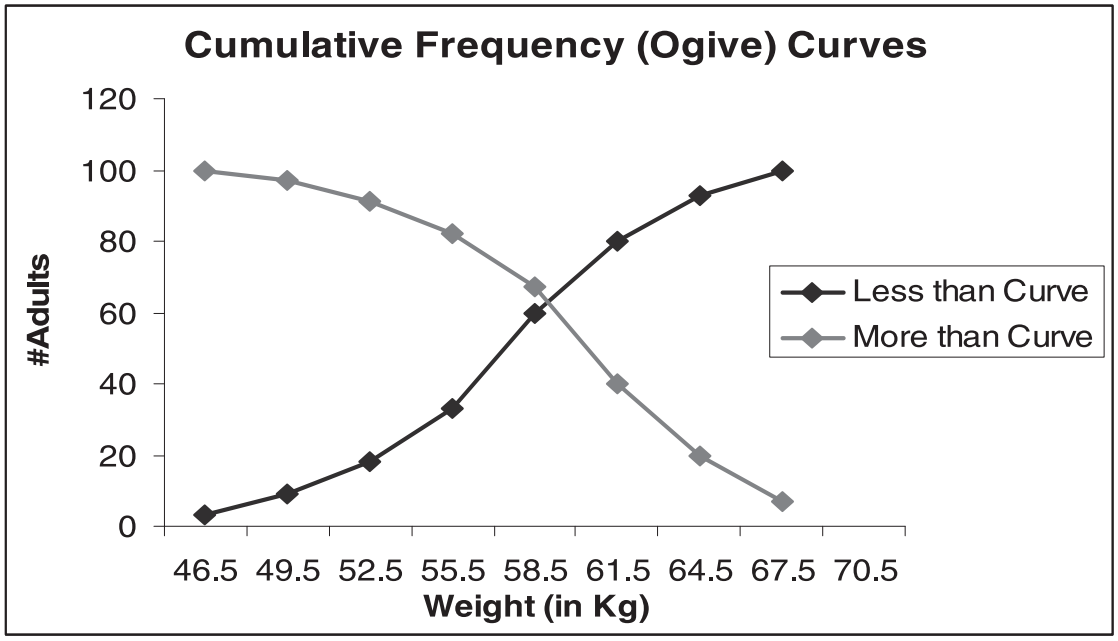


Figure 2.1: Cumulative frequency curves for weight data

2.8 Summary

In this unit we have understood some of the basic terms, such as primary and secondary data, raw (ungrouped) data and grouped data, used in Statistics. We also studied methods of presenting data in tabular forms. We studied structure of a statistical table and one way, two way and three way tables. We then studied various methods of classification. We further studied cumulative frequency distribution and methods of drawing ogive curves.

2.9 Self Test

1. Take 500 gm of peas with pods and prepare a frequency distribution of number of peas in a pod. Also prepare appropriate cumulative frequency distribution.

2. Read a table in a local newspaper. Identify the parts of the table in it.

3. Give a sketch of a blank statistical table and label its parts.

4. In 1980--81, the total production of principal oil--seeds (in '000 tons) in India was as follows: groundnuts 3702, linseed 434, mustard 1103, castor 105, seas mum 433. Next year the production each of first three items increased by 36% and of the remaining items by 10% and 12% respectively. In 1982-83, there was an increase compared to preceding year of 8% in groundnuts, 12% in linseed, 10% in mustard, 50% in castor and 10% in seas mum. In the next year the figures were respectively 5558.5, 622, 1502.1, 208.25 and 547.5. Tabulate the above information in an appropriate two-way table.

5. (Choose the correct alternative) A sample survey is conducted on family size of 25 households. These data are given below. Study the data and answer following questions:

Data on Family Size

2	7	2	2	5	3	4	3	5	4	7	4	6
3	4	5	5	3	4	6	6	4	6	4	2	

(i) Frequency of family size 5 is ---

a. 4 **b.** 5 **c.** 2 **d.** 25

(ii) # households of family size less than 5 is ----

a. 20 **b.** 15 **c.** 5 **d.** 9

(iii) The number of householders with maximum family size is ----

a. 2 **b.** 7 **c.** 4 **d.** 25

(iv) Total persons in all families is equal to ---

a. 99 **b.** 100 **c.** 98 **d.** 125

(v) Minimum family size of a household in population can be ---

a. 0 **b.** 1 **c.** 2 **d.** cannot be predicted.

6. Tabulate the following information:

The number of students in a college in the year 1998 was 510 of these 480 were boys and the rest girls. In 2003, the number of boys increased by 100 and that of girls increased by 300 as compared to their strengths in 1998. In 2003, the total number of students in the college was 1200, the number of boys being double the number of girls.

7. A classification of the population of India by livelihood categories (agricultural and non-agricultural) according to 1951 census showed that out of a total of 3,56,628 thousand persons, 2,49,075 thousand persons belonged to agricultural category. In the agricultural category 71,049 thousand persons were self-supporting; 31,069 thousand were earning dependents and rest were non-earning. The number of non-earning persons and self-supporting persons in the non-agricultural category were 67,335 thousand and 33,350 thousand respectively. The others were earning dependents. Tabulate the above information.

8. Tabulate the following information.

Candidates appearing at the F. Y. Commerce examination can offer English either at Higher Level or at Lower Level. They have to offer either Geography or Mathematics as an optional subject. Out of 1,000 candidates from College ABC, 45% have offered English at Higher Level. 80% of the candidates with English at Lower Level have offered Mathematics. The difference between numbers of candidates offering Mathematics and Geography in the entire college is 600. In College XYZ there are 1500 candidates of whom equal numbers have offered Mathematics and Geography. The number of candidates with the combination English at Higher Level and mathematics is 200, which is 25% of the number of candidates offering English at Lower Level.

9. Explain the terms (i) class limits and class boundaries (iii) class width (iv) frequency (v) frequency density and (vi) relative frequency.

10. Write a note on inclusive and exclusive methods of frequency distribution and illustrate it with suitable data..

11. Prepare a frequency distribution of data on shoe size

4	9	8	9	8	9	9	6	5	4	8	4	6	4	8	6
8	5	4	4	5	5	8	10	9	6	4	5	10	6	8	9
4	8	5	7	9	4	8	8	10	5	4	8	9	7	4	4
7	6	5	5	6	6	8	9	8	7	4	5	10	6	8	9
4	5	7	8	8	7	8	8	7	5	4	8	9	7	4	4

12. During a study on health status of tribes living in a river basin, researchers collected data on their body weights. Prepare an appropriate frequency table. These data are given below.

Weight of tribal adult (in kg)

46.11	75.25	63.12	71.99	47.56	42.15	31.78	45.07	65.94	34.00
42.55	65.11	67.57	52.01	53.37	51.73	32.49	37.52	46.19	58.64
54.61	57.69	36.78	56.08	71.08	42.45	62.61	51.52	55.72	46.36
36.81	45.19	53.25	58.04	46.86	41.52	57.50	47.08	62.50	76.85
33.93	68.50	40.95	48.76	41.01	54.92	42.70	64.53	56.49	56.47
53.93	56.53	52.16	40.86	45.14	55.85	40.61	50.70	55.20	41.77
49.17	45.22	38.53	53.07	45.60	40.19	27.88	54.25	49.14	68.47
48.57	40.67	44.12	42.15	6.07	60.53	50.91	58.14	75.06	72.08
48.12	61.30	49.56	62.79	59.21	38.36	43.45	36.62	28.76	43.88
42.83	47.01	45.85	53.68	48.55	53.65	44.02	52.87	45.69	53.60

13. Following is an extract of the data on internal marks (out of 10), in a unit test, secured by F.Y. B. Sc. students of a college affiliated to University of Pune. Prepare appropriate frequency distribution and write your findings.

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

Unit 3 : Charts and Diagrams

3.1 Overview

In this unit we discuss the graphical representation of data. Graphical representation also helps in comparing two or more data sets. Presenting data in graphs and charts is an art. The same data can be presented in two different ways using the same chart, but one of them turns out to be better. In this unit we will learn some basic charts which are extensively used in presenting statistical data.

3. 2 Learning Objectives

After studying this unit you will be able to –

- Know the types of charts,
- Draw appropriate charts,
- Read the charts.

3.3 Introduction

In the last two units we have studied what statistics is all about, the different types of data and methods to represent these data. We can also draw graphs, charts and pictures of the data. Many times the visual impact is more impressive than either the textual representation or tabular presentation of data. In this chapter we will study some selected charts that are used extensively in data analysis. A graphical presentation of data is very common. It is one more method of presenting data. We will give below some illustrative graphs. It is necessary to have a title for the chart drawn and also index keys. If uncommon abbreviations are used in charts then explain the same. We will study following charts:

(1) Pie Chart

(2) Simple Bar Diagram

(3) Multiple Bar Diagram

(4) Component Bar Diagram

(5) Percentage Bar Diagram

(6) Histogram

(7) Frequency Polygon and Line graph

3.3.1 Pie Chart

Pie chart is a circular diagram. Drawing a pie chart manually is tedious, but on a computer you can make use of graphics packages or Excel. We Study the method of drawing pie chart with the help of some data. One such example is given below. Here the population under consideration is subdivided into non-over lapping components. Draw a circle with appropriate radius. Next partition the circle into segments (slices) so that, the size of each component is the percentage of the total represented by the category the component denotes. The natural query would be how to measure an area of each component. The circle has 360^0. So, decide angle of a sector in proportion to percentage share of the component. The appropriate angle thus can be worked out for each component.

Table 3.1: Distribution of patients according to type of disease

Disease ▶	G. I.	Chest	E. N. T.	Diabetes	Heart	Total
Number	***1200***	***260***	***400***	***700***	***50***	***2670***
Percentage Share	***47***	***10***	***15***	***26***	***2***	***100***
Angle (in ^{0})	***169.9***	***35.1***	***53.9***	***94.4***	***6.7***	***360***

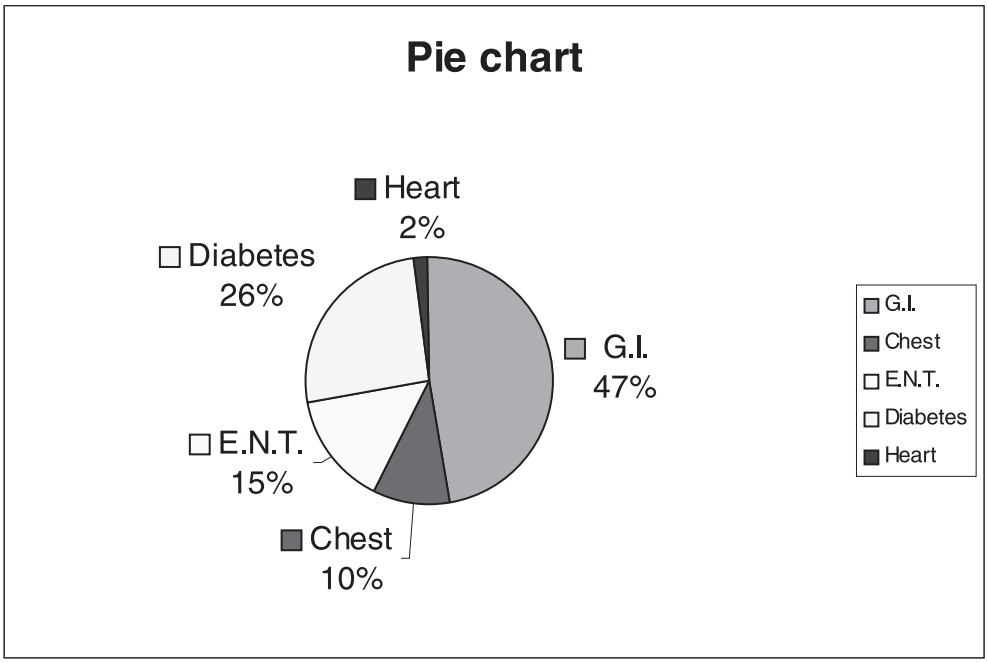


Figure 3.1: Pie chart of data in Table 3.1

3.3.2 Simple Bar Diagram

The data on attributes such as sex, profession etc, are graphically represented by a simple bar diagram. In simple bar diagram draw vertical bars of equal width, one bar corresponding to each value of the attribute. The heights of these bars must be proportional to the frequency of the corresponding value. All the bars must stand on the same base line. You have to be careful while choosing the scales. Sometimes these bars are also drawn horizontally. Study bar-diagram -1. This is a bar diagram of the number of suicides reported in Pune city as reported in a local newspaper in 2001.

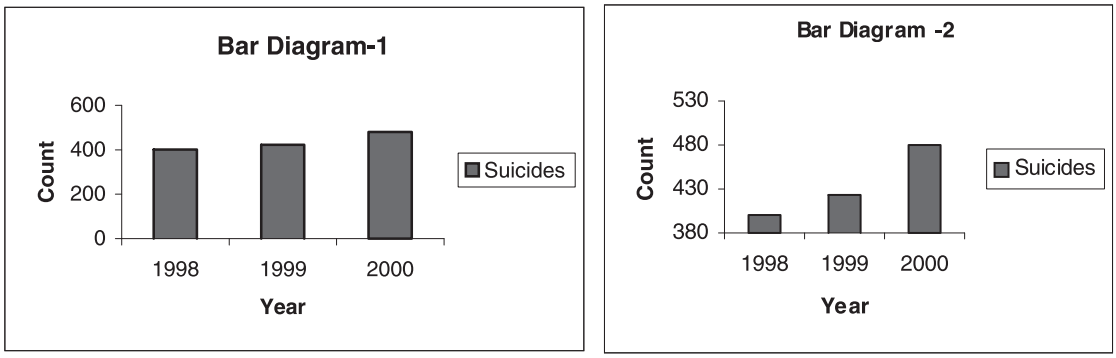


Figure 3.2: Bar-diagrams

In Bar-diagram-1, notice that along y-axis, we have given the scale for the frequencies starting from 0. Sometimes when all the frequencies in the data are large, the zero of the y-axis is shifted to an appropriate large number. Thus in Bar-diagram-2 for the same data the zero on the y-axis is shifted to 380. Note that this gives a dramatically different visual effect to the same data. Both diagrams show the increase in the number of suicides over the three years. However in diagram two this increase looks dramatically large than what it really is. **This constitutes a misuse of the graphical representation.**

3.3.3 Multiple Bar Diagram

Assume that you have data one attribute. If your interest is to compare the two populations **for the same attribute** then think of multiple bar diagram. It is thus an extension of simple bar diagram. For each value of the attribute place the bars corresponding to the two populations side by side. Study data in Table 3.2 where the attribute of interest is the total number of accidents in an year per 200 km, in two neighbouring states A and B.

Table 3.2: Number of accidents in different years on national highways

Year → *State* ↓	1999	2001	2003
A	150	170	230
B	160	200	240

Number of accidents in two states on national highways

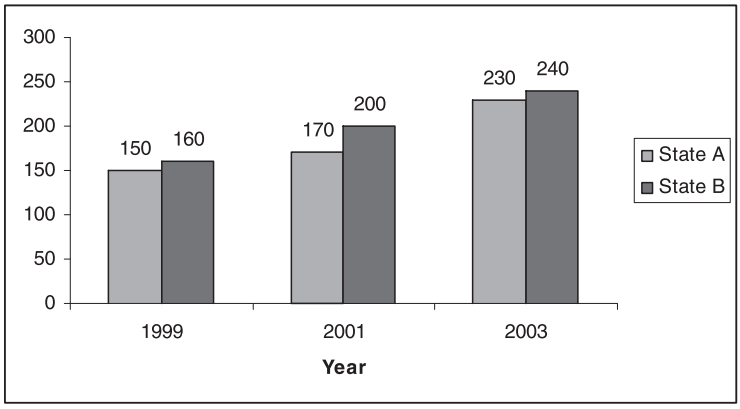


Figure3.3: Multiple bar diagram

You can compare these figures but do it with a pinch of salt. One can raise a series of valid questions.

3.3.4 Component Bar Diagram

When the multi-character data possesses additional features we may recommend component bar diagram. Here instead of placing bars side by side place them on top of the other. Consider the data in Table 3.3

Table 3.3: Distribution of Bacterial Population (in10^5/g of soil)

	Sugarcane	Eucalyptus	Grape
Vegetative stage	150	140	80
Flowering stage	120	100	60
Seed-setting stage	60	80	70

Its component bar diagram is shown below.

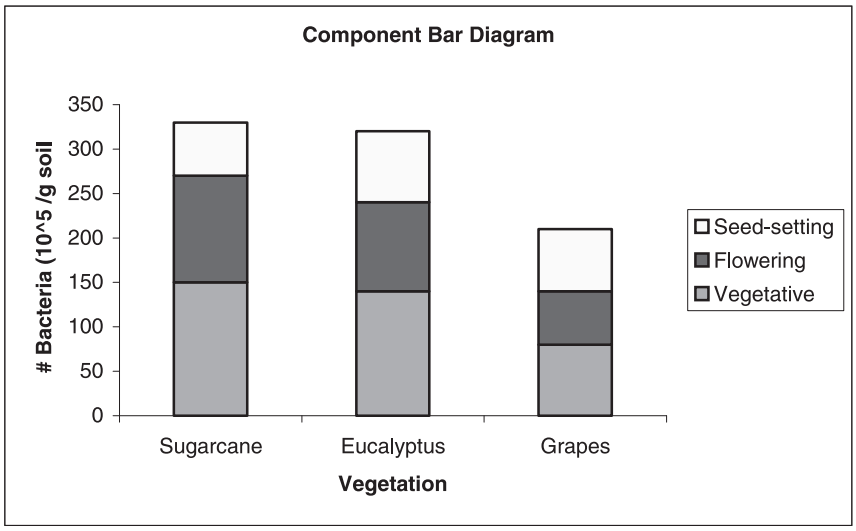


Figure 3.4: Component bar diagram of bacterial population

3.3.5 Percentage Bar Diagram

If numbers to be compared in component bar diagram vary considerably then convert them into percentages and draw component bar diagram for these percentages. Such a diagram is called percentage bar diagram.

3.3.6 Histogram

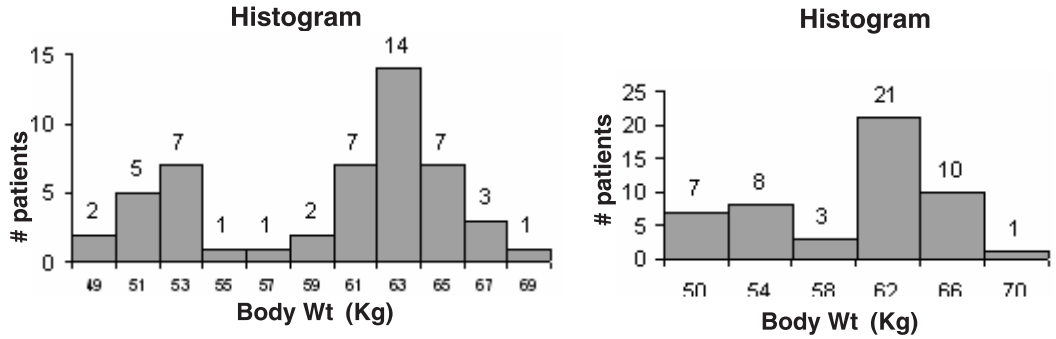
When the raw data are classified based on class-intervals and data are condensed in the form of frequency distribution then use of histogram is justified. It is useful when each group consists of a number of measurements-that are made on continuous type of data. Usually the raw data is classified into intervals of equal width and the number of values belonging to each sub-interval is noted. These are then represented by a rectangle with height proportional to the frequency of that subinterval. More the number of sub-intervals less would be the count in each subinterval. Never emphasize unnecessary minor fluctuations by creating more intervals. If there are too few subintervals then the true variability in the data will remain unnoticed. In histogram it is true that we lose track of individual observations and the choice of subintervals is subjective. Nevertheless histogram is most popular and powerful tool used in a graphical presentation of data. To know more about it consider following data. We draw two histograms for the same data. Drawing of histograms can be done using computer packages. We describe here the procedure of

drawing histogram manually. First mark the points corresponding to various class-limits by using a convenient scale on the x-axis. Then choose a suitable scale along the y-axis and erect columns on the segments between two consecutive limits of the class-intervals. The height of the column on the segment between l_{i-1} and l_i s is equal to f_i (frequency of i^{th} class interval. The diagram is called histogram.

Table 3.4: Body weight (in kg) of 50 heart patients

49.0	49.5	50.5	51.0	51.0	52.0	52.0	53.5	53.5	54.0
60.0	60.5	61.0	61.0	61.5	62.5	62.5	63.0	63.0	63.0
64.5	64.5	64.5	65.0	65.0	51.0	51.0	52.0	52.0	52.0
57.5	58.0	59.5	60.0	60.0	62.0	62.0	62.0	62.5	62.5
63.0	63.5	63.5	63.5	64.0	65.0	66.0	66.5	67.0	69.0

Notice the difference between **Figure 3.5 A** and **B** histogram shown below. The difference has roots in number of intervals formed. There are no fixed rules of deciding number of class intervals. If the experimenter has some information on the nature of the variable being studied then use the information even while preparing histogram.



(A) **(B)**

Figure 3.5: Histograms for the data given in Table 3.4

We have earlier remarked that sub-intervals must be of equal width. But it may not be always possible and perhaps not desirable. If such is a case then plot frequency density $\frac{f}{h} = \frac{frequency}{width}$ of each class against the class-interval and draw the histogram.

3.3.7 Frequency Polygon and Line Graph

The data for histogram can also be used for drawing a frequency polygon. The drawing of frequency polygon is simpler and quicker. It is mainly used to compare two or more frequency distributions. In frequency polygon frequency is plotted at the midpoint of the class-interval of the corresponding rectangle in a histogram. A straight line is drawn to join the successive midpoints (class marks) marked in consecutive rectangles. Since polygon is a closed curve, it is customary to include a class-interval (of the same width) at the beginning and end, so that we can draw it. The curve we get is called a **frequency polygon.**

A **line graph** is used to represent a trend over a period of time. You may also use it for continuous type of data. There are many situations wherein the data may be taken repeatedly over time.

Example: We give below extract of the data collected in socio-economic survey of a tribal community residing in Barhanpur, M.P. These data were collected and analyzed by Dr. N. C. Gujarathi and M. B. Kulkarni in 2004.

Table 3.5: Age and sex distribution of the villagers in Barhanpur

Age (in years)	*Males*	*Females*
1- 5	*72*	*80*
5 – 10	*76*	*64*
10 – 15	*72*	*60*
15 – 20	*45*	*42*
20 – 40	*150*	*140*
40 – 60	*100*	*80*
60 – 80	*30*	*40*
Total	*545*	*506*

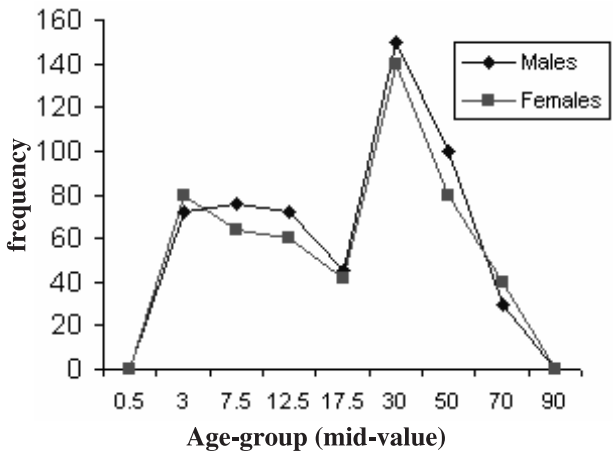


Figure 3.6: Frequency polygon

Example: The data on average height of Indian and U.S.A. students is reported in the **Table 3.5**. Its line graph is shown in **Figure 3.7**.

Table 3.5:
Average height (in cm) of an Indian and U.S.A. child

Age in months	Indian	U.S.A.
At birth	30	31
6	40	42
12	55	60
24	70	76
36	90	98
48	100	108
60	115	124
120	130	140
240	165	173

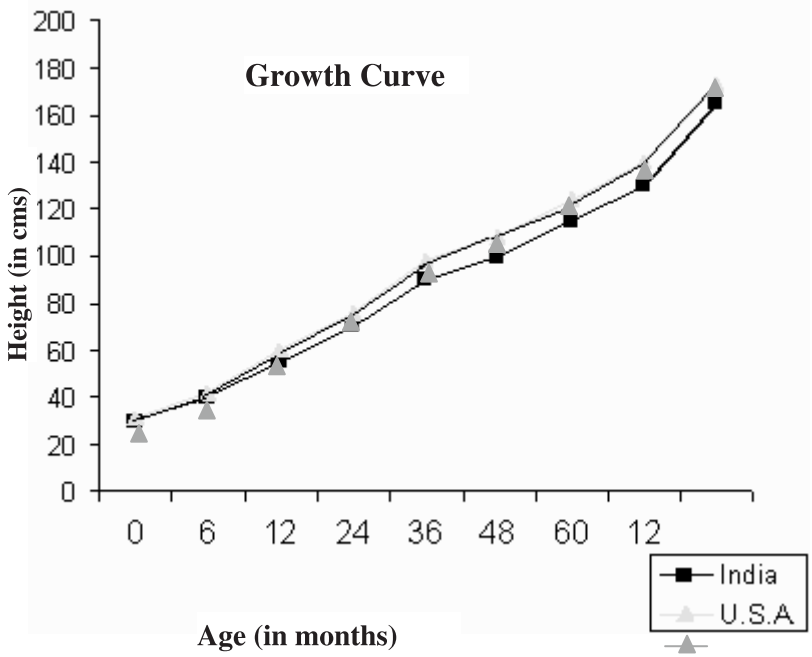


Figure 3.7: Line graph

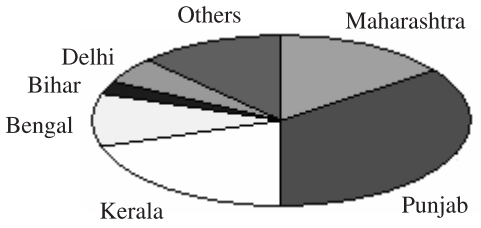
3.4 Summary

In this unit we have studied various types of charts and have understood its usefulness in data analysis. studied its usefulness . Drawing of a chart given appropriate data can certainly be handled using various software's. It is the researcher who has to the responsibility to make proper choice of the chart.

3.5 Self Study

1. Study the following data and chart and choose the correct alternative

State	Maharashtra	Punjab	Kerala	Bengal	Bihar	Delhi	Other
# Tourists	300	700	400	200	50	100	250



Pie chart of Tourists

Total number of tourists from Maharashtra and Punjab is equal to ---

(a) 300 (b) 700 (c) 1000 (d) 2000

(i) Percent number of tourists from Punjab is approximately equal to ---

(a) 70 (b) 35 (c) 50 (d) 700

(ii) Percent number of tourists from states other than Kerala is ---

(a) 20 (b) 80 (c) 40 (d) 400

(iii) Angle (in degrees) in Pie diagram for Kerala is approximately equal to ---

(a) 72 (b) 298 (c) 20 (d) 36

(iv) # Tourists from Kerala and Punjab is ---

(a) More than all other states combined. (b) Less than all other states,

(c) Exactly equal to all other states. (d) More than 50% of all the tourists.

2. Draw appropriate graphs and comment on it.

a. Obtain the data on Hemoglobin level (in gm %) of 100 normal and diabetic individuals as per their sex..

b. Obtain ABO blood group distribution of 40 individuals each of different tribes.

c. Collect anthropometrical data on infants/children belonging to different socio-economic groups. Compare it.

d. Obtain distribution of monthly expenditure on different commodities from different economic groups and compare them.

3. Forest area (in million hectares) for some states of India are recorded below:

States	Forest area (in million Hectare)
Maharashtra	4,0
M.P	10.8
Karnatak	3.0
Orissa	4.8
A.P.	4.9
Meghalaya	2.0
Rajasthan	1.6

4. In a village consisting of 95 families, 70 families use wood as a fuel whereas 20 use either rock oil or gober gas or both and remaining use of LPG gas. Represent these data by an appropriate chart.

5. Distribution of 10,000 individuals as per their educational status is given below. Draw appropriate diagram.

Educational Level	# of Individuals
Illiterates	6,000
UPTO IV std	1,500
UPTO VII std	1,000
Upto HSC	800
Graduate	500
Post graduate	200

6. In different zones of India 88 Elephant corridors are identified as given below:

North-East	22
Southern India	20
Central India	20
Northern West Bengal	14
North Western India	12

Represent the above data by a suitable diagram.

7. Monthly variation in wind velocity in a dry zone area is given below:

Month:	1	2	3	4	5	6
Wind velocity m/s	2	2	2	2	5	4.5
Month	7	8	9	10	11	12
Wind velocity m/s	4.2	3	2	1.8	2.2	2.2

8. Income generated per hectare of forest in different forest zones is reported below. Draw appropriate chart.

Income in Rs/hectare/year	**Zone**
1800	South
600	North
1400	East
1000	West

Unit 4 : Measures of Location (Central Tendency)

4.1 Overview

The methods of data representation as given in units 2 and 3 give the overall picture of the data. However many times we need numerical descriptive measure to know more about the distribution of the variable. The measures should be useful for user for understanding the tendency of the data and they must also be useful for inferential procedure. Here we discuss three popular measures of location of the data (also called as measures of central tendency of the data), viz, arithmetic mean (A.M.), median and mode.

4.2 Learning Objectives

After studying this unit, you will be able to

- Understand meaning of the term average
- Choose appropriate average amongst mean, median and mode
- Calculate average(s) for ungrouped (raw) and grouped data
- Interpret the average
- Use graphical techniques to estimate median and mode
- Understand limitations of the average

4.3 Introduction

In our daily life we come across number of situations wherein we make use of averages knowingly or unknowingly. For example, proprietor of a cafeteria keeps the daily record of number of customers arriving at it. By looking at the data at the end of month he can roughly conclude the average number of customers to whom he is giving service. These records are useful to determine the business policy. Information on the average number of customers arriving at ATM centers is necessary for the bankers. In the present education system a student's performance is judged by average marks scored in the final examination. Performance of a motor-cycle or a car is judged by the average mileage (in km) per liter of petrol. A person interested in studying a reasonable problem has some questions in mind. In order to facilitate objective understanding of the problem, relevant data are collected, either generated by experiments or collected from various resources. She then looks for appropriate statistical techniques. Any conclusion drawn from such data must be acceptable to a "common sense". In order to make a sound assessment a researcher examines the data. We have seen earlier that the data can be displayed in the form of tables, graphs or charts. A picture is worth a thousand words. But it has also limitations. For example, it may not give precise answers or it cannot be used for statistical inference. Since mass of data is difficult to comprehend, summary statistics is a must. Summary statistics indicate location, spread and symmetry of the data. To begin with we introduce popular measures of location that are routinely used. We will briefly review and study them here.

4.4 Measures of Central Tendency

What is average? Many of us believe that we know lot about it. Everybody knows that if you want to find average, simply add all the observations and then divide the sum by number of observations to get what is known as arithmetic mean (A.M) or simply mean. You perhaps know the formulae that are used to find A.M. We will learn more about different averages.

Whenever a sample survey is conducted or an experiment is performed we get data on the observed units belonging to the population under consideration. By the term population we mean the variable under study. For example, we may be interested in estimating the average expenditure incurred on food items or knowing average weight of an Indian baby at birth. A botanist may be interested in estimating average sepal length or width of a flower species, whereas a microbiologist may be interested in estimating average size of a cell. A computer professional may be interested in estimating average time required to compile a program on various systems. In such studies we invariably study only part of the population. This part of the population on which the data are observed is called a sample. Since the population characteristics (parameters) are usually unknown, our aim would be to estimate these parameters using sample values. Therefore it is prudent to have a representative sample. Hence statisticians insist upon "randomness" of the sample. If the sample is random then it can help us to study the properties of the population. It is possible to devise many types of averages. What are the desirable characteristics of an average? First of all it must represent the data properly. It must be based upon all observations and addition or deletion of few observations should not affect it seriously. Moreover it must be capable of further mathematical treatment.

4.4.1 Mean

One of the most common and useful measures of central tendency is the arithmetic average of a set of measurements. In routine language we refer to it as mean or arithmetic mean (A.M). We usually have sample information, so we obtain sample mean $\bar{x}$ and use it as an estimate of population mean μ (Greek letter mu). It is the center of a set of sample measurements (analogues to center of gravity used in physics where a unit mass is put at each data value). If at all population mean μ is known then we do not need sample mean. In statistical inference we plug the value of sample mean in place of unknown population mean. **Note that while calculating the mean, all numbers must represent the value of the same variable under study.**

Definition: The arithmetic mean of a set of n measurements x_i (i =1, 2…n) is equal to the sum of n measurements divided by n.

For ungrouped data $A.M = \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = (x_1+x_2+x_3+\ldots+x_n)/n$ eq (4.1)

For discrete frequency distribution, where each value x_i occurs f_i times

$$A.M = \bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \quad \text{eq (4.2)}$$

In case of a continuous variable, where the data are available in the form of a frequency table (grouped data), find class-mark and frequency of each class. We will thus have pairs (x_i, f_i) for i = 1, 2... k classes. A.M for such grouped data is given by

$$A.M = \bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = (f_1x_1+ f_2x_2+ \ldots+ f_kx_k) / (f_1+ f_2 + \ldots+ f_k) \quad \text{eq (4.3)}$$

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

4.4.2 Merits and Demerits of Arithmetic Mean

Merits:

- It is precisely defined,
- It is easy to calculate and simple to understand,
- It is based on all the observations,
- It gives equal importance to all the observations,
- It is preferred for statistical analysis because of its mathematical properties,
- It possesses sampling stability, i.e. it is lest affected by sampling fluctuations,

Demerits:

- A too large or too small observation will seriously affect arithmetic mean,
- A.M cannot be determined graphically,
- It cannot be calculated if you have open-end classes such as less than 1000 or more than 10,000.

Example: Ungrouped data

When drugs are recommended to a patient, one of the important quality characteristics of the drug is its disintegrating time. An experiment was conducted in laboratory to estimate disintegrating time (in seconds) of four different but equivalent drugs, say A, B, C and D. We can compute arithmetic mean of disintegrating time for each of the drug using following data. A drug having least average disintegrating time is to be recommended for future work.

Table 4.1: Disintegrating time (in seconds) of drugs A, B, C and D

Sr. No	Drug A	Drug B	Drug C	Drug D
1	20	21	25	19
2	23	18	24	15
3	25	17	25	16
4	19	16	26	24
5	25	19	25	28
6	25	20	24	12
7	21	22	25	15
8	23	20	21	16
9	20	25		21
10	18			22
Total	**219**	**178**	**195**	**189**
A.M.	**219/10 = 21.90**	**178/9 = 19.78**	**195/8 = 24.38**	**189/10 = 18.90**

Thus the disintegrating time for drug D < drug B < drug A < drug C. On the basis of A.M. drug D appears to be better. (Do you agree with this conclusion? Think about it).

Note that combining the observations on all the four drugs and reporting the combined mean would not make much sense because each drug forms its own separate population.

Example: Discrete frequency distribution

A geneticist interested in studying number of seeds set in a pod collected following data. Using these data we will find the sample mean.

Table 4.2: Data on number of seeds set per pod

# Seeds/pod (x)	0	1	2	3	4	5	6	8	9	Total
# pods (f)	2	3	15	20	25	25	20	5	5	120
f*x	0	3	30	60	100	125	120	40	45	553

Here sample mean = $\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$ =553 / 120 = 4.61 (approx)

It may happen that sample mean is not even a possible value of the variable under study. It is a mathematical quantity used as a summary statistic.

Example: Grouped data

Following are the extracts from a nutritionist's survey aimed at estimating malnutrition in an area. It gives distribution of height (in cm) of 100 students.

Table 4.3: Distribution of Height (in cm) of 100 students

Class interval (Height in cm)	*Number of students*
$145 < x \leq 149$	2
$149 < x \leq 153$	3
$153 < x \leq 157$	15
$157 < x \leq 161$	40
$161 < x \leq 165$	20
$165 < x \leq 169$	15
$169 < x \leq 173$	5

We first prepare following table and then find sample mean of these data. Recall the definition of a class-mark. It is a mid-point of a class-interval. For class 145-149, class-mark is (145+149)/ 2 = 294/2 = 147 and so on. Refer Table 4.4, for computation of mean.

Table 4.4: Computation of mean for grouped data

Class-marks (mid values x_i)	Frequency (f_i)	$f_i x_i$
147	2	294
151	3	453
155	15	2325
159	40	6360
163	20	3260
167	15	2505
171	5	855
Total	100	16,252

Using equation 4.3, $\bar{x} = \dfrac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = (f_1x_1 + f_2x_2 + \ldots + f_kx_k) / (f_1 + f_2 + \ldots + f_k) = 162.52$ cm

4.4.3 Some more Properties of A.M.

1. **Combined mean of two series:** Suppose we are given two data sets say A and B on the same variable. Let us assume that mean of A, based on m observations is $\bar{x}$ and of B, based on n observations is $\bar{y}$. Then mean of the combined data based on (m+n) observations is given by

$$\bar{x}_c = (m\,\bar{x} + n\,\bar{y}) / (m + n)$$

Suppose in a class consisting of 60 boys and 40 girls, aptitude test is conducted. Mean marks scored by Boys are 70 and that of girls are 62. Our interest is to report mean marks of the whole class. Using formula for combined mean, mean marks of the class would be

$$\bar{x}_c = (m\,\bar{x} + n\,\bar{y}) / (m + n) = (60 \times 70 + 40 \times 62) / 100$$
$$= 6680 / 100 = 66.8 \text{ Marks}$$

The above formula can be generalized for 3 or more data sets. This is left as an exercise.

2. **Change of origin and scale property:** Let us define new variable $\mathbf{U = (X - a) / h}$, where $h \neq 0$. Then we can obtain mean of X by the relationship $\bar{X} = a + h\,\bar{U}$

If h = 1, we have U = X- a, it reduces to a change of origin property. Similarly, if a = 0 and $h \neq 0$ then we have change of scale property. When to use change of origin and scale property? If values of X are too large or too small then use of this property is strongly recommended. Note that choice of h and a is subjective. However, for ease of calculations, choose a close to mid-value of the distribution and let h be the class-width. We will make use of this property to find mean of the variable in the following example.

Example: Following is the frequency distribution of waiting time (in seconds) of 200 customers in a cyber cafe.

Waiting-time (In seconds)	0-20	20-40	40-60	60-80	80-100	100-120	Total
# Customers	20	30	70	50	20	10	200

To find mean waiting time, prepare following Table 4.5 (U = (X – 60) /20)

Table 4.5: Mean using change of origin and scale formula

Class-interval	Midpoint (x)	Frequency (f)	U = (X – 60) /20	f *u
0 -20	10	20	-2.5	-50
20 -40	30	30	-1.5	-45
40 -60	50	70	-0.5	-35
60 -80	70	50	0.5	25
80 -100	90	20	1.5	30
100 -120	110	10	2.5	25
Total		200		-50

$\bar{u} = \sum fu / \sum f = -50 / 200 = -0.25 \Rightarrow \bar{x} = 60 + 20 \times (-0.25) = 55$ seconds.

Verify that $\bar{x} = \sum fx / \sum f = 11000 / 200 = 55$ seconds.

4.4.4 Median

A second measure of central tendency is the median. It is a **positional** average. It does not change even if there are extreme observations. It does not depend on the **exact** values of all the observations but only depends on their order of magnitude. It is in fact that value which divides the data in to two equal halves.

Definition: The median of a set of n measurements x_i (i = 1, 2 …n) is defined to be that value, below and above which 50% of observations fall.

Some authors denote median by $\tilde{\mathbf{x}}$ (called as x curl) or simply by **Me**.

1. Case of ungrouped data: Suppose x_i (i = 1, 2…n) are n observations given to us. In order to obtain median of such ungrouped data

1. Arrange all these observations in ascending (or descending) order,
2. Locate the middle observation(s),
3. If the number of observations (n) is odd equal to 2k+1 then the value of $(k+1)^{th}$ observation will be the median.
4. If n is even equal to 2k, then the mean of k^{th} and $(k+1)^{th}$ observation will be the median.

For example, median disintegrating time for drugs A, B, C, D of the data as given in table 4.1 are respectively 22, 20, 25, and 17.5. The relationship for median disintegrating time (in seconds) can now be stated as

Drug D < Drug B < Drug A < Drug C.

However these four medians are computed on samples of different sizes. How to account for these different sample sizes for drawing the conclusions for these data needs further study.

2. Case of discrete frequency distribution: To determine median for a discrete frequency distribution, follow the following steps:

1. First find less than cumulative frequency,
2. Find total frequency (N),
3. If N is odd equal to 2k+1, determine the value of $(k+1)^{th}$ observation.
4. If N is even equal to 2k, determine the mean of values of k^{th} and $(k+1)^{th}$ observation. It is Median.

For example, median for following data can be obtained.

Data on number of seeds set per pod

# seeds/pod (x)	0	1	2	3	4	5	6	8	9	Total
# Pods (f)	2	3	15	20	25	25	20	5	5	120
Less than C.F.	2	5	20	40	65	90	110	115	120	

Here N = 120 so k = 60 and k+1 = 61. Since both the 60^{th} and 61^{st} observations correspond to the value of x is 4 and hence median would be 4 seeds/pod.

3. Grouped data (continuous frequency distribution)

For a grouped data we first obtain the median class, i.e. the class that would include the median. Next we obtain the median value using the following interpolation formula:

$$Median = l_1 + \left[\frac{\frac{N}{2} - C_p}{f}\right]h \qquad \text{eq. 4.5}$$

where l_1 = the lower limit of median class

N = total frequency,

C_p= less than cumulative frequency of pre-median class,

f = frequency of median class,

h= class-width of median class,

Example: Data given below is on diameter of individual tree trunks measured at breast height (DBH), of trees belonging to a particular species. Forest is interested in knowing the distribution of the variable DBH, as it can be used to estimate biomass of a tree.

Table 4.6: Diameter at breast height of 80 trees

DBH (cm)	mid-point	Frequency	Cumulative frequency
10-14	12	11	11
14-18	16	20	31
18-22 (Median class)	20	30	61
22-26	24	15	76
26-30	28	4	80
Total		80	

Here N/2 = 40, which indicates that median class is 18-22,

Using equation 4.5, we get **Median** = 18 + [(40 – 31) / 30] x 4 = 19.20 cm

How to interpret median? For above data median is 19.20 cm. It means fifty percent of the trees have diameter at breast height less than 19.20 cm and 50% of them have diameter more than 19.20 cm. Thus median is a positional average.

4.4.5 Merits and Demerits of Median

Merits

- It is easy to understand and calculate,
- It is not affected by a few extreme values in the data,
- Even if frequency distribution has open-end classes (at either ends), median can be determined, In this case we do not have a middle value of the class and so mean cannot be found. However the median can be found as long as any open-ended interval does not include more then half the observations.
- It can be determined graphically (using ogive curves),

Demerits

- Median is not determined using all the values; it being a positional average it is least affected by changes in extreme observations. This can be looked upon as its demerit,
- It is well defined but may not be unique.
- Just knowing the medians of two data sets we cannot determine the median of the combined data.

4.4.6 Mode

Definition: The mode of a set of n measurements x_i (i = 1, 2…n) is defined to be the value of X that occurs most frequently.

It is thus easy to calculate. It may not be unique.

Case 1: Mode for ungrouped data

In case of ungrouped data reported in Table 4.1, mode for disintegrating time for drug A, B, C, and D is respectively 25, 20, 25, (15 and 16).

Notice that there are two modes 15 and 16 for the data on drug D. This is because both these values appear twice in the data whereas the each of the remaining values appears only once.

Case 2: Mode for discrete frequency distribution

Here we apply the definition of mode. Locate the value of the variable having maximum frequency and declare it as mode. For example, for the data in Table 4.2, both the values 4 and 5 occur with maximum frequency 25 and therefore this distribution is bimodal, mode being 4 and 5. Thus again we see that there are two modes. Note that mode may not exist.

Case 3: Mode for grouped data

In case of grouped data we first obtain the modal class having the maximal frequency and then use the following formula:

$$\mathbf{Mode} = \mathbf{l} + \frac{\mathbf{f_m} - \mathbf{f_0}}{2\mathbf{f_m} - \mathbf{f_o} - \mathbf{f_1}} * \mathbf{h} \qquad \text{eq 4.6}$$

where

l = lower limit of modal class,

f_m = frequency of modal class,

f_o = frequency of pre-modal class,

f_1= frequency of post-modal class, and

h is the class width of the modal class

For the data on diameter at breast height, mode = 18 + (10/25)*4 = 19.6 cm

4.4.7 Merits and Demerits of Mode

Merits:

- It is easy to understand and simple to calculate,
- For qualitative and also for quantitative data it can be defined,
- It is not affected by extreme observations,
- It can be obtained graphically (using histogram in case of grouped data),
- If there are open-end classes at either ends, mode can be computed as long as the open ended class **does not have the maximum frequency,**

Demerits:

- It is not rigidly defined,
- It may not be unique.
- It is not based on all the observations.

Note that in the above list, first five points stand for merits and next three for demerits of mode. Mean, median and mode are all equal when the distribution is symmetric. But it may happen that for non-symmetric data mean and median can be quite different from each other. It calls for a further study of the data. General guidelines can be stated at this juncture. If you have “enough” number of observations draw histogram. If the histogram indicates that the data are symmetric

then use mean, median or mode as a measure of central tendency. These measures will be close to each other. Does there exists a relationship between these three measures of central tendency? For a symmetric distribution, mean = mode = median. What happens if distribution is asymmetric? An empirical relation does exist. For a moderately asymmetric unimodal frequency distribution the relationship between these measures is

Mean – Mode = 3(Mean – Median) eq .4.7

Figure 4.1 : Bell-shape histogram

4.4.8 Graphical Presentation of Median and Mode

We have already stated that arithmetic mean cannot be determined graphically, but median and mode can be approximated using graphs.

(a) Graphical determination of median

In case of a grouped data, first prepare cumulative frequency distribution. Draw less than type cumulative frequency curve. Draw a line parallel to x-axis from point N/2 on y-axis. It will meet at a point on ogive curve drawn. From the point, drop a perpendicular on x-axis. The point at which the perpendicular and x-axis intersect is the value of the median. For following data set median is estimated graphically.

Table 4.7: Data on marks scored by 600 students

Marks less than	# Students
10	20
20	40
30	80
40	200
50	350
60	470
70	530
80	560
90	590
100	600

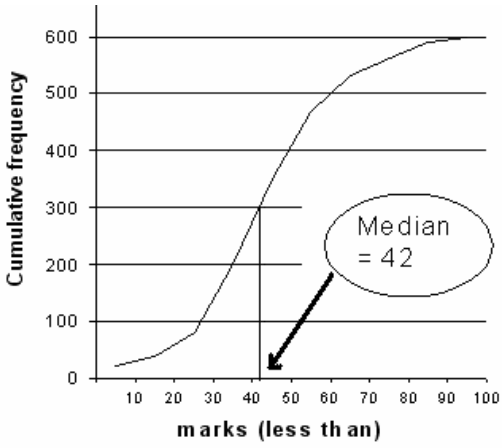


Figure 4.2: Median using less than ogive curve

If both ogive curves are drawn, then draw a perpendicular on x-axis from the point of their intersection, again you will get estimate of median. Again we draw both the ogive curves for the same data. We get same value of median (=42).

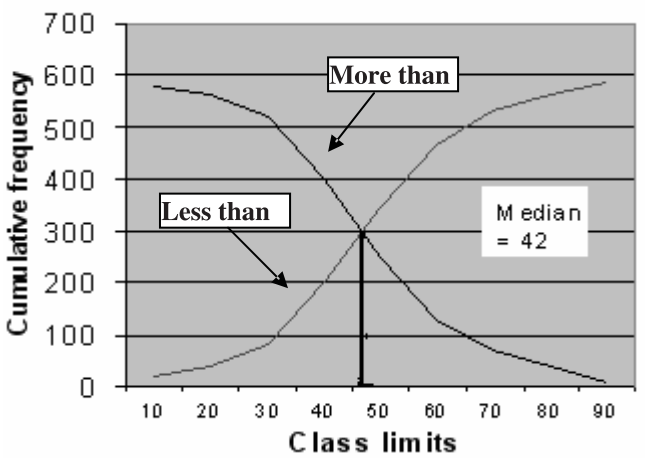
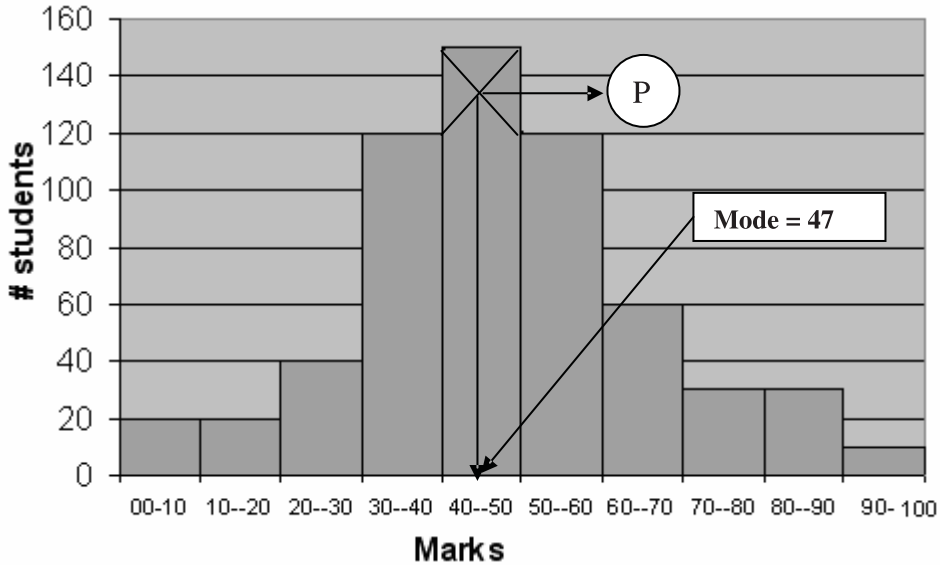


Figure 4.5: Median using both the ogive curves

(b) Graphical determination of mode

Mode can be determined graphically using histogram.

The first trivial step is to draw a histogram. Locate a class-interval with maximum frequency. If the class-interval with maximum frequency is at the end, graphical method fails. Study the histogram drawn below. In it a point "P" is determined by drawing two lines as shown in the graph. X-coordinate of point P is the value of mode. Draw a perpendicular from point P on x-axis to get an estimate of mode. For the data displayed in the histogram mode equal 47 approximately.



4.5 Other Averages

We will briefly mention few more averages. These are also used in various situations. For details of it students may refer reference books.

1. **Weighted Arithmetic Mean:** It is not always prudent to assign equal importance to all the observations. There may be situations where we need to assign appropriate weights to observations according to its importance. In such a situation use of weighted mean is advocated. Such sort of weighted averages are used in constructing index numbers.

   Definition: If there are k observations $x_1, x_2 \ldots x_k$ with corresponding weights $w_1, w_2 \ldots w_k$ then the weighted mean is given by $\bar{x}_w = \sum w_i \bar{x}_i / \sum w_i$.

   Example: The average yield of a paddy crop from two field stations is 50 and 62 quintals per hectare respectively. The averages were based on 40 and 20 hectares respectively. The area on which paddy crop is experimented varies considerably, hence we recommend use of the weighted average in this case. Here it would be

   $$\bar{x}_w = (50 \text{ x}40 + 60 \text{ x } 20) / 60 = 54 \text{ quintals per hectare.}$$

2. **Geometric Mean (G.M):** In our daily activities we may be interested in knowing average of bank interest rates, or average change in population growth, or average rate of returns on share, etc. If you have the data reported either in ratios or percentage or you have a geometric series then appropriate average would be geometric mean.

   Definition: If there are n observations $x_1, x_2 \ldots x_n$ then the geometric mean is given by $G = (x_1 x_2 \ldots x_n)^{(1/n)}$ = n^{th} root of $(x_1 x_2 \ldots x_n)$

   In simple words, it is the n^{th} root of product of individual values. To compute it, first find Log $G = \sum \log(x_i)$, find antilog to get G. Now-a-days scientific calculators can be used to find G.M directly.

   Example: Relative change in population for 4 consecutive years is 1.02, 1.04, 1.04 and 1.03. The average relative change in the population is given by use of G. M.

   $$G.M = (1.02 \text{ x } 1.04 \text{ x } 1.04 \text{ x } 1.03)^{0.25} = (1.147361)^{0.25} = 1.034964$$

   Average percent increase in population is (G.M. -1) x 100 = 3.4964

Merits and demerits of G. M:

Demerits:

- It is not easy to understand and calculate,
- It may not exist,
- It may reduce to zero, even if one observation is zero,
- It cannot be estimated graphically,

Merits

- It is rigidly defined,
- It is based on all observations.

3. **Harmonic Mean (H.M):** Harmonic mean is appropriate to determine average speed, average rates.

   Definition: The harmonic mean is the reciprocal of the mean of the reciprocal of values. It is thus given by **H.M. = $n / \sum(1 / x_i)$**

   Example: Suppose price of a quality A of the commodity is 10 units per rupee, whereas quality B of it is 20 units per rupee. If a shopkeeper sales the mixture of quality A and quality B that are mixed in equal proportions for 15 units per rupee, is he making extra profit? Use harmonic mean to answer such question. It is easy to verify than the price of mixture should be 2 / (0.1 + 0.05) = 2 / 0.15 = 13.33 units per rupee. It clearly shows that the shopkeeper is enjoying extra profits.

Merits and demerits of H.M are obvious.

Merits:

- It is rigidly defined
- It is based on all observations

Demerits

- It may not get defined if x = 0
- It is not simple to calculate and easy to comprehend.
- It cannot be estimated graphically.

4.6 Summary

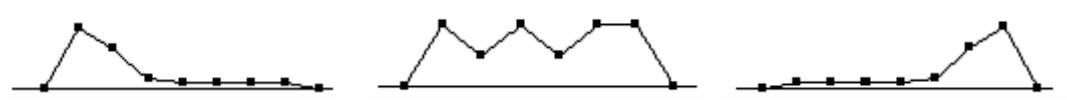
In this unit we have studied mean, median and mode as measures of location. Other measures of central tendency such as geometric mean (G.M) or harmonic mean (H.M) are also mentioned. These measures are also useful in certain situations. One of the frequently asked question is "which measure of location should I use?" The glib answer is – it depends upon study matter. Many times A.M. or median is appropriate. Interested students should refer reference books for more insight into all such measures. Earlier we have remarked that for a symmetric distribution mean = mode = median. Is converse true? No, it is not. The statement that "mean, median and mode occur in a dictionary order" is also not valid. Students should play with data and verify the validity of such statements. You can prove these statements either mathematically or disprove them by giving one counter example. Students can also verify that A.M. > G.M > H.M. Remember that analyzing data set is a skill. First study the data carefully. You may require editing the data or may doubt the validity of observations reported to you. Consult the subject expert if necessary and then proceed for further analysis. Computations of these measures can easily be done using computers. The skill of interpreting results and reading the data need to be developed.

4.7 Self Test

1. A researcher wants to graphically represent the relationship between month of the year and average number of suicides in a district. Which is the most suitable representation?

a. Ogive curve **b.** Pie chart **c.** Histogram **d.** line graph

2. Following are the frequency polygons for three groups. The X-axis is common to all.

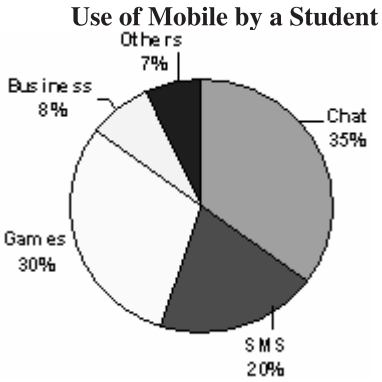


Which of the following statements is true?

a. Mean A > Mean B > Mean C **b.** Mean A < Mean B < Mean C

c. Mean A < Mean B > Mean C **d.** Mean A > Mean B < Mean C

3. The pie chart shows daily usage pattern of mobile phone by a student.



On a given day, if he has spent 7.5 minutes more in chatting than the time he spent in playing games, how much time would he have spent on sending SMS?

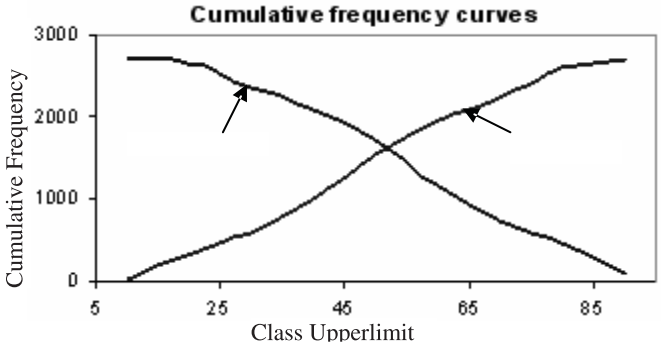
a. 5 minutes

b. 15 minutes

c. 20 minutes

d. half hour

4. The figure shows cumulative frequency of employees by salary. The point of intersection gives,



a. Mean

b. Median

c. Mode

d. Geometric mean

5. Number of children in 8 families is 2,3,0,5,8,1,3,1. To organize these data we should use frequency distribution for ungrouped data because ---

a. Data are quantitative

b. Observations are whole numbers

c. Values are not clustered

d. Number of possible values is small

6. The choice of an appropriate measure of central tendency is really crucial. It depends on the question that you ask and the information that you collect. For example suppose we have data on survival time (in years) of ten units. These are 9, 9, 25, 30, 35, 40, 49, 55, 70, 78. The average survival time (in years) of these 10 units is ---

a. 40 **b.** (35+40)/2 **c.** 9 **d.** none of the above

Discuss your answer.

7. Point of intersection of less than type and more than type ogive curves is ---

a. ($N/2$, Q_1) **b.** ($N/2$, mean) **c.** (mean, mode) **d.** (median, $N/2$)

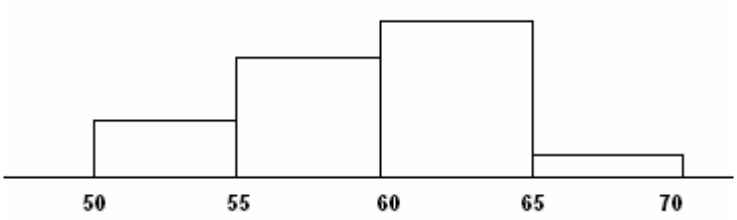
8. Most appropriate value of the mode based on the histogram is ---

a. 62.5

b. 61.5

c. 63.5

d. None of the above.



9. For a set of 10 positive observations on a variable X, A.M., G.M. & H.M. are computed, then which of the following triplet for (A.M., G.M., H.M.) is possible?

a. (50, 45, 40) **b.** (40, 50, 45) **c.** (40, 45, 50) **d.** (45, 40, 50)

10. Let x_1, x_2, x_3 be a speed of a train per unit time for t_1, t_2, t_3 duration of time respectively, then we can compute average speed of a train per unit time by using --

a. A. M. **b.** H. M. **c.** G. M. **d.** Median

11. In a group of 800 persons, 55% are men. If 55% of the women are literate then the number of illiterate women is ---

a. 360 **b.** 240 **c.** 198 **d.** 162

12. In a class, the average age of 50 boys is 16 years & 6 months; & that of girls is 15 years. The average age of whole class is 15 years & 6 months. The number of girls in the class is given by ---

a. 75 **b.** 82 **c.** 100 **d.** 150

13. Find appropriate measures of central tendency for following datasets.

Data set 1: Y- Mg level in females, X - Age in yrs

X	20	30	8	15	40	66	40	25
Y	2.5	2	1.9	2	3	2.1	2.5	3.1
X	30	30	35	40	26	30	38	50
Y	2.8	2.8	2	2.06	2.2	2.16	1.85	1.85

Data set 2: Yearly Income in rupees of 20 individuals

10,000	16,000	40,000	12,000	14,000
120,000	140,000	8000	330,000	50,000
12,000	200,000	20,000	24,000	30,000
36,000	36,000	96,000	60,000	40,000

Data set 3: # Dry days during rainy season

0	2	2	1	3	0	2	2	1	3	5	4	3	0	0
0	0	0	1	1	0	0	0	1	1	3	3	4	0	1
2	0	1	2	1	2	0	1	2	1	2	2	2	0	2
4	5	8	2	1	4	5	8	2	1	2	1	1	1	2

Unit 5 : Measures of Dispersion

5.1 Overview

In this unit you will be introduced to the concept of dispersion. To begin with we will discuss need for measuring variation. We will introduce two measures of variation and will discuss their details. Using illustrative examples, we will explain the computational procedure with which you can find these measures. We will do it for ungrouped as well as grouped data. You will learn the interpretation of s. d. Finally we will study relative measures of dispersion and their utility.

5.2 Learning Objectives

After studying this unit, you will be able to

- Understand the concept of variability,
- Need of defining measures of variability,
- Use measures of variability – Range and standard deviation,
- Compute measures of variability for ungrouped and grouped data,
- Use relative measure of variation – percent C.V for checking consistency.

5.3 Introduction

In the unit 4, we have studied various measures of central tendency. These measures give us an idea about the middle or centre of a data set. Measure such as mean is expected to represent the characteristics of the entire group. But knowing the value of the measure is usually not sufficient to describe the nature of the data. For example, a class teacher may be interested in knowing the distribution of the marks of the students and the extent to which they differ from each other. For a quality control engineer, information on maximum and minimum values of the quality characteristic observed during the manufacturing process of a lot is of importance. An investor may be interested in knowing the fluctuations in prices of a share from average stock prices.

What is a variable? A thing, which varies, is a variable. It can take all possible values in its range. The averages are not meant to describe the differences between the observations or of the spread of the values in the data set.

For example consider the three dataset reported below in Figure 5.1

Figure 5.1: Same mean, but different variability

It appears that the mean value for all the three data sets is the same (= 5.5 units), however the spread of the data values around this mean is different for the three data sets. A little inspection of the data shows that data set 2 with the squares has the maximum spread; where as the data set three with triangles has the minimum spread. The spread of the data values around the mean for the data set 1 is somewhere in the middle. Let us now consider a situation in a manufacturing industry. In case of industrial product the manufacturer invariably insists on less variability in their finished product – all units manufactured must be very close to average value, as specified by research and development department. More variability in the finished product may result into more rejection rate of such products. More rejection would mean either rework on the rejected units; if at all rework is feasible or it means adding to scrap volume. Thus, more variation in the quality of a product will result in more losses for the company. Is more variation always undesirable? No, not necessarily. For example, consider another scenario. An entrepreneur is interested in selecting an applicant for the position of trainee engineers (or programmers/ statisticians, etc). For this purpose he designs a test to be given to all the applicants. If the test is too easy or too hard then most applicants would get scores, which are too high or too low, and hence the test would not be good for choosing a suitable candidate among all. A good test would be the one, which results in to high variability in the test result data of the applicants. More the variation in test grades of the candidates more will be the opportunity for the employer to select a team of best candidates for the position in the organization. It is a matter of judgement to decide which type of variation is desirable? If, you as a customer enter in a sari shop and if, all the saris are of same type (no variation) then you are unlikely to visit the shop again. In manufacturing process industries, lesser the variation, better would be the quality.

Summary of data in the form of average provides what usually would happen but nothing can be said about reliability or accuracy of the observed value. Let us study one more case. We report here an extract of the data collected in a rubber industry. It was noticed that rejection of the manufactured rubber tubes was considerable and therefore the manager launched statistical enquiry. A quality control engineer collected the data on weight of rubber tubes that were manufactured in shift 1 (Morning shift), shift 2 (afternoon shift) and shift 3 (night shift) on the same machine. These data are reported in Table 5.1

Table 5.1: Weight (in gm) of a rubber tube manufactured in a Rubber Industry

Sr. No.	Shift 1	Shift 2	Shift 3
1	230	260	220
2	235	250	225
3	260	240	235
4	267	240	250
5	270	250	255
6	255	260	265
7	245	255	270
8	250	245	280
9	256		
10	232		
A.M	**2500/10 = 250**	**2000/8 = 250**	**2000/8 = 250**
Median	**252.5**	**252.5**	**252.5**

Notice that mean weight of rubber tubes manufactured in each shift is the same, viz., 250 gm and median weight is also the same (=252,5 gm); but inside story is likely to be different. This can be noticed by studying following diagram.

Figure 5.2: Variability in weight of rubber tubes

In all the three shifts, for the data reported, mean weight of a rubber tube is the same but it is the variation in weight that really was disturbing factor for the manufacturer.

In statistical studies, usually the variation in the characteristics being studied matters a lot, since knowing of average value of it is not enough. The characteristic can be height, weight, blood pressure, blood glucose level, biomass of a tree, per acre yield of a crop in experimental plots or in agricultural fields, income or expenditure pattern of a family, etc. It is known that observations on the study variable vary a lot. All values are not alike and that is why study of statistical techniques is necessary. Once we recognize and accept that existence of variability, then the question arises, "Can we quantify or measure it?" In this chapter we explain the concept of variation, suggest a methodology to measure it. We will also see the interpretation of it.

The above two illustrations make it amply clear that measure of location is far from sufficient. We do need measures of variability. Numerous measures of dispersion exist. We will study only two such measures. We will also define relative measures of dispersion. The requisites of a good measure of dispersion are the same as that of measures of location, which we have already studied in unit 4. We therefore start discussion on measures of dispersion.

5.4 Range

Definition: The **range (R)** of a set of n measurements x_1, x_2… x_n is defined to be the difference between the largest and the smallest of all the measurements. Usually symbol R is used to denote range of the data. Thus **R = Maximum – minimum**.

It is possible that two or more datasets have the same range for the characteristic being studied. If interest is to compare these two datasets then a **relative measure of dispersion** is recommended. One such measure based on range is coefficient of range. It is defined as

Coefficient of Range = (Maximum – Minimum) / (Maximum + Minimum)

It can be used for comparing two datasets (populations) at the fixed time or the same population over two different time periods.

5.4.1 Raw Data (Discrete frequency distribution)

Suppose x_i (i = 1, 2…n) are the data on the characteristic X being studied. We must first arrange all the observations either in ascending order (or in descending order). Let $X_{(1)}$ and $X_{(n)}$ respectively denote minimum and maximum of all the observations. Then Range = $X_{(n)} - X_{(1)}$.

Refer data in Table 5.1.

For shift 1, range R = 270-230 = 40 gm and coefficient of range = 40/500 = 0.08

For shift 2, range R = 260-240 = 20 gm and coefficient of range = 20/500 = 0.04

For shift 3, range R = 280-220 = 60 gm and coefficient of range = 60/500 = 0.12

Thus for shift 2, we observe that range is smaller and coefficient of range is also smaller as compared to other shifts. In general, smaller the range less is the variation and smaller the coefficient of range, more is the consistency. We may express coefficient of range in percentage.

If the data are reported in the form of frequency tables (also referred as frequency distribution of data), we can obtain range for such dataset as above.

5.4.2 Grouped Data (Continuous frequency distribution)

Suppose we are given data in the form frequency distribution of continuous type. Then we can obtain range as difference between mid-values of last class-interval and the first class-interval. For example, if the data are reported in frequency distribution having class-intervals 10 - 20, 20 - 30…. 90 -100, then the range for such data would be 95 - 15 = 80 and similarly we can find coefficient of range.

5.4.3 Merits and Demerits of Range

Merits:

- It is easy to calculate and simple to understand.
- It is based only on two extreme observations.
- Range is appropriate for small data sets.
- Two or more datasets may have same range but they may still vary a lot.
- In statistical quality control, charts based on range are widely used.
- In weather forecasts, share markets, mutual funds range as a measure is used.

Demerits:

- Range is appropriate for small data sets.
- Two or more datasets may have same range but they may still vary a lot.

Looking at merits, rather at demerits of range as a measure of dispersion in data, it is necessary to have better measure than range. The standard deviation is another widely used and most common measure to describe dispersion in data. Some authors rightly remark that standard deviation is the heart and soul of the concept of variability in statistical inference. What is the role of a standard deviation? How is it defined? How to compute it for various types of data? Can we compute it always? How to interpret it? These are some of the frequently asked questions. We will answer these questions and will explain them with simple examples.

5.5 Standard Deviation

The standard deviation (S.D.) is a single number. It measures the spread of the distribution of data around its mean. Larger the value of the standard deviation more is the spread of the data around its mean; smaller the value of the standard deviation less is the spread of the distribution of the data around its mean. If you plot all the values, it will indicate spread of the data.

We will first define the concept of the variance and then standard deviation.

Definition of variance: The variance of n observations $x_1, x_2 \ldots x_n$ is defined to be the average of the square of the deviations of the observations from their mean. The variance is denoted by symbol σ^2 (σ is Greek letter sigma) and is given by the formula

$$\sigma^2 = \sum (x_i - \bar{x})^2 / n = [\sum (x_i^2) / n] - (\bar{x})^2$$

Some authors also use the notation $V(X) = \sigma_x^2$ to denote variance of X.

Definition of Standard Deviation (S.D): The standard deviation of a set of n measurements is defined as the positive square root of variance. It is usually denoted by symbol σ.

Procedure to calculate standard deviation (s. d.)

Case 1: Raw data

Here we are given n observations $X_1, X_2 \ldots X_n$ on the characteristic X being studied. To compute V(X),

- First find sum of all values. Divide sum by number of observations. This is arithmetic mean of all values.
- Find Find $(X_i - \bar{X})$ and $(X_i - \bar{X})^2$ for each i,
- Find sum of $(X_i - \bar{X})^2$, and then divide the sum by number of observations.
- $\sigma^2 = \sum (x_i - \bar{x})^2 / n = [\sum (x_i^2) / n] - (\bar{x})^2$

This is the variance of the observed data; its positive square root is the standard deviation σ.

Remember to state units of s. d; the units of s. d. are the same as that of X. Tabular form, as shown below, simplifies the computation of s. d.

Example: Computation of s. d. for raw data

Table 5.2: Weight of Rubber tubes (in gm)

Unit no	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
1	230	-20	400	52900
2	235	-15	25	55225
3	260	10	100	67600
4	267	17	289	71289
5	270	20	400	72900
6	255	5	25	65025
7	245	-5	25	60025
8	250	0	0	62500
9	256	6	36	65536
10	232	-18	324	53824
Total	2500	0	1824	626824
Mean	250			

Using the formula for variance, $\sigma^2 = \sum (x_i - \bar{x})^2 / n = [\sum (x_i^2) / n] - (\bar{x})^2 = \mathbf{182.4}$ and s.d. (X) = σ = 13.5056 ≈ 13.51 gm, say.

Thus for the data in Table 5.2, mean weight of tubes is 250 gm with a s.d. of 13.51 gm.

Case 2: Grouped data

For discrete frequency distribution as well as for continuous frequency distribution, formula for computation of variance is given below:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i^2}{\sum_{i=1}^{k} f_i} - (\bar{x})^2 \text{ , where } \bar{x} = \frac{\sum_{i=1}^{i=k} f_i x_i}{\sum_{i=1}^{k} f_i}.$$

For discrete frequency distribution x_i is the observed value of X and for continuous frequency distribution, it is the mid-value of i^{th} class-interval. We assume that we have k values of random variable X or the data are grouped into k classes.

Example: Discrete frequency distribution

The data on number of non-confirming units manufactured in a batch of 100 is reported below.

Table 5.3: Data on 100 boxes each containing 200 units of which X are non-confirming units

X	0	1	2	3	4	5	Total
# boxes	30	45	18	5	1	1	100

In order to find s.d. of X, you may prepare following Table 5.4

Table 5.4: Computation of standard deviation

X	f: # boxes	fx	fx^2
0	30	0	0
1	45	45	45
2	18	36	72
3	5	15	45
4	1	4	16
5	1	5	25
Total	100	105	171

Here mean number of non confirming units =1.05 with

$V(X) = 17.1 - (1.05)^2 = 15.9975$. Hence S.D.(X) = 3.9997 = 4, say

Example: Continuous frequency distribution

In case of continuous frequency distribution, it is many times convenient and also desirable to transform the variable under consideration, say X, to a new variable U by the transformation $U = (\mathbf{X - a}) / \mathbf{h}$ where a and h are two suitably selected real numbers ($h \neq 0$). It is easy to deduce that $\bar{\mathbf{X}} = \mathbf{a + h}\,\bar{\mathbf{U}}$ and $\mathbf{V(X) = h^2 * V(U)}$. This is an important property of variance. It means variance is independent of change of origin but it depends on change of scale. How to choose values of h and a? Any convenient values can be chosen. To ease the calculations choose 'a' close to central value of the data and h as the class-width. This was convenient in good old days when calculators and personal computers were not available. We will illustrate its use for the data on the age-distribution of researchers (Table 5.5). We will find mean and s.d. of these data using change of origin and scale property.

Table: 5.5: Computation of s. d. for continuous frequency distribution

Age (in years)	25 -35	35 -45	45 -55	55 -65	65 -75
# Researchers	25	50	80	45	30

Let us prepare following Table 5.6 to compute s. d.

Table 5.6: Computation of s.d.

Age (in yrs)	*Mid-value x*	*Frequency (f)*	*U = (X – 50) / 10*	*fu*	fu^2
25-35	30	25	-2	-50	100
35-45	40	50	-1	-50	50
45-55	50	80	0	00	00
55-65	60	45	1	45	45
65-75	70	30	2	60	120
Total		230		5	315

Mean of U = 5/230 = 0.0217; mean of X = 50 + 10(0.0217) = 50.21 yr

$V(U) = 315/230 - (0.0217)^2 = 1.3691$; hence $V(X) = 10^2*(1.3691) = 136.91$

S.D. (X) = $\sqrt{V(X)}$ = 11.70 yr.

It shows that mean age of researchers is 50 years with a s. d. of 11 years. It is indicative that very few students are not entering in the research profession, a thing to worry about.

5. 6 Merits and Demerits of Standard Deviation

Merits:

- It is simple to understand.
- It is precisely defined,
- It is based on all observations,
- It gives more weight to extreme observations, this is sometimes viewed as a demerit (particularly by economists or business community),
- It is least affected by sampling fluctuations,
- For comparing consistency of two or more data sets, it can be used,
- It is capable of further mathematical treatment, e.g., it is possible to derive formula for combined variance for two or more datasets.
- It is used extensively in statistical inference problems

Demerits:

- It is not easy to calculate. (However, with availability of personal computers, this is no more so.

5.7 Formula for Combined Standard Deviation (without proof)

We have earlier remarked that standard deviation is capable of further mathematical treatment, that is, if the means and standard deviations of two or more groups are given then the standard deviation of the combined group can be computed. We state its formula without proof. Proof is easy and it is left for you.

For two sets of observations having number of observations, mean and standard deviation **m,** $\bar{x}$ **,** σ_1 and **n,** $\bar{y}$ **,** σ_2 respectively, the standard deviation of the combined group, σ is given by $\sigma = \sqrt{[m(\sigma_1^2 + d_1^2) + n((\sigma_2^2 + d_2^2)]/(m+n)}$ where $\mathbf{d_1} = (\bar{x} - \bar{X}_C)$ and $\bar{X}_C$ is the combined mean of two groups, similarly $\mathbf{d_2} = (\bar{y} - \bar{X}_C)$ is defined.

Remark: It is easy to generalize the formula for combined variance.

Formula for combined variance of k datasets

For k sets of observations having number of observations, mean and standard deviation $n_i, \bar{x}_i, \sigma_i$ respectively, the standard deviation σ, of the combined data, is given by σ where

$$\sigma^2 = \sum_{i=1}^{k} n_i(\sigma_i^2 + d_i^2) / \sum_{i=1}^{k} n_i \quad \text{where } d_i = (\bar{x}_i - \bar{x}_c) \text{ for } i = 1, 2, \ldots, k$$

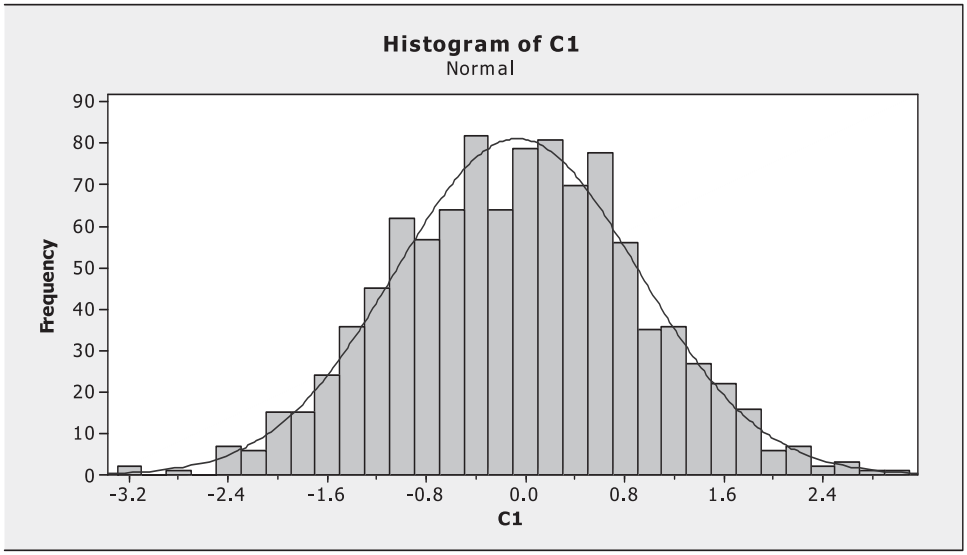
5. 8 Interpretation of Standard Deviation

You may still wonder how to interpret standard deviation (s .d.)? We will answer this query but before that let me ask you simple question, "How do you interpret weight of an individual?" The question is simple. You know for sure that more the weight (reported in kg), bigger will be the individual, smaller the weight smaller will be the individual. You are familiar with the unit like kg. Think similarly while learning s.d. It is a positive number that tells of the variability in the data; a bigger S.D. means more variation; 0 means no variation at all i.e. all observations exactly equal to the mean. With availability of calculators, computers and statistical software, computation of various measures is a simple task. The real work begins thereafter. How to interpret a standard deviation? What it represents? It simply indicates how the data is scattered around its mean. In other words, it indicates on an average how far off an observation will be away from the mean. More the standard deviation more is the heterogeneity in the data and vice versa. The data sets whose frequency distributions looks like symmetric bell shaped are called "normal" data sets. In case of "normal" data sets the majority of the observations can be expected to lie within one standard deviation of the mean. In general we can expect that the interval

Mean ± S. D. will contain approximately 68% of the measurements,

Mean ± 2*S. D will contain approximately 95% of the measurements,

Mean ± 3 *S. D will contain approximately almost all the measurements,



For the data reported in Histogram of C1, mean is zero and S. D. is one.

Suppose we are given 100 observations from normal population having mean 50 and s.d. 10. In the above case we can expect that roughly 68 of the observations to lie between [40, 60], 95 of them to lie between [30, 70] and almost all to lie between [20, 80] Thus information on mean and s.d. can be used (i) to judge presence of extreme observations or outliers if any and (ii) to postulate the model from which the sample might have been taken. (At this juncture, you may ask the question, "What happens if the distribution of measurements is not bell-shaped but something other?" It is a good question, but we cannot discuss further details at this stage. Interested students should study Tchebysheff's theorem to learn more about it.

In practice, we come across several problems. For example, we may be required to compare two or more datasets and answer the question "observations in which of the two data sets are more homogenous?" Can this question be answered by considering only the standard deviations of the two datasets? You can certainly compare two or more datasets, using standard deviations, provided their mean values are more or less same. But means and standard deviations both can differ. How to compare two datasets in this situation? In order to answer the question we must think of a relative measure of dispersion – called coefficient of variation.

5.9 Coefficient of Variation (C.V)

Coefficient of variation is defined as the ratio of standard deviation to mean; i.e., **C.V. = s.d. / mean**. It is usually expressed in percentage and is referred as percent coefficient of variation. Notice that since s.d. and mean are expressed in the same units, percent C.V. is independent of units in which the observations are collected. How to use this relative measure of dispersion and for what purpose? If we want to compare two or more data sets (or groups or populations) then we find mean and s. d. of each data set and compute percent C.V. for it. Smaller the value of C.V. more is the homogeneity. For example, in case of biological populations, it is said that natural variation is of the order of 15 to 20 %. This is what is reported in the literature but of course it is not a rule. If you have more percent C.V. then the reasons for it must be investigated. Remember that percent C.V. is the best measure of relative dispersion.

Example: Following are the data on run scored by two batsmen A and B in ten innings during a certain cricket tournament

Batsman A:	42	38	57	73	81	49	20	70	106	54
Batsman B:	19	31	48	53	61	90	40	62	40	80

Is batsman A better run scorer than B? Is he more consistent compared to B?

Solution: We compare the batsmen by their average scores and percent coefficient of variation of scores for better run getting and consistency. Verify that average runs scored by A and B are respectively 59 and 53 and corresponding standard deviations are 23.22 and 20.45. Clearly batsman A is better than batsman B. verify that percent C.V for A and B is respectively 39.5 and 38.6. Since smaller the percent C.V. more is the consistency, B appears to be more consistent but notice that both values of C.V are very close. It is better to conclude that both are equally consistent.

In the world of investments, the coefficient of variation guides the investor to determine how much risk he is assuming in comparison to the amount of return he can expect from his investments. The lower the ratio of standard deviation to mean return, the better is risk-return tradeoffs. (It is assumed that the variable for which you find C.V. will always take positive values).

We have studied two important measures of dispersion – range and standard deviation. Obtaining range (R) of a data is simple, but that of s. d. is not so simple. We can use the check on calculation of s d. as R ≈4 *(s.d). Remember you will not get accurate value of s.d., only its **order of magnitude** can be checked. Measures of location and dispersion are certainly important, but there are certain other features of a distribution such as symmetry and flatness or peakedness. We will study them in unit 6.

5.10 Summary

In this unit we have studied concept of variability. We have defined range and s.d. as measures of dispersion. We know the methods to compute these measures for various types of data. We have also studied percent C.V. as a relative measure of dispersion and now we can employ these known statistical techniques to analyse the datasets.

5.11 Self Test

Instructions: In case of multiple choice questions, mark all correct answers; if no answer is correct then report accordingly.

1. If each value in the data set $\{x_1, x_2, \ldots x_n\}$ is doubled then range of the data

a. Remains unchanged **b.** Will be doubled

c. Will be 50% of original range **d.** Cannot be obtained.

2. Study following data sets and answer following questions:

Set A = {1, 2, 3, 4, 5}; Set B = {10, 20, 30, 40, 50}; Set C = {11, 21, 31, 41, 51}

2. (a) Which of the following statements is correct ?

a. Mean of C = 1 + mean of B, and S.D (C) = S.D. of B

b. Mean of C = 1 + mean of B, and S.D (C) = 10*S.D. of B

c. Mean of C = 1 * mean of B, and S.D (C) = S.D. of A

d. Mean of C = 10 + mean of A, and S.D (C) = 10+ S.D. of A

2. (b) Which of the following statements is correct ?

a. Mean of B = 10 * mean of A, and S.D (B) = 10*S.D. of A

b. Mean of B = 10 * mean of A, and S.D (B) = 100*S.D. of A

c. Mean of B = 10 * mean of A, and S.D (B) = ($\sqrt{10}$)*S.D. of A

d. Mean of B = 10 * mean of A, and S.D (B) = S.D. of A

3. A certain data has a mean of 4 meters and a standard deviation of 0.7 mm. Its percent C.V. is ---

a. 0.07 / 4 **b.** 0.0007 / 4 **c.** 0.7 / 4 **d.** 7 / 40

4. For the two data sets, the numbers of observations are 10 and 15. Further their means and s.d.' s are 100, 10 and 200, 25 respectively.

4. (a) Combined mean of these two sets would be ---

a. 150 **b.** 160 **c.** (100 + 200)/(10 + 15) **d.** (1000 + 3000)/2

4. (b) Combined variance is equal to ---

a. (10*100 + 15*625)/(10+15) **b.** (10*10+15*25)/(10+15)

c. (100 + 625)/ 2 **d.** None of the above

5. Two sets have same means (=100) and same s.d.' s (= 25). Then ---

a. Combined mean = 100 and combined s. d = 625

b. Combined mean = 100 and combined s. d = 25

c. Combined mean = 100 and combined s. d = 50

d. Combined mean and combined s. d. cannot be obtained.

6. Monthly consumption of electricity of a certain family in a year is given below:

210 207 315 250 240 232 216 208 209 215 300 290

Find range, coefficient of range, s. d. and percent C.V.

7. The frequency distribution of daily expenditure of families is given below

Expenditure	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
No. of families	14	21	32	28	15

Find range, coefficient of range, s. d. and percent C.V.

8. Find coefficient of range and percent C.V. for following data:

Rainfall	20 -24	25 -29	30 -34	35 -39	40 -44	45 -49	50 -54
Number of Years	2	5	8	12	10	7	6

9. Determine which class is better and consistent:

Marks	4	5	6	7	8	9
Class A	10	13	11	7	5	4
Class B	9	11	10	10	5	10

10. A machine capability study was made on a Brown and Sharpe single-spindle screw machine. The number of items inspected (sample size), their mean diameters and standard deviations reported were as follows.

Sample size	Mean diameter (mm)	Standard deviation (mm)
4	2.8325	0.2479
6	2.8333	0.2687
5	2.8520	0.2786
4	2.8400	0.2581
5	2.8820	0.2721
6	2.8533	0.2925

Show that the combined mean and combined standard deviation of all samples is 2.84932 mm and 0.2724 mm respectively.

11. Time taken (in minutes) per customer by a counter employee is shown below:

Clerk I	5	5	3	4	2	5	4	5	3	5	2
Clerk II	3	3	5	4	5	5	3	5	3	5	5

It is claimed that A is better than B and is also consistent. Do you accept the claim? Justify your answer.

12. 12. Let N = {1, 2, 3…} denote a set of natural numbers. Find mean and s. d. of first 100 natural numbers.

{Hint: You may use formula $\sum_{i=1}^{n} i = n(n+1)/2$ and $\sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$ }

13. Let X be a variable with $\bar{X}$ and σX as arithmetic mean and standard deviation respectively. Define U = (X – A) / h, where A & h are any constants and h ≠ 0, then with usual notations --

a. $\bar{X} = A$ and $\sigma_x = \sigma_u$ **b.** $\sigma_x = |h| \sigma_u$ and $\bar{X} = A + h \bar{U}$

c. $\sigma_u = h^* \sigma_X$ and $\bar{X} = A + h \bar{U}$ **d.** $\bar{X} = \sigma_u + A$ and $\sigma_X = A$

14. Mrs. Bokil gave birth to 5 daughters in alternate years. The standard deviation of the ages (in years) of these 5 children is approximately equal to ---

a. 4 **b.**2.8 **c.** 5 **d.** cannot be calculated

15. The data on the scores awarded to fifteen students in a test are as follows:

1, 2, 15, 15, 17, 16, 19, 19, 20, 23, 24, 25, 25, 30, 401

A student calculates all of the descriptive measures of location and dispersions on these data. Later on he discovers that an error was made and one of the 25' s is actually 30.

15. (a) Which of the following measure will not be changed from the original computation?

a. Mode **b.** Mean **c.** Median **d.** SD

15. (b) Which of the following measure(s) will change from the earlier computations?

a. Range **b.** Coefficient of range **c.** Variance **d.** SD

16. Write a short note on use of percent coefficient of variation.

17. Identify whether the statements below are true or false. Justify your answers.

You are given a data {10, 20, 30, 30, 40, 50}

(i) If you add 10 to each observation, it adds 10 to the average,

(ii) If you change the sign of each observation, the sign of s. d. will also change,

(iii) If you multiply by 2 to each observation, variance will get doubled and s. d will get multiplied by $\sqrt{2}$,

(iv) If you add 10 to each observation, relationship between mean, median and mode will remain unchanged,

(v) Percent C.V will change if each observation is multiplied by 100.

18. Following data sets have the same average of 150. Which one has

(i) smaller range,

(ii) smaller s. d.

(iii) smaller C.V ?

Dataset I {175, 150, 140, 125, 170, 130, 160}

Dataset II {125, 130, 140, 150, 150, 150, 150, 160, 170, 175}

Dataset III {150, 140, 160, 150}

Dataset IV {200, 100, 150, 150, 140, 160}

19. The time required (in minutes) for writing a successful program is the variable under consideration. Two students Swanand and Ashish are asked to write 10 programs and submit them. The data on time required are as follows:

Swanand	10	15	24	8	12	10	10	7	8	10
Ashish	8	12	30	10	12	15	8	9	5	10

Analyze above data and comment on the results.

20. The number of insects trapped using two different treatments (methods) are reported below. Which method is consistent? Discuss your results.

Sampling	1	2	3	4	5	6	7	8
Trap A	40	56	60	35	120	100	80	110
Trap B	60	70	50	25	100	80	120	100

It was later reported that last four sampling occasions were conducted during nighttimes. Reanalyse these data in the light of this new information.

Unit 6 : Moments, Skewness and Kurtosis

6.1 Overview

In earlier chapters we have studied some quantitative techniques extensively used in exploratory data analysis. It includes presentation, classification and tabulation of data. It also includes finding out measures of location and dispersion, for a given dataset. Suppose you obtain all the three important measures of location, viz, mean, median and mode for a dataset. The question that arises is -"In what order these measures would occur? Is any order possible? You have also studied range and s.d. as two important measures of dispersion. If values in the dataset are concentrated near its mean value, the spread of these values will be less. In this unit we will study the concept of moments. Based on moments and the various measures studied earlier, we will study two more aspects of a data, viz., skewness and kurtosis. If the data are symmetric about a value (i.e. median), then there is no skewness in the data. Thus, skewness of a data is a measure of asymmetry of the data. Usually the frequency distribution of the data has a peak at the mode. Question arises as to how sharp in this peak? We compare this peak with the peak in a bell shaped data with the same mean and variance. If the two peaks are of the same height then we say that there is no kurtosis in the data. Thus we can say that the kurtosis is a measure of the peakedness of the data. In this chapter we introduce measures of skewness and kurtosis.

6.2 Objectives

After studying this chapter, you will be able to -

- Explain the concept of moments,
- Define skewness based on moments,
- Obtain measure of skewness for different types of datasets,
- State meaning of kurtosis and related terms,
- Compute measure of kurtosis
- Explain the nature of data.

6.3 Introduction

We will begin our study with introduction of the concept of moment. The term 'moment' is well known in mechanics. We reproduce it here for information. Moment is a product of force and distance of the line of action of the force from the point about which it is measured. In statistics also we use the term moment. It is analogous to the term used in mechanics. For example, for a grouped data, relative frequency of each class can be compared with the 'force' and the deviation $(x_i - \bar{x})$ as the distance. For raw data the frequency of each value is one and hence relative frequency is 1/n. We will first formally define the two terms (i) raw moments and (ii) central moments.

6.3.1 Definition

Suppose we are given $x_1, x_2 \ldots x_n$ observations on a characteristic X under consideration. The r^{th} moment of X about **any point A**, denoted by $\underline{m'_r}(A)$ for raw data is defined as

$$\mathbf{m'_r(A) = \frac{1}{n}\sum_{i=1}^{i=n}(x_i - A)^r}$$

For grouped data, $m'_r(A)$ is defined as $\mathbf{m'_r(A) = \dfrac{\sum_{i=1}^{i=k} f_i(x_i - A)^r}{\sum_{i=1}^{i=k} f_i}}$ for r = 0, 1, 2 …..

Case 1: The moments about A = 0 are called raw moments. The subscript r is called the order of the moment. For r = 0, we have $m_0' = 1$ (always) and for r = 1 we get m'_1 = sample mean. We can find moment of any order, however for all practical purposes, first four moments suffice.

Case 2: For $A = \bar{x}$, then we get central moments. These are called central moments (or moments about mean) for the given data. Let m_r denote central moment of order r. It is easy to check that for r = 0, $m_r = 1$ (always) and for r = 1, $m_1 = 0$ always. (Why?). This is because sum of deviations of observations taken from A. M is always zero. For r = 2, we get second central moment, which is variance of the data. For r = 3, 4 … we can obtain higher order raw moments or central moments.

A natural question arises, "Can we interrelate raw and central moments? What is the relationship between the two? What is the use of these moments?" We will answer these questions one by one.

6.3.2 Nature of Relationship Between Central and Raw Moments

Recall $\mathbf{m_r = (1/n)\sum(x_i - \bar{x})^r}$, for r = 0, 1, 2, …

You can apply binomial theorem and then expand R. H. S. of above relation. Then use the definition of raw moments. You will get following results. We state these results (without proof). These are the relationships in which central moments are expressed in terms of raw moments of lower order. Proof of the following results is simple. It is left as a self-study in the interests of students.

1) $m_1 = 0$ always 2) $m_2 = m'_2 - (m'_1)^2$

3) $m_3 = m'_3 - 3m'_2m'_1 + 2(m'_1)^3$ 4) $m_4 = m'_4 - 4m'_3m'_1 + 6m'_2(m'_1)^2 - 3(m'_1)^4$

Thus if raw moments are known then the central moments can be obtained (and conversely); but usually the interest lies in knowing mean value (first raw moment) and central moments upto order four. This is because these are enough to indicate the nature of the dataset.

6.3.3 Properties of Central Moments

1. The central moments are invariant to the change of origin.

Let $\mathbf{U = X - A}$, then since that $\mathbf{\bar{X} = A + \bar{U}}$, r^{th} central moment of U, denoted by

$\mathbf{m_r(u) = (1/n)\sum(u_i - \bar{u})^r = (1/n)\sum(x_i - A + A - \bar{x})^r = m_r(x)}$

2. Effect of change of scale

Let $\mathbf{U = X / h}$ then $\mathbf{\bar{X} = h\,\bar{U}}$. Hence r^{th} central moment of U, denoted by

$\mathbf{m_r(u) = (1/n)\sum(u_i - \bar{u})^r = (1/n)\sum[(x_i / h) - (\bar{x} / h)]^r = (1/h^r)m_r(x)}$

We can use this property to find central moment of X, from that of U.

3. You can combine above two properties and apply it while obtaining moments.

We have defined r^{th} central moment of a random variable X as $\mathbf{m_r = (1/n) \sum(x_i - \bar{x})^r}$. Define **U = (X – A)/h**, then $\mathbf{\bar{X} = A + h \bar{U}}$ and you will get $\mathbf{m_r (X) = h^r [m_r(U)]}$, where $m_r(U)$ and $m_r(X)$ denotes r^{th} central moments of U and X respectively. We will illustrate its use with the help of illustrative example.

6.4 Skewness and Kurtosis

There are occasions when in addition to measures of location and measures of dispersion, researchers are interested in shape of the distribution of the variable under consideration. For example,

(1) Economists are interested in analyzing data on income distribution of a family or land holdings in villages.

(2) Nutritionists are interested in the study of distribution of Energy intake (in Kcal) or Protein intake (in gm) of sampled population from a particular area.

Note that when the data are symmetrically distributed around a value, that value will be the mean as well as the median of the data. In fact the first central moment of the data is always zero. For the symmetric data even the third central moment will be zero (Why?). Thus non-zero value of third central moment means skewness and hence lack of symmetry. If there is skewness then there are two possibilities. The data (frequency distribution) may be having a long tail to the right and hence larger positive deviations from the mean. This type of skewness is referred as positive skewness. On the contrary, if the frequency distribution has a long tail to the left and hence larger negative deviations from the mean, it is said to have negative skewness. No skewness means lack of asymmetry – thus meaning complete symmetry. Usually, in a positively skewed data, mean exceeds median and mode. In a negatively skewed data curve value of the mean will usually be less than the median and the mode.

6.4. 1 Coefficient of Skewness

We can specify several indexes of skewness. We list them below.

(1) Karl Pearson's Coefficient of skewness

$$\mathbf{S_1 = (A.M. – Mode)/ S. D.}$$

This is based on measures of central tendency. But if mode is not uniquely defined then this measure is also not well defined. In this case you can use the next measure,

(2) Karl Pearson's measure of Coefficient of skewness

$$\mathbf{S_2 = 3(A.M. – Me) / S.D.}$$

(3) Bowley's coefficient of skewness, SkB, (based on quartiles)

$$S_K^B = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

where Q_1 and Q_3 are respectively lower and upper quartiles. Below Q_1, 25% observations lie and above Q_3, 25 % observations lie. Determination of Q_1 and Q_3 is easy. It is done similar to Q_2 (which is same as median).It is particularly useful if you have open-ended classes (such as less than or greater than a particular value)

(4) Coefficient of skewness (based on moments)

$$\gamma_1 = \sqrt{\beta_1}\ , \text{where } \beta_1 = \frac{m_3^2}{m_2^3}.$$

Remember that sign of γ_1 is important. It can be either positive or negative. Attach a sign that of m_3 to the coefficient of skewness. Study the three figures given here.

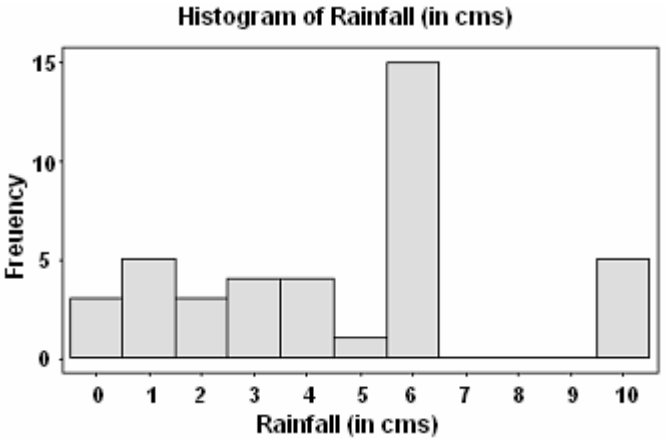


Figure 6.1: Mean < Median < Mode (Negative skewness)

They display negatives skewness (Fig 6.1), zero skewness (Fig 6.2) and positive skewness. (Fig 6.3) for various types of data. The original data are given in the self-study section at the end. (Ex-10). In Fig 6.1 we have plotted data on Rainfall (in cm) at a locality, in Fig 6.2, data on height of 100 students is plotted and in Fig 6.3, data on recovery time of patients admitted in a hospital are plotted. These three graphs respectively demonstrate negative skewness, no skewness and positive skewness of the data under consideration.

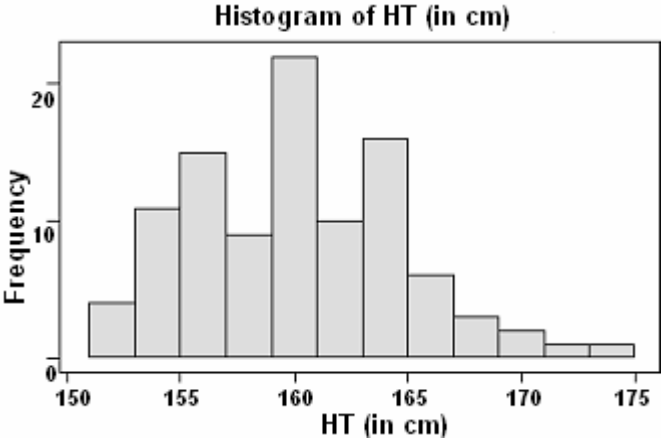


Fig 6.2: Mean = Median = Mode

Coefficient of skewness is a pure number (no units) and is independent of change of origin and scale. (You may prove it as a part of self-study).

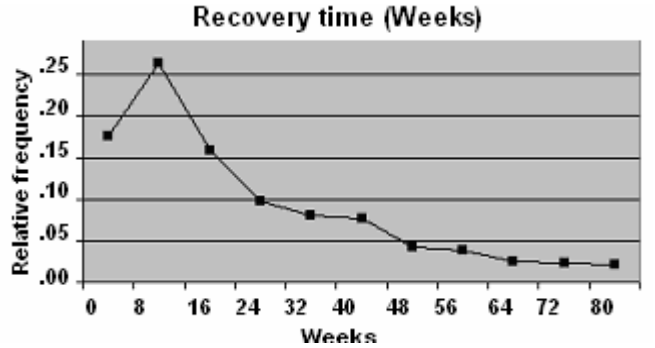


Figure 6.3: Mode < Median < Mean (+ve skewness)

6.4.2 Coefficient of Kurtosis

We have seen that skewness informs us about degree of asymmetry. Thus to know the shape of the distribution and the size of variation around the mean, skewness can be used, But what about the concentration of measurements at the central part of the distribution? It is kurtosis that measures the concentration of distribution at the central part. Kurtosis deals with the relative peakedness of the distribution around the mode. Kurtosis is a Greek word. It means "bulginess'. **Even though it can be defined for all types of data sets, it makes sense to talk about the kurtosis of the data set when it is at least approximately symmetric with the peak in the middle.**

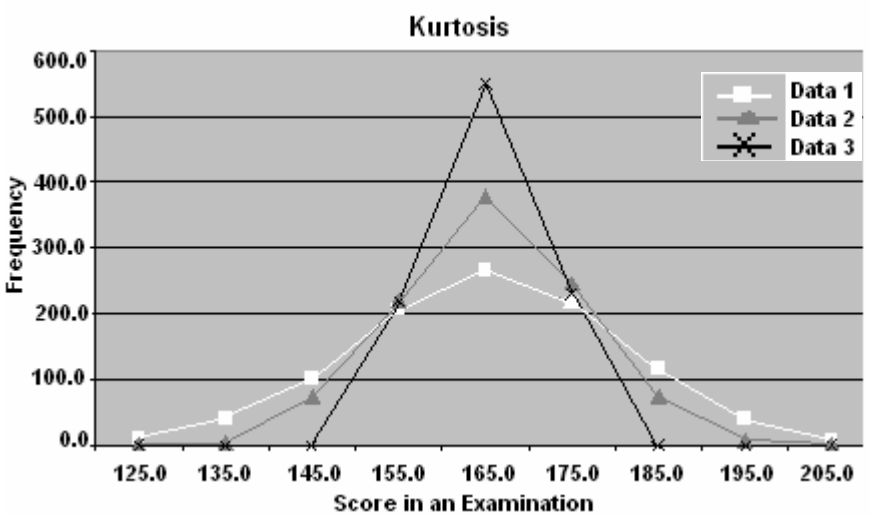


Fig 6.4: Kurtosis – Platykurtic, Mesokurtic and Leptokurtic curve

Study the Figure 6.4. The data on scores of students from three different regions of a city are plotted below. It can be seen that the average score is more or less same in all the three data sets, but the concentration near the mean value is different for each of the data set.

We will define only one measure of kurtosis, a measure based on central moments. We denote it by γ_2 and can be obtained as follows:

$\gamma_2 = (m_4 / m_2^2) - 3$, you may denote (m_4 / m_2^2) by β_2, so that $\gamma_2 = \beta_2 - 3$. In fact it can be shown that for the bell shaped data sets the value of β_2 is 3. Thus γ_2 is a measure of departure in peakedness for the data to that of the bell shaped data.

If $\beta_2 < 3$, the data is Platykurtic,

$= 3$ the data is Mesokurtic, and

> 3, the data is Leptokurtic.

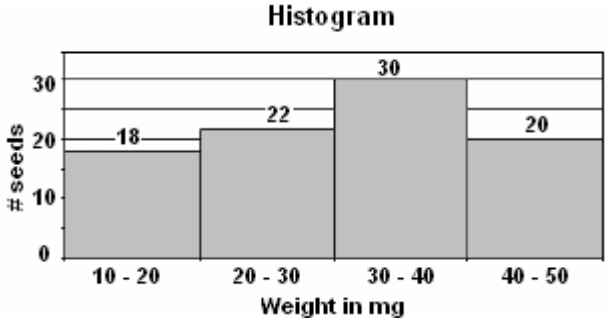
As in case of coefficient of skewness, β_2 is a pure number and is independent of change of origin and scale. β_1 and β_2 both are related. The relationship between them is given by $\beta_2 \geq \beta_1 + 1$.

6.5 Numerical Example

Following is the weight distribution of seeds. We will analyze these data and will find various measures of location, dispersion and also discuss its nature.

Weight in mg	10 - 20	20 - 30	30 - 40	40 - 50
Frequency	18	22	30	20

It is always desirable to draw appropriate graph/chart.



1. Here we will draw histogram. It shows that the distribution of weight is likely to be negatively skewed.

2. Further let us find measures of central tendency.

Verify that mean = 30.8 mg; Me (Median) = 31.7 and Mode = 34.4

Thus mean < median < mode, it indicates that distribution of seed weight is likely to be negatively skewed

3. Let us find measures of dispersion

Range = 45 -15 = 30 an S. D. = 10. 4326 (close to approximation 30/4 = 7.5)

Range and s.d. are quite large. Variation appears to be considerable.

4. Percent C.V = 100 S.D /mean = 33.90; a large coefficient of variation. It indicates

heteroscedasticity in the weight of seeds in the population.

5. In order to compute moments, skewness and kurtosis it is better if we prepare table such as give below: Let us define U = (X - 35) /10. Recall the notations used. Here A = 35 and h = 10.

Table 6.1: Computations or skewness and kurtosis

LCL	UCL	F	x_i	u	uf	u^2f	u^3f	u^4f
10	20	18	15	-2	-36	72	-144	288
20	30	22	25	-1	-22	22	-22	22
30	40	30	35	0	0	0	0	0
40	50	20	45	1	20	20	20	20
Total		**90**		**-2**	**-38**	**114**	**-146**	**330**

Raw moments of U	Central moments of U	Central moments of X
m'_1 = -0.4222	$m_1 = 0$	$m_1 = 0$
m'_2 = 1.2667	$m'_2 = 1.0884$	$m_2 = 108.84$
$m'_3 = -1.6222$	$m_3 = -0.1683$	$m_3 = -168.3$
m'_4 = 3.6667	$m_4 = 2.1864.35$	$m_4 = 21864.35$

***Mean of X** = 30.78 mg **S. D.(X)** = 10.4326 mg **C.V. %** = 33.90*

Median of X = 31.67 mg and **Mode** = 34.44 mg

Coefficient of skewness: S1 = -0.35; S2 = -0.26; β_1 = 0.02197 and γ_1 **= -0.1482:**

Coefficient of Kurtosis: β_2 = 1.8457 and γ_2 **= -1.15**:

Comment: Distribution of weight of seeds is **negatively skewed**; the distribution has **platykurtic** curve, it is likely to have a **flat top** and has a **long left tail.**

6.6. Summary

In this chapter we studied concept of moments, skewness and kurtosis. In it we have studied various measures to determine skewness and kurtosis of a given data.

6.7 Self Test

1. Multiple choice questions: (Read the following paragraph and answer)

"Measurements were taken on the quality characteristics (X) of a product. The data on 100 units showed that mean = 10.5, median = 11, mode = 13, Q_1= 10.2, Q_3 =11.2 and percent C.V for X was 4."

(i) The distribution of quality characteristic X is ---

a. positively skewed **b.** Negatively skewed

c. zero skewness **d.** inconclusive.

(ii) Coefficient of skewness based on quartiles is ---

a. -3/5 **b.** 3/5 **c.** 5/3 **d.** -5/3

(iii) 1.3 Coefficient of kurtosis is

a. > 0 **b.** < 0 **c.** zero **d.** cannot be obtained

(iv) 1.4. The distribution is likely to be negatively skewed because ---

a. mean = median = mode **b.** mean < median < mode

c. $Q_1 < Q_3$ **d.** Q_1 < mean < s.d.

(v) 1,5 Approximate value of s.d. is ---

a. 0.42 **b.** 42 **c.** 4.2 **d.** 4

2. Pearsonian Coefficient of Skewness ϒ1 & Kurtosis ϒ2 based on moments for 1000 observations on variable Y are respectively 1.5 & 4.5. Define a new variable **U = (X – 1) / h**, where h > 0 then for this new variable U, (ϒ1, ϒ2) is given by ---

a. (1.5 , 4.5) **b.** (1.5**/h**, 4.5**/h**)

c. ((1.5 –1)**/h**, (4.5 –1)**/h**) **d.** (4.5, 1.5)

3. Which of the following statement(s) is not correct?

a. Skewness, kurtosis & dispersion are the properties of a distribution expressing lack of symmetry, relative peakedness & variation of distribution respectively.

b. Skewness, kurtosis & dispersion of a distribution can be measured respectively by (mean-mode)/σ, β_2 & C.V.

c. Multiple bar diagram can be used to study Skewness and dispersion of a distribution.

d. Histogram is used to study kurtosis of a distribution.

4. For a certain data first four central moments are 31.4, 95, -247 and 18890. The distribution of the variable is likely to be ---

a. negatively skewd and platykurtic

b. negatively skewed and leptokurtic

c. positively skewed, but platykurtic

d. symmetric and mesokurtic

5. Which of the following measure is used to measure the kurtosis of a frequency distribution ?

a. (A.M. – mode)/ σ **b.** $[(Q_3 – Q_2) - (Q_2 – Q_1)] / (Q_3 – Q_1)$

c. γ_2 **d.** Only **a.** and **c.**

6. Let μ_r' and μ_r respectively denote r^{th} raw and central moment of a random variable. If $\mu_1 = \mu_3 = 0$ and $\mu_2 = 2$, then μ_3' is ---

a. 0 **b.** 6 **c.** 3 **d.** cannot be computed.

7. Following are the data on number of seeds set in a pod. Prepare appropriate frequency distribution, draw graph and analyze these data.

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

8. For a moderately skewed distribution mean is 17.2, s.d. is 50 and median is 16.7; obtain the coefficient of skewness and mode.

9. Derive the relationship between r^{th} central moment in terms of raw moments of raw moments. Verify your results for r = 1, 2, 3 and 4.

10. Analyze the data set on

(a) Daily rainfall (in cm) for a period of 40 days observed at certain locality

10	3	1	6	6	10	4	1	6	1
1	1	6	2	6	6	4	0	2	6
0	5	6	4	6	3	10	6	6	0
3	3	2	6	6	10	6	6	4	10

(b) Height of 100 students in cm.

167.6	155.7	165.1	155.6	164.6	160.0	160.0	154.4	155.0	154.5
156.9	155.6	173.3	171.4	153.1	160.4	153.6	162.8	151.9	159.7
160.0	163.6	164.2	156.8	156.5	160.6	160.6	166.7	161.7	164.9
164.1	158.9	163.8	163.1	162.7	160.6	153.8	156.4	153.6	161.3
160.0	159.7	159.5	158.7	166.3	160.0	155.4	165.5	159.9	160.9
160.5	158.8	160.0	157.2	164.2	159.8	163.4	154.1	156.5	158.2
161.7	169.4	162.1	155.0	159.7	163.6	158.4	164.3	160.3	167.2
164.7	151.9	159.8	159.9	164.3	157.6	162.3	155.8	153.4	162.0
157.9	152.7	165.7	160.1	164.8	155.6	154.6	161.0	155.6	158.9
156.8	164.4	170.9	168.4	166.8	162.7	151.5	154.7	163.5	153.8

Unit 7 : Correlation and Regression

7.1 Overview

The quantitative methods described so far are for summarizing the information with only one kind of data (characteristic or variable). But we need something more when two or more variables are to be studied simultaneously. In this unit we study methodology of studying two variables simultaneously. The nature of relationship between two variables is a subject matter of this unit. In it we will study scatter diagram, Karl Pearson's coefficient of correlation based on moments and Spearman's correlation coefficient based on ranks. We will also specify the computational details, interpretation and limitation of correlation coefficient. In the latter part of this chapter we will use the method of least squares to build regression equations. We will also explain these concepts with the help of simple examples and illustrations.

7.2 Objectives

After studying this unit, you will be able to -

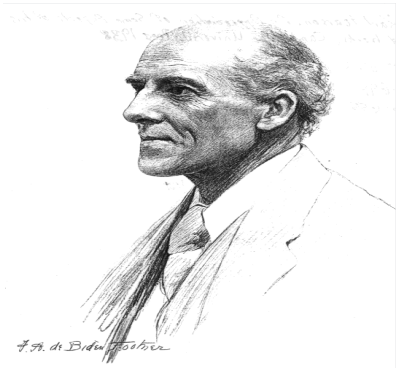
- Draw and interpret scatter diagram,
- Explain the concept of correlation coefficient,
- State properties of Pearson's Correlation Coefficient,
- Obtain correlation coefficient for a data,
- Define Spearman's Rank Correlation Coefficient,
- Apply method of least squares to obtain regression equation(s),
- State and use properties of regression coefficients,
- Interpret correlation and regression coefficients.
- Understand uses and limitations of correlation and regression.

7.3 Introduction

So far, we have studied elementary statistical methods to summarize data on one characteristic (variable). But there are many situations wherein we need data on two or more variables on each unit on which the data are collected. This is because we are interested to know whether the changes in one variable are associated or dependent on changes in another variable. Researchers are usually interested in the study of relationship between two or more variables.

Let us denote two variables by X and Y. For example, X can be the dose of a fertilizer and Y yield of a crop for a field. We expect that higher the dose of a fertilizer usually should give higher yield of a crop. Of course, this phenomenon is expected to be true within a certain band of fertilizer values. This is a case of **positive correlation**. If increase in values of one variable is associated with increase in values of other variable, we call it as a case of positive relationship.

For example as income increases, expenditure on luxury items may increase. Height and weight of growing children is positively correlated. It is also possible that increase in one variable is related with decrease in the other variable. It is then the case of **negative correlation.** For example, (i) age of a scooter and its fuel efficiency may be negatively correlated. (ii) More the time a student spends on watching T.V. serials or cricket matches his performance in examination may decrease. This is certainly not a rule, but is a general impression. It needs to be checked and supported with real data. In the theory of economics it is mentioned that as the price of a non-essential commodity increases, its demand goes down. If increase or decrease in one variable has no effect on changes in the other variable, then we simply say that the variables under consideration have **zero correlation.** Consider the case of salt consumption by an individual. We do not expect that as the income increases, consumption of salt will increase or decrease. We therefore would expect zero correlation between income of an individual and salt consumption. **Here the largeness or smallness of the value of a variable is a relative term. What is really understood by positive correlation is that above average values of one variable often correspond to above average values of the second variable. Similarly by negative correlation we mean above average value of one variable often correspond to below average value of another variable.** British Biometrician Karl Pearson (1867-1936) measured the heights of 1,078 fathers and their sons at maturity. He has defined index of correlation (r). It always lies between -1 to +1. Value of -1 indicates a perfect negative linear relationship, value of +1 indicates a perfect positive linear relationship, and intermediate values indicate the strength of relationship and also the direction. The question that you may ask is, "Can we guess the nature of relationship between the two variables?" Yes, we can do so with reasonable judgment if we plot **scatter diagram of the data.**



Karl Pearson (1867-1936)

A plot of Y Vs X values is called a scatter diagram. If a correlation is +1 or -1 all points (x_i, y_i) fall on a line with positive slope or on a line with negative slope. In general the data points scatter around a line showing the main trend. We give below examples of various scatter diagrams.

If two variables on the unit being studied are related, it is possible to use value of one variable to predict the value of other variable. This is done by regression equation. In this unit we will first study correlation and then regression.

7.4 Scatter Diagram

Many times our interest is in the study of relationship between two different measurements or observations in a dataset. One simple way to examine such a relationship is to draw a scatter diagram. Remember following points it.

1. It is the visual display for bivariate data (X, Y),
2. It displays the relationship between two variables measured on the same individual,
3. It is customary to plot explanatory variable or quantity on the x-axis and response variable on the y-axis. If you have no theory to justify which of the two variables is cause or effect, it does not matter which one you place on X or Y axis,
4. Scatter diagram may indicate presence of unusual observations or non-linear relationship. Study it carefully.

Examples:

Smoking index and health status: It is many times remarked that tobacco chewing or cigarette smoking has a serious effect on physiological parameters. Such statements can be checked using **appropriate** dataset.

Skin-fold thickness and weight: Here you may wish to explore the relationship between the two variables. Both are responses of age and other factors. You may put either one on either axis.

Scatter diagrams: We give below several examples of scatter diagrams. Study them. It will guide you to judge the relationship between two variables.

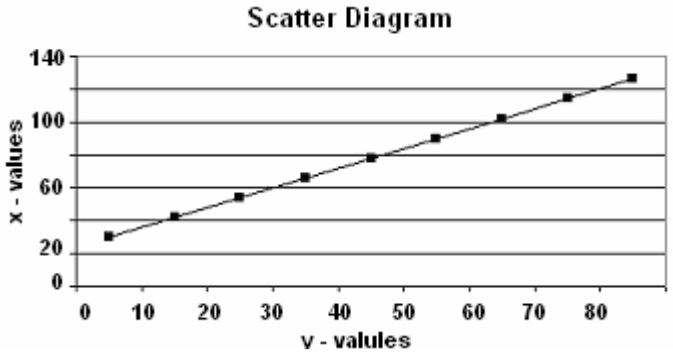


Figure 7.1 (a): r = 1

If variables are perfectly and positively correlated then all points lie on the rising line.

Figure 7.1 (b): r = - 1

If variables have are perfectly and negatively correlated then all points lie on the falling line.

Figure 7.1 (c): r = 0

Scatter diagram shows that X and Y are not correlated.

It is of interest to see how scatter diagram look like for various values of correlation coefficients. We give below some typical examples of scatter diagram. for the data on (i) monthly income and expenditure on food items of 40 families, Fig 7.1 (d), (ii) monthly income and relaxation time (in hours) per week of the household , Fig. 7.1 (e), and (iii) monthly income and yearly expenditure on education of these 40 families, Fig 7.1 (f). Various such datasets on wide variety of subjects are readily available.

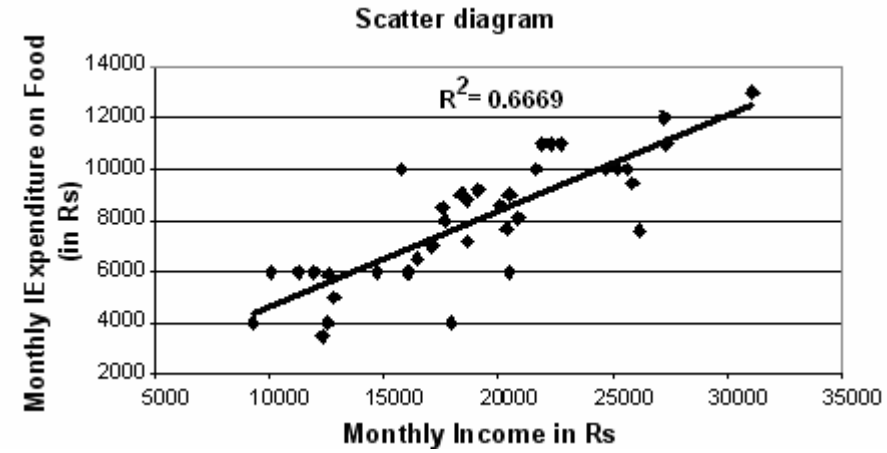


Figure 7.1 (d): r (Income, Expenditure) = 0.8166

Figure 7.1 (e): r (Monthly income, Weekly relax time) = - 0.8

Figure 7.1 (f)
r (Monthly income, Yearly Expenditure on education) = 0.0816

7.5 Karl Pearson's Correlation Coefficient, r (X, Y)

Given the sample data on pairs of (X, Y), it is easy to calculate correlation coefficient between them. The formula is simple.

Pearson's Correlation coefficient 'r' = [Cov(X, Y)] / [S.D.(X)*S.D.(Y)]

where $\mathbf{Cov(X,Y) = \frac{\sum xy}{n} - \bar{x}\bar{y}}$, $\mathbf{S.D.(X) = \sqrt{V(X)}}$, $\mathbf{S.D.(Y) = \sqrt{V(Y)}}$

Notice that Karl Pearson's correlation coefficient is based on moments. To compute it, we may first find

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $V(X) = \frac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2$, $V(Y) = \frac{\sum_{i=1}^{n} y_i^2}{n} - (\bar{y})^2$; hence find S.D.(X) and S.D.(Y). Further obtain $\mathbf{Cov(X,Y) = \frac{\sum xy}{n} - \bar{x}\bar{y}}$ and then using the above formula evaluate r (X, Y).

We will illustrate these computational details for the dataset on mathematical ability and coping level of students, in the following example.

Example: In one of the study on factors affecting "coping level" of an individual it was remarked that "Coping level" and "Mathematical ability" are highly related with each other. To check the claim of the researcher, undergraduate students of department of Psychology designed a questionnaire and collected relevant data. Part of these data on score in mathematical ability (X) and coping level (Y), for 10 students is reported in Table 7.1

Table 7.1: Mathematical ability (X) and Coping level (Y) of ten students

Student' no	406	431	511	514	522	540	542	602	605	619
X	40	70	80	90	110	130	140	150	160	180
Y	50	60	75	80	120	130	150	155	170	190

Figure 7.2: Scatter diagram of Coping level Vs Mathematical

It is always desirable that we draw a graph of (X, Y) pairs. The scatter plot (Figure 7.2) indicates that r (X, Y) is positive.

Table 7.2: Calculation of Pearson's correlation coefficient r(X, Y)

Sr. No.	Mathematical Ability (X)	Coping Level Y	X^2	Y^2	XY
1	40	50	1600	2500	2000
2	70	60	4900	3600	4200
3	80	75	6400	5625	6000
4	90	80	8100	6400	7200
5	110	120	12100	14400	13200
6	130	130	16900	16900	16900
7	140	150	19600	22500	21000
8	150	155	22500	24025	23250
9	160	170	25600	28900	27200
10	180	190	32400	36100	34200
Total	1,150	1,180	150,100	160,950	155,150

Here,

(i) $\bar{x} = 1150/10 = 115$; $\bar{y} = 1180/10 = 118$

(ii) **V(X)** $= (150100 / 10) - (115)^2 = 1785$; **S.D.(X)** $= 42.25$

(iii) **V(Y)** $= (160950 / 10) - (118)^2 = 2171$; **S.D.(Y)** $= 46.59$

(iv) **Cov(X, Y)** $= (155150 / 10) - (115*118) = 1945$

and hence r (X, Y) = 0.9880**,** a high positive correlation. This data indicates that mathematical ability and coping level of an individual is highly positively correlated. Why? What does it mean? Can it be accepted as a general rule? These are good questions but statisticians alone cannot answer them. It is necessary that such results be interpreted carefully. Subject experts and statisticians must work together and then conclusions should be drawn. We cannot and should not make sweeping statements on the basis of little evidence. If necessary more data must be collected and examined again. A word of explanation is necessary. In case of data, collected in sample surveys or generated in experiments, it is rare to get sample correlation coefficient of the **order of 0.9 or more** on either side of zero. If at all we get it, it is advisable to check the calculations and also the methods of collecting such data.

Is computed value of r statistically significant?

Does it differ from a zero?

Is statistical significance a function of sample size?

These important questions can be answered. But for more insight into these aspects, prerequisite is knowledge of some elementary probability theory and statistical inference, and topics such as testing of hypothesis. To know more about it, study reference books such as "A Course on Parametric Inference" by Dr. B. K. Kale and to know the test procedures refer books "Common Statistical Tests" by Kulkarni M. B., Ghatpande, S. B., Gore S. D. (1999) We will answer above questions while studying some common statistical tests later in this book.

Some more remarks on computations of r

We give below various formulae that are used to obtain correlation coefficient 'r'. All these formulae give the same answer. This is not at all surprising. It must happen because these formulae are equivalent. You can verify them easily.

Let X and Y denote the variables. Let $\mathbf{x = X - \bar{X}}$, denote deviation of X from mean of X and $\mathbf{y = Y - \bar{Y}}$, denote deviation of Y from mean of Y. Then we can compute r (X, Y) by using any one of the following formulae:

1. $\mathbf{r(X, Y) = Cov(X, Y) / [S.D.(X)*S.D.(Y)]}$
2. $$\mathbf{r(X, Y) = \frac{\sum_{i=1}^{i=n}(x_i y_i)}{\sqrt{\sum_{i=1}^{i=n} x_i^2 \sum_{i=1}^{i=n} y_i^2}}}$$
3. $$\mathbf{r(X, Y) = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{(\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2)(\sum_{i=1}^{n} Y_i^2 - n(\bar{Y})^2)}}}$$

7.6 Properties of Correlation Coefficient 'r'

1. The correlation coefficient is a **pure** number, **without any units**. It is not affected by interchanging the two variables,
2. It always lies between – 1 to +1,
3. r is **independent** of change of origin and scale, i .e., if $\mathbf{U = (X - a) / h}$ and $\mathbf{V = (Y - b) / k}$, then $r(X, Y) = \frac{hk}{|h||k|} r(U, V)$

7.7 Applications of Correlation in Various Fields

Pearson's correlation coefficient is most widely used in diverse fields of study. We can give a series of examples and list of correlated variables. For example,

(a) Income and expenditure of a family, price and demand are correlated variables.

(b) Pollution levels are correlated with population density, traffic intensity, wind velocity, temperature, etc.

(c) Yield and fertilizer dose are correlated variables.

(d) Germination time of a crop is related with maximum and minimum temperature, hours of sunshine, soil moisture at different levels, etc.

(e) Abundance of insects is related with habitat quality, plant density, grassland type, etc.

(f) Weight, height, age, arm circumference, head circumference are correlated to each other.

(g) Length and weight of a seed are correlated. Number of seeds set in a pod and their total weight is directly related with length of a pod.

(h) Correlation between dry fodder and milk yield, green fodder and milk yield, hours of grazing and milk yield are of importance in animal husbandry.

(i) Stress level, type of work, educational levels and degree of managerial responsibilities, spouse's behavior are correlated variables.

Remember that "r" is calculated in case of bivariate data. This fact should not be overlooked. For example, let X be the proportion of female employees in an organization, Y be the average number of absentees per month per employee. You may get a high positive correlation coefficient. But calculating correlation coefficient between X and Y is inappropriate. This is misuse of statistics.

Some more remarks on "r"

Remember that Pearson's correlation coefficient measures only **linear relationship** between two variables. It **can't be used** to measure nonlinear relationship. We illustrate this aspect with the help of two datasets. The first dataset is reported in Table 7.3. Try to guess the relationship between values of Y and X. There is a functional relationship between X and Y, bur r (X, Y) = 0.

Table 7.3: Case of r (X, Y) = 0, where Y = f(x)

X	-4	-3	-2	-1	0	1	2	3	4
Y	16	9	4	1	0	1	4	9	16

Study the scatter diagram in Figure 7.3 carefully. Notice that Y and X are perfectly related with each other, the relationship between them can be expressed by $y = f(x) = x^2$. It is an equation of parabola. It is easy to verify that for these data r (X, Y) = 0. This is left as a part of self-study. Let us study a data set reported in Table 7.4. It is about Work capacity (output) and Energy Intake (in Kcal) of an individual. Study the scatter diagram for this dataset.

Figure 7.3: Scatter diagram

Table 7.4: Data set on X: Energy Intake (In Kcal) and Y: Work output

X	Y	X	Y	X	Y	X	Y
1600	300	2600	340	3500	260	1800	350
2800	340	3600	220	2000	360	3000	330
3500	250	2200	360	3200	300	4000	275
2400	360	3400	280				

The diagram in figure 7.4 indicates that there is a **non-linear** relationship between food intake and work capacity of an individual. To begin with as intake increases, work output also increases, thereafter even if intake increases, work output does not change – it remains more or less at the same level, but if food intake increases beyond a certain threshold value, then work output decreases. If you find correlation coefficient between food intake and work output for these data, you will get it close to zero. It appears surprising. Why? Work efficiency does not increase in a linear manner as a function of food intake. In practice it is seen that up to a certain level it increases in a linear manner, but not all the time. Work output may decrease if food intake increases beyond a particular level. It is now accepted that two individuals belonging to the same age-sex group, similar in stature and engaged in same level of activity, can do the same amount of work even if one of them eats twice the other. It is not the amount of food that you eat but conversion efficiency of food to energy molecules (ATP bonds) that matter when we deal with work output.

Figure 7.4: Scatter diagram

We have seen that

- "Curvilinear relationship cannot be measured using Pearsonian coefficient of correlation".
- Correlation coefficient equal to zero does not necessarily imply that there is no relationship between the two variables being studied.

7.8 Spearman's Rank Correlation Coefficient (R_S)

There are many situations where we don't have exact measurements on the variables under consideration. In such a situation Pearson's correlation coefficient cannot be calculated. There are other situations where the data include a few extremely large or extremely small values. These values affect the moments seriously and hence also the value of the Pearson's correlation coefficient. In such cases we could like to have a coefficient of association that is more robust. In such a situation Spearman's correlation coefficient is appropriate. For example, In case of beauty contest – such as Miss India or Miss Universe, - suppose two experts rank n participants giving them ranks 1 to n. What is the extent of agreement between the two experts?

Ratings allotted or ranking given to a unit is qualitative, although there is order in those ranks. Spearman's rank correlation coefficient is simply Pearson's correlation coefficient for the ranked values of x and y. We will illustrate its use with the help of a simple example.

Example: Grader's rating

An experiment was conducted to study the relationship between the ratings awarded to a set of eight tobacco leaves with respect to their moisture content by an amateur student and the actual moisture content of the corresponding tobacco leaves. The data are reported below in a table form. If the moisture content of a leaf is on higher side then it is regarded as a poorer quality tobacco.

Table 7.5: Grader's rating and moisture content of a leaf

Leaf No.	1	2	3	4	5	6	7	8
Grader's rating	7	4	2	6	1	3	8	5
Moisture content	0.44	0.72	0.69	0.70	0.93	0.82	0.67	0.80

Do the data suggest an agreement between the grader's rating and the moisture content? Can we conclude that there is an association between the two variables under consideration?

Here two variables of interest are rank and moisture content. The former is already in rank form. The moisture content may be ranked similarly. Ranking can be awarded either in ascending or descending order. But the same procedure should be applied while ranking values of each of the variable or attribute. Here the Pearson's correlation coefficient reduces to the formula given below to get the Spearman's rank correlation coefficient (R_S) where d_i = difference between two rankings for the i^{th} unit.

$$\mathbf{R_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}}$$

For the above data you may prepare following table and find R_S.

Table 7.6: Computation of Spearman's correlation coefficient

Leaf	Rank (x_i)	Rank (y_i)	d_i = difference in ranks	d_i^2
1	7	1	6	36
2	4	5	-1	1
3	2	3	-1	1
4	6	4	2	4
5	1	8	-7	49
6	3	7	-4	16
7	8	2	6	36
8	5	6	-1	1
		Total	0	144

Verify that $\mathbf{R_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}} = 1 - \frac{6(144)}{8(64-1)} = -0.714$.

In this particular problem the grader was an amateur student who was being trained to become the grader in the business world of tobacco. It indicates that an amateur student is yet to learn the technique of assessing the tobacco quality. He needs perhaps more training on estimating moisture content and then grading the tobacco leaf.

In the above example note that the data on the student's rating was only in the form of rank where as the actual moisture content was a real measured variable. We converted the second variable also in the form of ranks before computing the correlation coefficient. Even when both variables are actual measurements we may still want to convert them into ranks and compute the Spearman's correlation coefficient instead of the Pearson's correlation coefficient. In the above example and in the formula for Spearman's correlation coefficient, we have assumed that there are no ties, i.e. each of the observation is different from the remaining and hence a separate rank to it can be awarded. This is acceptable theoretically, but in practice ties do occur. In such a situation you may obtain directly Pearson's correlation coefficient, for the data on rank of X and rank of Y after breaking the ties. Spearman's rank correlation coefficient defined above is in fact **the same** as the Pearson's correlation coefficient that can be obtained for the ranked observations.

7.8.1 Spearman's Rank Correlation Coefficient: (A case of Ties)

Example: During annual social gathering, personality competition was conducted, in an undergraduate college. Two judges, A and B awarded scores to 15 competitors in a personality test. These scores, say X and Y, are given in Table 7.6

Table 7.6: Personality Scores by two Judges A and B

Judge A: X:	30	40	40	50	60	80	60	50	30	70	60	50	90	30	50
Judge B: Y:	32	35	37	38	42	50	43	35	30	40	39	36	52	28	28

To compute 'R_S', prepare following table.

Table 7.7: Computation of R_S (A case of tied ranks)

Sr. no.	Rank (X)	Rank(Y)	X^2	Y^2	*XY*
1	2	4	4.00	16.00	8.00
2	4.5	5.5	20.25	30.25	24.75
3	4.5	8.0	20.25	64.00	16.00
4	7.5	9	56.25	81.00	67.50
5	11	12	21.00	144.00	132.00
6	14	14	196.00	196.00	196.00
7	11	13	121.00	169.00	143.00
8	7.5	5.5	56.25	30.25	41.25
9	2	3	4.00	9.00	6.00
10	13	11	169.00	121.00	143.00
11	11	10	121.00	100.00	110.00
12	7.5	7	56.25	49.00	52.50
13	15	15	225.00	225.00	225.00
14	2	1.5	4.00	2.25	3.00
15	7.5	1.5	56.25	2.25	11.25
		Total	1230.50	1239.00	1199.25

Notice that in case of X, observation 30 occurs thrice, 40 occurs twice, 50 occurs four times and 60 occurs thrice, hence observation 30 was awarded rank 2, 40 got 4.5 and 50 got 7.5 whereas 60 got 11, similarly verify that in case of Y, observations 18 and 25 both occur twice, and hence they should be awarded average ranks 1.5 and 4.5 respectively. Now you can obtain Pearson's correlation coefficient 'r' to these ranked data. It is left as an exercise for you. This is nothing but Spearman's rank correlation coefficient. In this example it is equal to 0.8709; It is positive and high; it indicates that two judges agree with each other.

Notice that Spearman's correlation coefficient will not change at all if you change original values either of X or Y without changing their respective ranks.

7.9 Linear Regression (Bivariate data)

Regression analysis is the study of relationship among variables. It is a simple method, and is widely used, for investigating and establishing a functional relationship among variables. The relationship is expressed in the form of an equation. The response variable, or the variable to be predicted is on the left side of equation and on the right hand side we have one or more predictor variables. These predictor variables are also called as independent variables and the response variable is called the dependent variable. The equation that we are referring is called a regression equation and is of the form

$$Y = b_o + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

where b_0, b_1, b_2, …, b_p are unknown constants. These constants are called as regression coefficients and these are to be determined from the available data.

The simplest regression equation is $Y = b_o + b_1X_1$. Here Y is the response variable and can be predicted from a predictor variable X_1. If there is only one predictor variable X_1, then suffix 1 may be dropped and we will use symbol X only. Suppose a researcher has interest in exploring the relationship between monthly expenditure on luxury items and income of a family. He therefore collects a data from 100 families, on monthly income (X) and expenditure on luxury items (Y). He can build a relationship $Y = b_o + b_1X$. This equation can be used to predict monthly expenditure on luxury items of a family given monthly income of that family.

We can prepare a list of situations in which simple regression equation can be used. One such list is given in the following table.

Table: Situations in which regression equations can be applied

Sr. No.	X	Y
1	Supply of a commodity	Price of the commodity in the market.
2	# Hours of study	Marks in the annual examination
3	Height of father	Height of a daughter
4	Level of a fertilizer dose	Per acre yield of a crop
5	Total food intake	Gain in body weight
6	Working hours in an organization	Stress level of the manager
7	Weekly Iron intake	Hemoglobin Level
8	Gestational age	Head circumference of an infant
9	Mother's age	Weight of a newborn baby
10	Family size	Monthly savings

You may be curious to know the genesis of the word regression. To regress means to go back. It was Sir Francis Galton (1822-1911) a well-known British anthropologist who introduced the term "regression". Later on Galton also remarked "the offspring of seeds did not tend to resemble their parent seeds in size, but to be always mediocre (more average) than parents – to be smaller than the parents, if the parents were large; to be larger than the parents if the parents were very small ". We continue to use the term regression even today, but it is of historical importance only. It is used extensively in almost all fields of research. In socio-economic surveys or in scientific experiments, our aim is to investigate how the changes in one variable affect another variable. This dependence of one variable upon the other variable must be modeled by an appropriate function. You can certainly think of several such functions. The simplest way is to link two variables by a straight-line relationship. If we plot Y Vs X values then the data could indicate such a straight line. This we call as a simple linear regression. An example of a very complex regression equation would be predicting weather conditions (such as rainfall) using a number of variables such as maximum and minimum temperature at various places, humidity, wind velocity, wind directions, hours of sunshine, etc. In the study of regression all such problems can be studied. In this book we will restrict ourselves to study of two variables only, one response variable and one predictor variable. The determination of a regression equation is important. It can certainly be used to

- Judging the importance of individual predictor variables,
- Predicting values of Y for a given set of values of predictor variables,
- Understanding interrelationships among the variables.

7.9.1 Linear Regression: Fitting a Straight Line

Description of the data

We assume that we are given a bivariate data (x_i, y_i) on n units. These data may be reported as follows:

Unit no	y	x
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
:	:	:
:	:	:
n	y_n	x_n

A scatter plot of such data can give an idea about the relationship between the variables being studied. Let us assume that the data points in the scatter diagram cluster around a straight line given the equation $y = b_0 + b_1x$. Here we have a dataset having n points and we have only two unknown constants b_0 and b_1. **Our object is to find that line which best fits the data.** We therefore look for "best" solution. Here the word "best" must be explained.

We want to have the best fit line for y using x. We therefore find the line such that the sum of squares of distance of the data point from the line in the y-direction is minimum. In statistical jargon such a line would be available if we determine values of b_0 and b_1 by using method of least squares. It is easy to show that the unknown constants b_0 and b_1 can uniquely be determined if we solve following two equations. (These are called normal equations and are obtained by using calculus):

$$\sum y_i = nb_0 + b_1 \sum x_i$$

$$\sum y_i x_i = b_0 \sum x_i + b_1 \sum x_i^2$$

Solving above two equations for $\mathbf{b_1}$ we get,

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}$$ and hence

$$b_0 = \bar{y} - b_1 * \bar{x}$$

Now we can write equation of a regression line of Y on x as $y = b_0 + b_1*x$.

The regression equation of Y on X can also be written as $\mathbf{y - \bar{y} = b_1 * (x - \bar{x})}$

It is easy to show that $\mathbf{b_1 = r\ [S.D.(Y) / S.D.(X)]}$. It suggests that as the standard deviation of X gets larger, it will pull the regression line down, flatten it out or decrease the slope. So when the X-variable has a large standard deviation relative to the Y-variable, slope of the regression line would be small.

If it is known that one of the variable is cause (X) and the other is effect (Y), then obtain regression equation of Y on X. It may happen that the two variables are associated but there is no cause and effect relationship then both the regression equations, ie. of Y on X and of X on Y can be obtained. Remember that while deriving regression equation of X on Y, we find a line such that the sum of squares of distance of the data point from the line in the x-direction is minimum. Let us denote such a line by equation $X = b'_0 + b'_1*Y$. It is easy to show that

$$b'_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} = \frac{S_{xy}}{S_y^2}$$ and hence

$$b_o' = \bar{x} - b_1' * \bar{y}$$

We will illustrate computational details with the help of an example.

Example: In a study of "Monitoring and Documenting Biodiversity of India", Biodiversity Study Group of Nashik (BSGN), collected data on abundance of various plant species at different locations nearby Nashik. Amount of effort (in hours) required for observing the plant species in the various localities is given below. It is of interest to estimate the relationship between # species observed and units of efforts put in.

Units of "effort"	10	12	14	20	24	26	30	40	11	11
# Species observed	15	18	18	20	22	24	28	30	17	16

Units of "effort"	13	14	24	26	21	21	20	30	40	35
# Species observed	19	20	24	26	21	23	18	30	32	31

In order to obtain regression equations of Y on X (or of X on Y), prepare a table such as given below, but even before that draw a scatter diagram and think of possible relationship between efforts required to observe the number of species in a given locality.

The scatter diagram indicates that more the effort, more are the number of plant species seen in a given locality. (This statement is valid obviously within certain range of units of efforts and depends on several other environmental factors).

Sr. No.	*Units of Effort (X)*	*# new speices observed (Y)*	X^2	Y^2	*XY*
1	10	15	100	225	150
2	12	18	144	324	216
3	14	18	196	324	252
4	20	20	400	400	400
5	24	22	576	484	528
6	26	24	676	576	624
7	30	28	900	784	840
8	40	30	1600	900	1200
9	11	17	121	289	187
10	11	16	121	256	176
11	13	19	169	361	247
12	14	20	196	400	280

Sr. No.	*Units of Effort (X)*	*# new speices observed (Y)*	X^2	Y^2	*XY*
13	24	24	576	576	576
14	26	26	676	676	676
15	21	21	441	441	441
16	21	23	441	529	483
17	20	18	400	324	360
18	30	30	900	900	900
19	40	32	1600	1024	1280
20	35	31	1225	961	1085
Total	**442**	**452**	**11458**	**10754**	**10901**

It is now easy to verify that

n = 20, mean of x = $\bar{x}$= 22.1, mean of y = $\bar{y}$= 22.6,

$$b'_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{1}^{n} x_i^2 - n(\bar{x})^2} = \frac{10901 - 20*22.1*22.6}{11458 - 20*(22.1)^2} = \frac{911.8}{1689.8}$$

= 0.539505

i.e, here S_{xy} = 911.8 and S^2_x = 1689.8 and hence b'_1 =0.539505

And therefore $b_0 = \bar{y} - b_1 * \bar{x}$ = 22.6 – 0.539505*22.1 = 10.6705

Thus we now can state regression equation of number of species observed (Y) on units of effort required (X) as

Y = 10.67 + 0.54*X

We must now interpret the regression coefficients in the above equation. The constant term represents that in a given locality about 11 species can be seen with perhaps no "effort" (zero effort), The coefficient of X indicates that increase in the effort required for observing **each additional** plant species in the locality under consideration.

The above regression equation can be used to predict the number of plant species that can be seen in a given locality for various values of effort. For example if effort is 10, 15 and 20 units then the average number of species that can be found are respectively 16.07, 18.77 and 21.47. A word of caution is necessary here. You can use such a regression equation within the range of observations. The regression equation should not be used for predicting purposes for values far outside the range of observations. For example, if in the above regression equation estimate of number of species that can be seen equals 64.67 for 100 units of effort. This is usually unlikely to happen. It is the number of species in a given locality depends upon several biological and environmental factors.

Verify that for these dataset $r^2(X, Y)$ = 0.91, which is equal to $(b_{1*}b'_1)$. Since r^2 is always positive, it implies that the slopes of the regression lines must be of same sign.

Since both the regression lines lie on the same plane, they will always intersect each other. The point of intersection is $(\bar{x}, \bar{y})$ and the acute angle θ between these lines is given by

$$\theta = \tan^{-1}\left\{\frac{|b_{yx} * b_{xy} - 1|}{|b_{xy} + b_{yx}|}\right\}$$

Remember following points:

1. It is necessary to test the significance of a regression coefficient. For testing the significance of the regression coefficients (H_o: $\beta_1 = 0$ Vs H_1: $\beta_1 \neq 0$,) use statistical test. For details, refer to a book "Common Statistical Tests" by Kulkarni M. B., Ghatpande S. B. and Gore S. D. (1999).
2. You can also test H_o: Population correlation coefficient $\rho = 0$ using sample data.
3. Do not use a regression equation to predict Y for x values outside the range of observed cases. Linear relationship may not hold good.
4. The r^2 value is a measure of model fit. Is this model good at predicting values of Y? r^2 tells us the fraction of the V(Y) accounted for by the regression line. Remember $0 \leq r^2 \leq 1$. More the r^2 better would be prediction.
5. Refer: Anscombe's Quartet (F. J. Anscombe, "Graphs in Statistical Analysis," American Statistician, Feb 27, 1973, pp.17-21). It gives scatter plots for 4 different datasets, each with different values for X and Y. Each data set has same mean of X, of Y, the regression line is $Y = 3 + 0.5X$, same $r^2 = 0.67$. Moral of the story is – the numbers don't tell you everything. If datasets have the same numerical summary, this does not mean that they also have the same picture. The combination of a graph and numbers give you a better idea of what the data looks like.
6. If Y is to be predicted using X as a predictor variable, use regression equation of Y on X.
7. Verify that $0 \leq bb' = r^2 \leq 1$; it means b and b' must have same sign.
8. You can transform variables (X, Y) to (U, V), by making change of origin and scale. Let $U = (X-a)/h$ and $V = (Y-b)/k$, then $b_{yx} = (k/h)*b_{vu}$ (assume h and k both positive). You may first obtain regression equation of V on U and then reconstituting obtain regression equation of Y on X.
9. We can study multiple regression models involving 3 or more variables and also non-linear regression models. For details refer standard books on regression analysis such as Regression Analysis by Feund, Wilson and Ping Sa, Academic Press, 2006.
10. You can use EXCEL package to plot data, and obtain regression equation.
11. We have not discussed procedure for inference of regression coefficients and for predicted values. For it use reference books.

7.9 Summary

In this unit we have studied important statistical techniques – scatter diagram, Pearson's correlation coefficient, Spearman's Rank correlation coefficient and elements of regression analysis. We then studied properties of correlation coefficients and of regression coefficients. We have also specified various situations wherein these techniques can be used keeping in mind the limitations.

7.10 Self Test

1. In case of two variables X and Y which of the following statements are true?

a. If the covariance between two variables X & Y is equal to zero, then correlation coefficient between these variables equal to zero.

b. If the covariance between two variables X and Y is equal to zero then they are independently distributed.

c. For independent variables X and Y, correlation & covariance are both always zero.

d. r (X, Y) = r(Y, X)

2. If X and Y are any two random variables with s.d. unity each and covariance between them is equal to 0.60, then r(X/2, Y/2) is given by ---

a. 0.60 **b.** 0.30 **c.** 0.15 **d.** 0.75

3. Which of the following statement is true?

a. Regression coefficient always lies between 0 and ∞.

b. Correlation coefficient always lies between 0 and 1.

c. Shift of origin has no effect on the value of $r_{X,Y}$.

d. Shift of origin affects the absolute value of correlation coefficient.

4. Which of the following statement are false?

a. b_{yx} and b_{xy} are either both positive or both negative.

b. b_{yx} and b_{xy} both can exceed unity.

c. b_{yx} and b_{xy} both can be less than unity.

d. b_{yx} = r V(X)/V(Y).

5. If r = o then it means ---

a. two variables are independent.

b. two variables have functional relationship.

c. regression lines pass through origin.

d. X and Y do not have linear relationship.

6. Suppose correlation coefficient between rainfall as measured in cm and production of jowar measured in metric tone is 0.50. What is the correlation coefficient of rainfall measured in mm and production measured in Kg?

a. 0.50 **b.** 0.05 **c.** (0.005*2.2)/1000 **d.** Can't be derived

7. Let $Y = \overline{Y} + b_{yx}\left(X - \overline{X}\right)$ & $X = \overline{X} + b_{xy}\left(Y - \overline{Y}\right)$ be regression equation of Y on X and of X on Y respectively. Then which of the following statement is false?

a. The point of intersection of two regression lines is $\left(\overline{X}, \overline{Y}\right)$

b. The acute angle between the two regression lines is $\theta = \tan^{-1}\left\{\frac{\left|\mathbf{b_{yx} * b_{xy} - 1}\right|}{\left|\mathbf{b_{xy} + b_{yx}}\right|}\right\}$

c. Both regression coefficients cannot exceed unity simultaneously.

d. Both regression coefficients have opposite algebraic signs.

8. In spite of a progress in science and technology, at the time of marriage "Janmakundali" is seen. Following are the data on the score awarded to janmakudlis of bride and bridegroom (X) and their satisfaction score (Y) as judged by couple itself. Draw scatter plot, find r (X, Y) and also explore the relationship between the two variables under consideration.

Table 7.6: Data on Satisfaction score and Score in the "Janmakudali"

Y	X	Y	X	Y	X	Y	X
19	19	65	28	14	9	11	7
40	19	65	29	16	7	12	28
42	29	65	25	16	12	12	28
42	29	66	29	6	8	18	28
47	31	66	30	7	10	18	9
49	32	66	25	7	8	18	17
50	36	67	17	28	17	19	10
70	34	75	16	30	29		
88	31	94	19	36	17		

9. Collect data on height of father, mother and their children. Develop regression equations to predict

(i) height of a son using data on (a) mother's height, (b) father's height,

(ii) height of a daughter using data on (a) mother's height, (b) father's height.

10. Following are the data on girth at breast height (in cm) and height of a eucalyptus tree. Plot scatter diagram and find r(GBH, HT). Find regression equation of height of a eucalyptus tree on its girth at breast height.

GBH	**HT**	**GBH**	**HT**	**GBH**	**HT**	**GBH**	**HT**
5.0	1.95	22.0	6.15	13.0	4.60	10.5	4.95
9.0	3.88	4.0	2.30	4.0	2.70	6.0	2.90
12.0	5.15	11.5	4.60	6.0	2.93	14.0	5.40
13.0	5.88	17.0	4.80	10.0	4.08	12.0	3.90
10.0	3.68	4.0	2.80	14.0	4.10	4.0	2.30
8.0	2.94	13.0	4.25	7.0	2.90		

11. Teachers A and B assessed papers of 40 students independently and awarded marks. These marks are given below. Analyze the data sets. Can you predict the marks that teacher B would award if you know the marks awarded by teacher A? If yes, explain the procedure that you may employ.

A	B	A	B	A	B	A	B
16	10	14	10	21	10	16	8
9	8	9	10	5	4	23	20
15	15	27	26	19	8	16	9
22	24	23	20	9	11	15	13
24	21	19	17	19	14	24	24
18	14	12	15	20	17	27	22
18	14	8	12	11	15	12	8
25	15	20	14	17	14	12	12
21	13	18	12	26	20	11	13
17	16	16	17	22	20	16	16

Unit 8 : Probability

8.0 Overview

In this introductory unit on probability you will be introduced to basics of probability. In it you will study deterministic and non-deterministic experiments. You will also study sample space and events associated with it. You then will be introduced to classical definition of probability, axioms of probability. After learning basic probability you will be introduced to theorems on probability and then you will be introduced to concept of conditional probability and independence of two events. Knowledge of counting techniques is a prerequisite for study of theory of probability.

8.1 Learning Objectives

After studying this unit, you will be able to

- Describe deterministic and non-deterministic experiments (models)
- List the elements of Sample space (finite and countably infinite)
- Identify events and work on algebra of events
- Describe probability models and explain classical definition of probability
- State Axioms of probability
- Find probabilities of various events and solve simple problems
- Describe and evaluate conditional probability
- Explain concept of independence and solve related problems.

8.2 Introduction

Introductory lessons about Theory of Probability begin with simple illustrations of (i) tossing of a coin, (ii) rolling of a six faced die, (iii) drawing of a ball from an box containing balls of different colours and other examples of a similar type. These examples are used to bring out the distinction between "random" experiments and the non-random experiments like the "experiment" of allowing a rigid body to fall from a height. If the coin is fair or the die is not loaded then it would be impossible to predict the outcome of the experiment before it is performed. In the case of the experiment of allowing a rigid body to fall from a height we can predict the final result with certainty before the experiment is carried out. The study of probability is concerned with study of random experiments, i.e. with experiments, which have more than one possible outcome, and for which it is impossible to predict the outcome before the experiment is performed. The probability theory is used to measure the degree of uncertainty involved in the problem. In our day- to- day life, when confronted with situations, we take a certain course of action. A XIIth standard passed student has to decide the future course of study. A newly recruited employee has to decide whether to opt or not to opt for medi-claim policy. You may have doubts whether your decisions are reasonable or not. The study of the probability of the occurrence of an event will help you to clear these doubts. The theory of probability has wide applications: in the biological, engineering and social sciences, and also in education.

Before getting into the probability, the concept of random experiments, outcomes and sample space needs to be discussed. This is the topic of the next section.

8.3 Random Experiments

Experimentation is an integral part of any learning process. For some of the experiments, the results are specific and known in advance. Such experiments are called **deterministic or non-random experiments**. Examples of deterministic experiments are

- Dropping of a pebble from a height.
- Introducing a spark of electricity in a cylinder containing a mixture of hydrogen and oxygen and noting the end product.
- Heating water in a pot for a sufficiently long time to a temperature in excess of 100^{o} C.

You can prepare a list of deterministic experiments. In such experiments, the outcomes are certain and can be predicted even before the experiment is performed. The results of experiments are labeled as **outcomes.** A particular repetition of an experiment is known as a **trial**. In the study of theory of probability our interest is in random experiments.

Random Experiment: An experiment in which the set of all possible outcomes are known, but the exact outcome of a particular trial is unknown prior to conduct of the experiment, is called a random experiment. It is possible that you have some idea or guess or intuition how the experiment will turn out but you cannot predict its outcome with certainty. Hereafter whenever we use the word experiment, we mean it to be a random experiment.

Sample Space: Each trial of the experiment will always result in an outcome. We can think of the set of all possible outcomes of a trial. This set is called a sample space and is usually denoted by either letter S or by a Greek letter Ω (Omega). If the number of elements in the sample space is finite then it is a finite sample space. For example, in case of an experiment of tossing of a coin the sample space S = {Head, Tail}. It is a finite sample space. It contains finite number of elements. The number of elements of S may be countably infinite. Such a sample space is called countably infinite sample space. For example, suppose you decide to toss a die till you observe '1'. Total number of tosses required would be 1, 2, 3 and so on. In other words, here sample space S = {1, 2, 3….}. Counting of the elements is possible but it may not end. This is an example of countably infinite sample space. We can also think of an uncountable infinite sample space. We will study it later. As you must have observed elements of the sample space are reported in curly brackets.

Exhaustive outcomes: If each trial of the experiment results in any one of the outcomes from the list of possible outcomes, then the list of outcomes is said to be exhaustive.

Mutually Exclusive Outcomes: If the occurrence of one of the outcomes prevents the occurrence of all other outcomes in that particular trial, then the outcomes are said to be mutually exclusive.

Examples of a random experiment:

1. A coin is tossed. The possible outcomes are head and tail. The outcomes are exhaustive and mutually exclusive.
2. A six-faced die is tossed once. The possible outcomes are 1, 2, 3, 4, 5 and 6. Here S = {1, 2…6}. The outcomes are mutually exclusive and exhaustive.
3. Observe change in BSE (Bombay Stock Exchange) index. The possible outcomes are increase, decrease and no change. Again these outcomes are exhaustive and mutually exclusive.
4. A student attempts an examination till he passes. The possible outcomes are 1, 2, 3…. The outcomes are exhaustive and mutually exclusive.
5. Consider the experiment of tossing of two coins together. Let H represents a head and T represents a tail. Since the possible outcomes are HH, TT, TH and HT, the sample space S will be S = {HH, TT, TH, HT}.

You may describe the sample space using a Venn diagram. You can also use a Tree diagram for the same. For example, consider the experiment of tossing of a six-faced die. Figure 1 is a Venn diagram (a closed geometric figure such as square or rectangle or circle). Alternatively you can make use of a tree diagram (see Figure 2). In it a branch of the tree represents each outcome. For example, in case of tossing of a coin two outcomes H and T are possible.

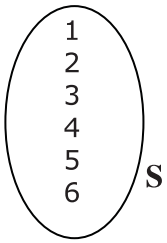


Figure 1: Sample space S

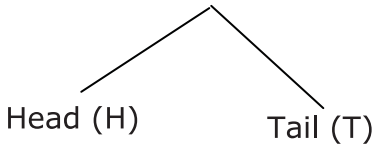


Figure 2: Tree diagram of sample space

Many times our interest lies in a subset of outcomes of the sample space. Therefore we need to study concept of events and algebra of events.

Events and Algebra of events A subset of the sample space is called an **event**. It is a collection of one or more outcomes (sample points) of a random experiment. We will use upper case letters with or without suffix, such as A_1, A_2, B, C, P, Q to denote events. Using the same sample space we can define several events. All events defined are essentially subsets of the sample space. Consider an experiment of tossing two coins together. The sample space S of this experiment is S = {HH, HT, TH, TT}. Let A denote the event of getting at least one head, then A = {HH, HT, TH}. If B denote the event of getting at the most one head then B = {HT, TH, TT}. If C denote the event of getting neither head nor tail then C= {}. This is empty set and will be referred as a **null event**. The sample space is also a subset of itself and is referred as a **sure event**. An event can be simple or compound. A **simple event** (also called as elementary event) contains only one of the outcomes of a sample space. **A compound event** is an event that has two or more of the outcomes of a sample space.

Example

Consider the experiment involving two dice together. The sample space is given by S = {(1,1), (1,2), ….(6,6)}. Define an event A_1 as that of getting '1' on both the dice. A_1 = {(1,1)}. Define A_2 as the event of getting '1' on at least one die. A_2 = {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (3,1), 4,1), (5,1), (6,1)}. Here A_1 is a simple event and A_2 is a compound event. From the definition of an event, it follows that the sample space S as well as the empty set ϕ are also events. The sample space is referred to as ***a sure event***, whereas the empty set ϕ is referred as **a null event**.

Operation (algebra) on events (Union and intersection)

All the operations on sets are applicable to events. The basic operations on sets are union and intersection. Let A and B denote two events associated with the same sample space (hereafter whenever we discuss two or more events, it will be assumed that they are associated with the same sample space unless otherwise stated).

$A \cup B$: The event $A \cup B$ (read as A union B) is the set of all sample points, which belong to either A or B or to both A and B.

$A \cap B$: The event $A \cap B$ (read as A intersection B) is the set of all sample points, which belong to both A and B. In the Venn diagram (Figure 3), the area indicated by arrow represents the intersection of two events.

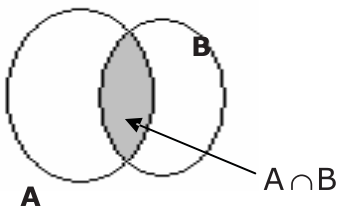


Figure 3: Venn diagram showing $A \cap B$

In the following Venn diagrams (Figure 4) the area of A, B $A \cap B$ and $A \cup B$ of is shown.

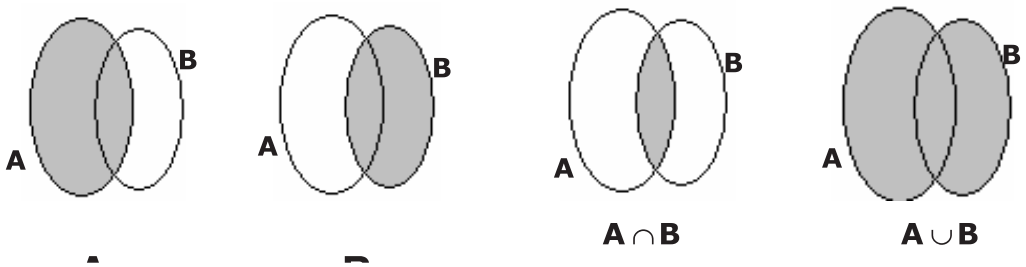


Figure 4: Venn diagram showing A, B, $A \cap B$ and $A \cup B$: shaded area

The concept of union as well as intersection of two events can be generalized to three or more events.

Complement of an event: We have defined an event as a subset of the sample space S. It may or may not occur. If event A does not occur the observed outcome of the experiment does not belong to A. Hence such an outcome must belong to compliment of the set A. The event defined by the set A^c = (w |w does not belong to A} is called the compliment of the event A. It is also denoted by A'.

Mutually exclusive (disjoint) events: Two events A and B are called **mutually exclusive (disjoint)** if they have no sample point in common, i.e if $A \cap B = \Phi$. It means mutually exclusive events cannot occur together. Consider a collection of all mutually exclusive events. If the union of all these events gives the entire sample space, then such events are called **exhaustive**. It means that at least one of the events is certain to occur. Two events A and B are said to be **mutually exclusive and exhaustive** if $A \cup B = S$ and $A \cap B = \Phi$. It means A and B constitute a partition of the sample space S. It is easy to see that A and A^c are mutually exclusive and exhaustive events.

Example

From three couples, three persons are chosen. Let A denote the event that two males and one female are chosen, B denote the event that two females and one male are chosen and C denote the event that a couple is chosen. Clearly A and B are mutually exclusive events since $A \cap B = \Phi$. But A and C as well as B and C are not mutually exclusive events.

We list below some identities satisfied by the operations $\cup$ and $\cap$

1. $A \cup B = B \cup A$ (commutative rule)
2. $A \cap B = B \cap A$ (commutative rule)
3. $A \cup \phi = A$ (ϕ is the additive identity)
4. $A \cap \phi = \phi$ (ϕ is the multiplicative identity)
5. $A \cup (B \cup C) = (A \cup B) \cup C$ (associative rule)
6. $A \cap (B \cap C) = (A \cap B) \cap C$ (associative rule)
7. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (distributive rule)
8. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ (distributive rule)
9. $(A^c)^c = A$
10. $(S^c) = \phi$
11. De Morgan's laws: $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$

(These laws can be generalized to 3 or more events).

8.4 Probability

The world around us is full of uncertainties. We can think of measuring the extent of uncertainty. In this section we understand how to measure this uncertainty. Probability theory deals with the study of measurement of uncertainty. The most important aspect of any probabilistic investigation is the assignment of numerical value (called probability) to events. At the initial stage, some simplifying assumptions are made about the structure of the underlying phenomenon. These assumptions together with mathematics and logic help us to develop probability model of the phenomenon. In this course we will study few such models. Before we develop and study these models, we have to understand some basic concepts.

We first should understand meaning of occurrence of an event and the interpretation of probability of an event. In case of tossing of a fair die, sample space S = {1, 2, 3, 4, 5, 6}. Suppose A denotes the event that die shows even number. We will say that event A has occurred if we observe the outcome of a die as 2, 4 or 6. If '2' occurs we say that event A has occurred. If we observe any element other than 2, 4 or 6 we will say that A^c has occurred. Of course event A may or may not occur. There is uncertainty about occurrence of event A. The uncertainty is related to the occurrence of an event when a random experiment is to be performed. If event is certain to occur (with 100% guarantee) then we say that it has probability one. If event is certainly not to occur, we say that its probability is zero. We therefore expect that the probability of an event should have a value in between zero and one (both inclusive). If probability of an event A, denoted by P(A), is 0.80, it is much more likely to occur than if its probability is say 0.15. Events with larger probabilities occur more frequently than events with smaller probabilities.

8.5 The Relative Frequency Approach of Determining Probability

Let f_A be the number of occurrences or frequency of occurrences, of an event A in n repeated trials of the experiment. Then the probability of A is defined as the limit of the ratio f/n, as the number of trials n goes to infinity. The probability of an event can be thought of as a relative frequency of the occurrence of that event in infinite number of repeated trials of the experiment. For example, suppose you collect data on number of male and female children in your area. Suppose out of 1000 children there are 520 male children. So according to the relative frequency approach, probability of male child in your area can be estimated as 520/1000= 0.52

This approach is simple to understand but it has two limitations:

(1) Most experiments cannot be repeated infinite number of times and

(2) We cannot guarantee that the experiment can be repeated under identical conditions.

We therefore assume that P(A) = f/n, when n is sufficiently large. Here, P(A) denotes the probability of the occurrence of the event A. This is an approximation of theoretical probability.

Example 8.1

A market survey was conducted in Nashik city to check whether the advertisement of a new product has changed buying habits of Nashilkites. In a sample survey of 1000 respondents, only 150 informed that they have opted for new product after the advertisement campaign of the company was over in there locality. Here the relative frequency of opting for a new product is 150/1000 = 0.15. Remember that 0.15 is the estimate of the probability that a randomly selected individual will opt for the new product as a result of an advertisement in the locality.

8.6 The Equally Likely Approach

In this approach we assume that every outcome of an experiment has got equal chance of occurrence. Consider a random experiment having n possible outcomes. Suppose that each of these n outcomes is equally likely. Thus, each possible outcome has probability $1/n$ of occurring, when the experiment is performed. Thus equally likely approach defines the probability of the event A as $P(A) = n(A)/n(S)$ where, $n(A)$ is the number of basic outcomes in the event A and $n(S)$ is the number of outcomes in the sample space S.

Examples

1. Consider the experiment of throwing a dice. The sample space is S = {1,2,3,4,5,6}. Let A be the event of occurring an odd number, that is, A = {1,3,5}. We assume all outcomes are equally likely and hence we claim that $P(A) = 3/6$.
2. Consider the experiment of tossing a fair coin once. Suppose A is the event of getting head. Assume equally likely approach of defining the probability. Here $P(A) = ½$, where A is the event of getting a head. Also $P(B) = ½$, where B is the event of getting a tail. Notice that $B = A^c$,
3. Consider the experiment of selecting a card from a well-shuffled pack of cards. Assuming equally likely approach, we can say that each card has probability of 1/52 being selected.

In this approach, we will have to justify the assumption that all outcomes are equally likely. In the games of chance like tossing of a coin, throwing dice, or drawing a card from a well-shuffled deck of 52 cards we can make such assumption. But this is not always the case. For a student appearing for an examination, probability of passing the examination or failing in it need not be equal. Similarly suppose a farmer plants thousands of seeds in his field. In this experiment the probability of germination or no germination of seed cannot be assumed 0.5. (If such is the case then we all will be in a big trouble. Farmers and we all will have serious food problem). There has to be justification for assigning equal probability to each of the possible outcomes.

8.7 Axioms of Probability

Here we assume that S is a sample space associated with a random experiment E and A is an event in S. P(A) is called probability of an even A if the following assumptions are satisfied.

Axiom 1: P(A) is a real number such that $P(A) \geq 0$ for any event A on S.

Axiom 2: P(S) = 1

Axiom 3: For any two mutually exclusive (disjoint) events A and B defined on the sample space S $P(A \cup B) = P(A) + P(B)$

Using the above three axioms of probability, following theorems on probability can be easily proved.

Theorems on Probability (only statements)

(i) $0 \leq P(A) \leq 1$,

(ii) $P(A) + P(A`) = 1$,

(iii) $P(A) \leq P(B)$ when A c B

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

Example

Show that P(A) + P(A^c) = 1

For any event A defined on the sample space S, since A and A^c are disjoint and exhaustive events (that is, $A \bigcap A^c = \varphi$ and $A \cup A^c = S$), $P(A) + P(A^c) = 1$ or $P(A^c) = 1 - P(A)$. This is true because, $P(S) = 1 = P(A \cup A^c) = P(A) + P(A^c)$.

Example

Show that P(φ) = 0 (i.e., probability of an impossible event is zero).

Let A be any event. We know that $A \bigcap \varphi = \varphi$ and $A \cup \varphi = A$. By axiom 3, P($A \cup \varphi$) $= P(A) = P(A) + P(\varphi)$ and since P(A) $\geq$ 0 (by axiom 2) we get P(φ) = 0

Example

Show that if $A \subset B$ then P(A) $\leq$ P(B).

For any two events such that $A \subset B$, we can write B = $A \bigcup (B \bigcap A^c)$. Further A and $B \cap A^c$ are disjoint events, hence P(B) = P(A) + P($B \cap A^c$) and since P($B \cap A^c$)$\geq$ 0, we conclude that P(B) $\geq$ P(A).

Example

Consider a box containing 10 chits numbered 1 to 10. A chit is selected randomly. What is the probability the number on the sampled chit is divisible by 5.

Here sample space S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Each chit has the probability1/10. Define event A = {5, 10}. P(A) = P{5} + P{10} = 1/10 + 1/10 = 2/10. This is the probability that the number on the chit is divisible by 5.

Example 8.2

Suppose a box contains 1 chit of number 1, 2 chits of number 2, and so on up to 10 chits of number 10. The chits are similar but for the numbers written on them. Suppose you draw a chit at random from the box. What is the probability that the number on the selected chit is divisible by 5?

Note that in this case we can consider the whole space as S = {1, 2, 3 …, 10}. Further since number of chits having different numbers is not same for all the numbers, it is not difficult to see that all the 10 outcomes in *S* are not equally likely. However note that there are in all 55 chits (=1 + 2 +…+10) in the box and we can enlarge our whole space to represent these 55 chits. Since we mix all these chits thoroughly before drawing a chit, so all these 55 chits are equally likely to be drawn with probability 1/55 for each chit. Note that for this enlarged sample whole space, a number like 5 does not remain an element of this whole space, but represents a set of all those five chits, which carry the number 5, and so probability of selecting a chit carrying the number *i* is given by $p_i = i/55$. Let *A* denote the event that the selected number is divisible by 5, then A = {5,10}, and so P(A) = p{5} + p{10} = 5/55 + 10/55 = 3/11 Note that the final result is not 2/10, which would be the case if all the numbers 1 to 10 were equally likely.

Using the three axioms of probability several other theoretical results regarding probability can also be proved. We state it below for information.

1. $P(A) = P(A \cap B) + P(A \cap B^c)$

2. $P(A^c \cap B^c) = 1 - P(A \cup B)$

3. $P(A^c \cup B^c) = 1 - P(A \cap B)$

Example 8.3

Suppose one card is drawn at random from an ordinary deck of 52 playing cards. Let A be the event of obtaining "a black ace' and let B be the event "a club". Then P(A) =2/52, P(B) = 13/52 and P(A $\cap$ B)=1/52 and therefore. This is the probability either a black ace or a club card is drawn. $A \cap B$ is the event of drawing an ace of club from a well shuffled pack of 52 cards.

Example 8.4

In a class of 400 students, 150 passed in the subject of English, 100 passed in Mathematics, 50 passed in both mathematics and English, and the rest failed in mathematics and English. Find the probability that a student randomly selected from the class has passed (i) English, (ii) only in English, (iii) Mathematics, (iv) only Mathematics (v) Mathematics and English (vi) neither Mathematics nor English, (vi) either Mathematics or English.

In this example, total number of students = 400. Number of students passed in English = 150, number of students passed in Mathematics = 100, and number of students passed in both Mathematics and English = 50. Therefore, the number of students who passed only in English =100, number of students passed only in Mathematics = 50 and number of students who failed in both Mathematics and English = 200. Suppose E and M denote the events that a student passes in English and Mathematics respectively. Then, P (E) = 150/400, P (M) = 100/400, P (Passing in English only) = 100/400, P(Passing in Mathematics only) = 50/400 and P (not passing in English and Mathematics) = $P(E^c \cap M^c)$ = 200/400.

Verify that P (E $\cup$ M) = 200/400 = (150/400 + 100/400 – 50/400).

You may draw Venn diagram to get the idea of these probabilities.

8.8 Conditional Probability

In the earlier sections we have studied properties of the probability of an event. It is defined under certain axioms. If no restrictions are involved and if we calculate P(A) under these assumptions only, then it is the unconditional probability. However there are cases where determination of P(A) under the condition that some other event say B has already occurred needs to be calculated. Of course P(B) must be positive. We denote this by P(A|B) and read it as the probability of an event A given that B has occurred. Before we define it formally consider the example given below.

Example 8.5

Consider the data on students admitted to IT course.

Gender	Urban area	Rural area	Total
Male	300	200	500
Female	400	100	500
Total	700	300	1000

The relative frequencies are as follows:

Gender	Urban area	Rural area	Total
Male	0.3	0.2	0.5
Female	0.4	0.1	0.5
Total	0.7	0.3	1.0

Suppose we select a student at random from these 1000 students. Let M denote the event that a male student is selected, P (M) is clearly 0.5, similarly if F denote the event that a female student is selected then P (F) is also 0.5. Suppose you know that a student from urban area is selected, then what is the probability that a student is male? It will be 300/700. How we get it? This is the probability that a student is male given that he is from urban area. Since there are 700 urban students, of which 300 are males, we find

$P(M|U) = P(M \cap U) / P(U) = [300/1000]/[700/1000] = 300/700.$

Similarly $P(F|U) = 400/700$.

Definition Let B be an event with positive probability. The conditional probability P(A|B) of an event A given the event B is $P(A|B) = P(A \cap B) / P(B)$.

Remember that P(A|B) is the probability of an event A given that the other event B has occurred. It therefore satisfies all the properties of probability of an event A

To calculate P(A|B), we reduce the sample space S to the collection of outcomes that belong to event B. The conditional probability P(A|B) measures the fraction of those occurrences of B that also resulted in the occurrence of an event A.

Example 8.6

In the above example to find the probability of male student selected

$P(Male) = P(Male \cap Urban) + P(Male \cap Rural) = 0.3 + 0.2 = 0.5.$

Note that we also get $P(M \cap U) = P(U) \cdot P(M \mid U)$.

Similarly we can also get $P(M \cap U) = P(M)\, P(U|M)$

This is the unconditional probability. (We have used $P(A) = P(A \cap B) + P(A \cap B^c)$) where A represents the event 'male' and B represents the event of urban (B^c, the event of rural).

Example 8.7

There are 40 white balls and 50 black balls in an urn. 20% of white balls and 10% of black balls are marked. A person tosses a fair coin. If coin shows head, he selects a ball from white balls; otherwise he selects a ball from black balls. What is the probability that the selected ball is marked?

Define the events A and B as follows:

A: The ball is selected from white balls.

B: The selected ball is marked.

The person selects either a white ball or a black ball with probability 0.5 (tossing a fair coin). Since 8 white and 5 black balls are marked, $P(B|A) = 0.20$ and $P(B|A^c) = 0.10$. We want to obtain P(B). Notice that $B = (B \cap A) \cup (B \cap A^c)$.

Therefore $P(B) = P((B \cap A) \cup (B \cap A^c)) = P(B \cap A) + P(B \cap A^c)$

$= P(A)^* P(B \cap A) + P(A^c)^* P(B \cap A^c)$

$= 0.5^*0.2 + 0.5 * 0.1$

$= 0.15.$

In the above example we have made use of multiplicative law. We state it below.

8.9 Multiplicative Law

Suppose we have two events A and B defined on the same sample space. Then the conditional probability of A given B defined as $P(A|B) = P(A \cap B) / P(B)$, $P(B) > 0$, implies that that $P(A \cap B) = P(A|B)\, P(B)$ or $P(A \cap B) = P(B|A)\, P(A)$. This is the *multiplicative law* of probability.

Example 8.8

In a certain lot 0.5% of the units are defectives. Each unit is subjected to a test. The test correctly identifies a defective unit. The test also indicates that the unit is defective about 5 in every 100 non-defective units. A unit is randomly chosen from the lot fails in the test. What is the probability that the unit chosen is actually not defective?

Suppose A_1 denotes the event that the randomly chosen unit is not defective and A_2 denotes the event that it is defective. We are given $P(A_1) = 1 - .005 = .995$ and $P(A_2) = 0.005$; Let B denote the event that the randomly selected unit fails in the test. We want to obtain $P(A_1|B)$. By definition $P(A_1|B) = P(A_1 \cap B)/P(B)$.

$P(A_1 \cap B) = P(A_1)P(B|A_1) = (0.995)*(0.05) = 0.04975$ and

$P(B) = P(B \cap A_1) + P(B \cap A_2) = P(A_1)* P(B|A_1) + P(A_2)* P(B|A_2) = 0.05475$

(Verify that $P(B \cap A_2) = 0.005$); hence $P(A_1|B) = 0.9087$. It means almost 91% of the times unit declared defective will not be so.

Comment: In solving above example we have made use of partition of the sample space and Baye's theorem. We state it below for interested students.

Partition of sample space S

Suppose $A_1, A_2, \ldots A_k$ are k events defined on the sample space S. The events A_i, $i = 1,2,\ldots,n$ are said to constitute a *partition* of S if (i) $A_i \cap A_j = \phi$ for all $i \neq j$ and (ii) $A_1 \cup A_2 \cup \ldots \cup A_n = S$. Further let B be any event. We can write $B = B \cap S = B \cap (\bigcup_{i=1}^{n} A_i)$.

Thus since $B \bigcup A_i$ are disjoint sets we have $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$

8.10 Baye's Theorem

Let $A_1, A_2, \ldots, A_n$ represent a partition of the sample space S and let B be an event defined on S such that $P(B) \neq 0$ Then the conditional probability if A_i given B is

$P(A_i|B) = P(B|A_i)* P(A_i) / P(B)$, where $P(B) = \sum_{i=1}^{k} P(B \mid A_i)P(A_i)$.

8.11 Concept of Independence

Suppose A and B are any two events associated with the same sample space. Assume that B and B^c are events with positive probabilities. Let $\alpha = P(A \mid B)$ and. $\beta = P(A \mid B^c)$ If $\alpha > \beta$ we say that event A is more likely to happen with B as opposed to B^c. Also if $\alpha < \beta$ than event A is less likely to happen if B does than B does not. However if $\alpha = \beta$, then A is as likely to happen if B does or B does not happen. It means occurrence or non-occurrence of B does not in any way influence occurrence of A. This agrees with intuitive notion of causal independence of events A and B. Let us find α. Since $\alpha = P(A \mid B)$, we have $\alpha * P(B) = P(A \cap B)$.

Also $\beta = P(A \mid B^c)$ implies $\beta * P(B^c) = P(A \cap B^c)$.

Now since $\alpha = \alpha * (1) = \alpha * \{P(B) + P(B^c)\} = P(A \cap B) + P(A \cap B^c) = P(A)$. This gives us the definition of independence of two events of the sample space.

Definition: Let A and B be any two events of sample space S associated with random experiment E. They are said to be independent if $P(A \cap B) = P(A)P(B)$

Comment: Concept of independence can be generalized to 3 or more events. If A, B and C are any three events such that

- $P(A \cap B) = P(A)P(B)$
- $P(B \cap C) = P(B)P(C)$
- $P(C \cap A) = P(C)P(A)$ and
- $P(A \cap B \cap C) = P(A)P(B)P(C)$

Then we say that A, B and C are completely independent (or mutually independent).

Remember it may happen that A, B, and C are pair wise independent, but they are not completely independent. It is also possible that the condition $P(A \cap B \cap C) = P(A)P(B)P(C)$ holds good, but any two of these three events are not independent.

Example 8.9

The study of data of employee's turnover in an IT Company shows that 50% of its high skilled employees resign in less than 2 years. Ashish and Devendra are selected for the high skilled job. What is the probability that (a) both resign in less than 2 years, (b) exactly one resigns? Assume that the decision of Ashish and Devendra are independently taken.

Let A be the event that Ashish resigns and D be the event that Devendra resigns.

We are given $P(A) = P(D) = 0.5$. The probability that both resign within a span of 2 years is equal to $P(A \cap D) = P(A)*P(D) = 0.25$. To find the probability that exactly one quits but other does not, we have to obtain $P(A \cap D^c) + P(A^c \cap D)$. Since A and D are independent we can claim that events A and D^c (and also A^c and D) are independent. Using the definition of independence of two events, we get required probability as 0.50.

Example 8.10

Show that if A and B are independent events then (a) A and B^c are independent events, (b) A^c and B^c are also independent events.

Since $P(A \cap B^c) = P(A)*P(B^c \mid A) = P(A)\{1- P(B|A)\} = P(A)(1 – P(B)) = P(A)*P(B^c)$, A and B^c are independent events. Using De Morgan's laws, we can write

$P(A^c \cap B^c)=1- P(A \cup B) = 1- \{P(A) + P(B) - P(A)*P(B)\}= \{1-P(A)\}*\{1-P(B)\} = P(A^c)*P(B^c)$

Hence the result.

8.12 Note on Counting Techniques

Many of the basic concepts of probability theory can be considered in the context of finite sample space. It requires that students are aware about mathematical technique of counting. The knowledge of these counting techniques is a pre-requisite for study of theory of probability. Here we briefly summarize these techniques.

Multiplication Principle: *If sets A and B have m and n elements respectively, then there are m x n ways in which one can select an element of A and then an element of B. (This result can be generalized).*

Example 8.11

Suppose that a person has choice of 6 shirts and 5 trousers. Then he has 6 x 5 = 30 different choices of wearing a dress.

Example 8.12

A class consists of 40 boys and 25 girls. Two students – one each from boys and girls are to be selected. This can be done in 40 x 25 = 1000 different ways. Remember that the order in which you select does not matter.

Addition Principle: If disjoint sets A and B have m and n elements respectively, then there are m + n ways in which one can select either an element of A or an element of B. (This result can be generalized).

Example 8.13

Suppose there are 5 books available for Statistics and 10 books are available for Mathematics. Total number of ways a student can borrow a book is 5 + 10 = 15

Combination and Permutation

Many of the problems in probability are based on the assumption of equiprobable sample space. So it is necessary that we know some basics of permutation and combination. Consider set A = {a, b, c} and set B = {b, c, a}. You know that sets A and B are the same sets. The order of the elements within a set is not important. This is called an unordered set. Such an unordered set is also referred as a combination. If order within a set is important, then we refer such ordered set as permutation. In permutation, you have the same elements in the sets but the order in which these elements appear does matter. A = {1, 2, 3, 4} and B = {4, 3, 2, 1} are two different ordered sets, although the same elements belong to each of A and B.

Combination of r objects out of a set of n objects; (^nC_r)

Suppose a set contains n distinct objects. Number of possible combinations of r objects (r ≤ n) out of a set with n objects is denoted by ^nC_r and is given by $^nC_r = \frac{n!}{r!(n-r)!}$, where, n! = 1 x 2 x 3…x *n*, r! = 1 x 2 x 3…x *r* and (n-k)! = 1 x 2 x 3…x (n-k), r and n are positive integers. Remember that 0! = 1 by definition.

Example 8.14

Suppose you are given 12 points on a plane no three of which are on the same line. How many straight lines can be drawn by joining the points? How many triangles?

Select any two points out of 12 points and then line may be drawn. There will be thus in all $^{12}C_2 = \frac{12!}{2!10!} = 66$ different lines. Similarly $^{12}C_3 = \frac{12!}{3!9!} = 220$ triangles can be drawn.

Permutation of r objects out of a set of n objects (^nP_r)

Number of possible permutations of r objects out of a set with n objects is denoted by ^nP_r and is given by $^nP_r = \frac{n!}{(n-r)!}$, where $0 \le r \le n$ and r and n are positive integers.

Example 8.15

A committee consisting of 4 members is to be formed from a group of 12 members. How many possible committees can be formed? Suppose the first person selected will be President, 2nd person – Vice President, 3rd secretary and 4th will be a treasurer. In how many possible ways this can be done? A committee of 4 members from 12 members can be selected in $^{12}C_4 = \frac{12!}{4!8!} = 495$ different ways (if order of selection is not important).

In case of an ordered set of four members out of 12 a committee can be formed in ${}^{12}P_4$ = 12x11x10x9 = 11880 different ways.

In permutation order is taken into account and in combination it is not taken into account. So when in doubt always ask the question – Is order important? If it is then it is a problem on permutation, otherwise on combination.

Example 8.16

Five chairs are numbered 1 to 10. Two men and three women wish to occupy one chair each. First the women choose the chairs from amongst the chairs 1 to 5 and then the men select from the remaining chairs. Find the number of possible arrangements.

Three women can be arranged in the five chairs in ${}^5P_3$ ways. Then, two men can be arranged in the remaining 7 chairs in ${}^7P_2$ ways. Hence the total number of possible different arrangements is ${}^5P_3 * {}^7P_2$ = 1,51,200.

8.13 Self Test

1. You are given a set of experiments. List the outcomes and describe the sample space. Also prepare a list of the sample points that belong to particular events mentioned. Also classify the experiment as non-deterministic (random) or deterministic. Give reasons for the same.

a. Tossing of two coins together and the event of getting (i) at least one head (ii) at most one head (iii) exactly one head.

b. Tossing of three coins together and the event of getting (i) at least one head (ii) at most one head (iii) exactly one head.

c. Tossing of two dice together and the event of getting (i) sum on the uppermost faces less than 10 (ii) sum even (iii) sum a prime number.

d. A card is drawn from a pack 52 playing cards and its (i) suit (ii) color is noted.

e. A rigid body is allowed to fall freely from a certain height.

f. Water in a pot is heated over 100^0 C for a sufficiently long period of time and its end result is noted.

g. A electric spark is introduced in a cylinder containing mixture of hydrogen and oxygen.

h. The number of students attending a period from a batch of 50 is noted.

i. Astrologers classify a student as mentally retarded or bright after studying the horoscope. The experiment is conducted on 50 horoscopes.

2. During an opinion poll, 100 sampled voters were asked the question, "Will you vote in favor of ruling party or against?"

(i) Describe the experiment. Is it a random experiment? List possible elements of the sample space.

(ii) Describe any one event.

(iii) It is conjectured that 35% of voters may vote for ruling party, 40% may vote against the ruling party, while others have not made any decision yet and will therefore not opine. Suppose to voters are selected randomly. What is the probability that (i) both will favour ruling party candidate (ii) both will oppose, (i) exactly one of them will favour the ruling party candidate.

3. The number of spots turning up when a six-sided die is tossed is observed. Consider the following events.

A: The number observed is at most 3

B: The number observed is an odd number

C: The number 2 turns up.

a. Define the sample space for this random experiment and assign probabilities to the simple events.

b. Find $P(A)$ and $P(A^c)$

c. Are events *A* and *B* mutually exclusive?

d. Find $P(A \cup B)$ and $P(A \cap B)$

e. Find $P(A \mid B)$ and $P(A \mid B^c)$

4. Let A, and B are two events defined on a sample space. Find an expression in terms of union and/or intersections of A and B for the following events:

a. At least one of the events A or B occurs

b. A and B^c occur simultaneously

c. A occurs with either B or B^c

5. There are sixty employees working in a mall. Their details are as follows:

Sex↓ \ Edu.→	*Undergraduate*	*Postgraduate*	*Total*
Female	*45*	*15*	*60*
Male	*105*	*35*	*140*
Total	*150*	*50*	*200*

An employee is selected at random.

a. What is the probability the employee is male?

b. What is the probability the employee is either male or postgraduate?

c. What is the probability the employee is a postgraduate given that a male employee is selected?

6. Five candidates Patil, Jadhav, Deshmukh, Apte and Sonar are standing for election. Their names are drawn randomly.

a. What is the probability that Mr. Apte is drawn first?

b. What is the probability that the order is alphabetical?

7. Draw one card at random from a standard deck of cards. Consider the following events.

A = {x : x is a heart}, B = {x: x is an ace or 2 and x is red}, C = {x: x is red}

Find the following probabilities.

(i) $P(A)$ **(ii)** $P(B)$ **(iii)** $P(C)$ **(iv)** $P(A \cap B)$ **(vi)** $P(A \cup B)$

8. A doctor claims that a patient surviving a particular type of open-heart surgery is 0.85. If the doctor performs ten such operations, find the probability of the following.

(i) All ten patients survive; **(ii)** At most five patients survive

(iii) At least eight patients survive

9. A student plants ten seeds each of the two crops A and B in a pot culture trial. If it is known that the probability that seed of A will germinate is 0.8; the probability that seed of B will germinate is 0.2 then find

(i) probability that all the seeds of A and B will germinate,

(ii) probability that exactly all the seeds of one of the crop A or B will germinate.

10. A test in English is given to a class of 10 students. 3 of them scored A, 2 scored B and others scored C. A committee of 3 students is to be formed. Find the probability that the committee will be of

(a) only those who scored A.

(b) one each from A's, B's and C's.

11. Two cards are drawn from a well-shuffled deck of 52 cards. Consider the following events:

A: Both cards are queens, B: Both cards are red

C: One card is red and one is black D: A queen and a king is drawn.

Find P(A), P(B), P(C), P(D), P(A|B), $P(A|B^c)$

Unit 9 : Random Variables

9.1 Overview

In this unit you will be introduced to discrete and continuous random variables. In it you will study properties of probability mass function, probability density function, and cumulative distribution function. You will also learn how to compute mean and variance of the random variable being studied using the concept of expectation. We will discuss discrete as well as continuous random variables.

9.2 Objectives

After studying this chapter, you will be able to

- Identify discrete as well as continuous random variables
- Define probability mass function and cumulative distribution function of a discrete random variable
- Obtain probabilities of various events
- Define expected value of a random variable or its function
- Identify situations wherein use of standard discrete and continuous distributions would be appropriate
- Solve simple numerical problems.

9.3 Random Variables

Consider a random experiment of tossing three coins together. The sample space of this experiment is S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Suppose that the experimenter is interested in knowing total number of heads that turn up when the experiment is realized. It can take values 0, 1, 2 or 3. Suppose X denotes the total number of heads. If the experiment is performed repeatedly, then the value of X will vary from each trial. We therefore call X as a variable. We always use capital letters to denote the random variables (with or without suffix) and the values taken by the variable is usually denoted by corresponding small letters. Thus here X takes the values 0, 1, 2, or 3. This is the sample space of the random variable. In mathematical jargon we say that X is a function defined from S to R and write it as X: S --> R. Notice that

X(HHH) = 3	X(HHT) = 2	X(HTH) = 2	X(THH) = 2
X(HTT) = 1	X(THT) = 1	X(TTH) = 1	X(TTT) = 0

The random experiment had eight sample points. The sample space contains all these points. The random variable X in this case has taken only four possible values. It is possible that you may define another variable on the same sample space. For example, define Y as 1 if you observe at least one head and at least one tail, otherwise define Y as 0. Notice that

Y(HHH) = 0	Y(HHT) = 1	Y(HTH) = 1	Y(THH) = 1
Y(HTT) = 1	Y(THT) = 1	Y(TTH) = 1	Y(TTT) = 0

Here variable X (and also Y) takes isolated values. Total number of these values is countable. We will see later in this chapter that these values are taken with different probabilities. Therefore we call these variables as discrete random variables.

9.4 Discrete Random Variable

A random variable is discrete if it can assume either a finite or countably infinite number values.

Countable means that you can associate the values that the random variable can assume with the integers 1, 2, 3 and so on. For example, if you count how many times a particular attribute occurs, its possible values are usually integers such as 0, 1, 2 …; clearly the number of outcomes will either be finite or countable. It however does not mean that a discrete random variable must take integer values. It takes isolated values and these values are countable. If you find proportion of female children in a family having four children, then the possible values it can assume are 0, 0.25, 0.50, 0.75 and 1.

Typical examples of a discrete random variable are:

- Number of peas set in a pod.
- Number of accidents on Mumbai- Agra national highway in a day
- Number of misprints per page
- Number of male children in a family having n children
- Number of TV channels seen by the viewer in a day.
- Number of candidates interviewed before the first candidate is selected.
- Number of customers in a queue waiting for service during peak hours.
- Number of errors reported while compiling a C-program.
- Number of defective bolts in a sample of size 20 drawn from day's production.
- Number of goals scored in a foot ball match.

9.5 Continuous Random Variable

A random variable is continuous if it can assume all values in an interval (finite or infinite). Thus a continuous random variable can assume the infinitely large number of values corresponding to all the points on a line interval. .

Examples of a continuous random variable

- The height of an individual.
- The length of life of a component.
- The systolic and diastolic blood pressure of an adult individual.
- The residual life of an individual.
- The time required for roasting groundnuts.
- The amount of sugar in an apple.
- The waiting time for receiving service at a nationalized bank.

Note that all random variables are strictly speaking discrete. This is so because it can take values which are multiples of the least count of the measuring device. However this is only the

practical limitations of measurement. The continuous random variables them self can take any value in the interval. (In each of the cases listed above, the random variable under consideration can have any intermediate value in the possible range of the variable. Whether you can record or measure the particular value depends on several factors such as the measuring instrument. Nonetheless the variable must be termed as continuous random variable). There can be a variable, which is a mixture of discrete and continuous random variables. However we will not study such variables in this course. But remember that there can be a variable which is neither discrete or nor continuous.

9.6 Probability Distribution

9.6.1 Discrete Random Variable

A probability distribution is the distribution of unit probability mass among all possible outcomes of a random experiment. We have earlier seen that a relative frequency distribution distributes a total of one among different values of a random variable (response variable).

In case of a discrete random variable, we can report its probability distribution (also called as probability mass function) in a table form. It is also possible to report it in a graph or by a rule that associates a probability P(X = x) = P(x) with each possible value of x. (There is difference between probability mass function and probability distribution of a random variable; however we will not elaborate on it here). Before we study properties of pmf (probability mass function) of a discrete random variable, we will first define it formally.

9.6.2 Probability Mass Function

The function f(x) = P[X = x] defined for all values of x assumed by X satisfying the following conditions is called the probability mass function of X.

$$f(x) \geq 0 \text{ for all x and } \sum f(x) = 1$$

Any function that satisfies above two conditions qualifies to be a probability mass function of some random variable. You can draw a probability histogram to picture probability mass function. On horizontal scale show the possible values of X and the height of each bar is the probability for the value reported at its base. It is equivalent to relative frequency histogram.

Example 9.1: Recall the example of tossing of three fair coins. Let X denotes total number of heads. The possible values of X are 0, 1, 2 and 3. It is easy to report probability mass function of random variable X in the tabular form as follows:

x	0	1	2	3
P(x)	1/8 = 0.125	3/8 = 0.375	3/8 = 0.375	1/8 = 0.125

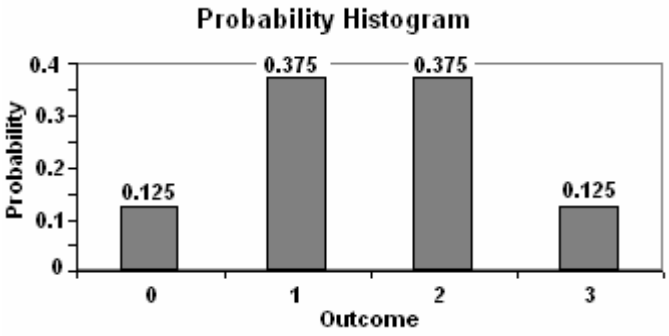


Figure 9.1: probability mass function Vs x-values

Example 9.2: Suppose you toss two fair dice with faces marked 1, 2,…6. and observe the sum on uppermost faces (say X). Verify that following is the probability mass function of the sum on uppermost faces:

x	2	3	4	5	6	7	8	9	10	11	12
P(x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Example 9.3: Verify whether f(x) = [4 - |x – 5|] / 16 for x = 2, 3, …8 is a probability mass function of some random variable X.? If it is, obtain P(X < 4), P(X ≤ 4) and P(X ≥6)

Notice that f(x) > 0 for x = 2, 3,…, 8 and f(x) can be written in tabular form as follows:

x	2	3	4	5	6	7	8	Total
f(x)	1/16	2/16	3/16	4/16	3/16	2/16	1/16	1

And since ∑f(x) = 1 we can claim that f(x) is a probability mass function of some random variable X.

It is customary to use notation p(x) to denote probability mass function of random variable X.

P(X < 4) = P(X = 2) + P(X = 3) = 3/16

P(X ≤ 4) = P(X = 2) + P(X =3) + P(X = 4) = 6/16

P(X ≥6) = P(X = 6) + P(X = 7) +P(X = 8) = 6/16

9.6.3 CDF (Cumulative Distribution Function)

We here introduce a fundamental concept of CDF. Consider the Example 9.1 above and denote

F(0) = P(X≤ 0) = 1/8

F(1) = P(X≤ 1) = 3/8 (= P(0) + P(1))

F(2) = P(X≤ 2) = 7/8 (= P(0) + P(1) +P(2))

F(3) = P(X≤ 3) = 8/8 (= P(0) + P(1) + P(2))

In general we can define F(x) = P(X ≤ x), where x is any real number.

This function F(x) is called CDF (Cumulative distribution function) of random variable X. It can be represented by a graph as in following Fig 9 .2. It is easy to see that we can obtain the probability mass function by differencing F(x). For example, p(2) = F(2) - F(1), p(3) = F(3) - F(2), and so on. In general p(x) = F(x) – F(x-1).

.

Example 9.4: Following is the CDF of a discrete r.v. X

x	F(x)
1	0.1
2	0.3
3	0.5
4	0.7
5	0.8
6 or more	1

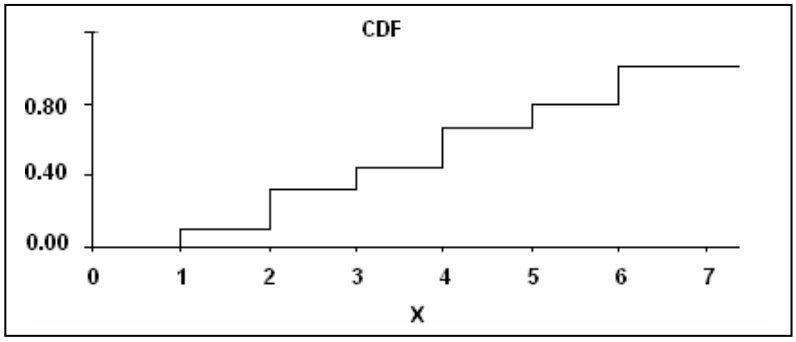


Figure 9.2: CDF as a step function

Properties of CDF (discrete random variable)

1. F(x) is defined for every real number x.
2. $0 \leq F(x) \leq 1$, since F(x) is a probability.
3. F(x) is a non-decreasing step function of x. (looks like staircase).
4. $F(-\infty) = 0$ and $F(\infty) = 1$
5. If a and b are any two real numbers such that $a \leq b$, then

   (i) $P(a < X \leq b) = F(b) - F(a)$

   (ii) $P(a \leq X \leq b) = F(b) - F(a) + P(X = a)$

   (iii) $P(a \leq X < b) = F(b) - F(a) - P(X = b) + P(X = a)$

   (iv) $P(X > a) = 1 - P(X \leq a) = 1 - F(a)$

6. CDF of a discrete random variable is usually a step function. This can be seen by plotting a graph of F(x) Vs x. It is a theoretical counterpart of a "less than" cumulative frequency curve.

Mean (μ) and Variance (σ^2)

Suppose X is a discrete r. v. and p(x) is its probability mass function. Then expected value or mean of X is defined as the weighted average of its possible values, weights being respective probabilities of the values of X. It is denoted by μ or E(X). Thus

$$\mathbf{E(X) = \text{mean of } X = \sum xP(X = x)}$$

If X takes finite values then existence of E(X) is guaranteed. If X takes countably infinite values then its existence needs to be established. (We will assume existence of mean, in the sense that the infinite series converges.).

Similarly variance of X is defined as $E(X - \mu)^2 = \sum (X - \mu)^2 \, xP(X = x)$ and is denoted by V(X) or by σ_x^2.

The square root of variance is called as standard deviation and is denoted by σ.

Median and Mode

The median of a probability distribution is that value of X for which $P(X \leq x) \geq 0.5$ and $P(X \geq x) \geq 0.5$.

Mode of a probability distribution is defined as that value of X at which probability mass function is maximum. The distribution of X can have two or more modes.

Example 9.5: The CDF of a r.v. X is given below. Using it obtain (i) pmf of X, (ii) $P(X \leq 2)$, (iii) $P(X \leq 4)$ and (iv) $P(X > 4)$, (v) E(X) and V(X), (vi) x such that $P(X \leq x) = 0.5$

x	1	2	3	4	5	6
F(x)	0.08	0.26	0.50	0.68	0.93	1.00

(i) We will report pmf (probability mass function) of X in its tabular form.

x	1	2	3	4	5	6
p(x)	0.08	0.18	0.24	0.18	025	0.07

(ii) $P(X \leq 2) = F(2) = 0.26$,

(iii) $P(X \leq 4) = F(4) = 0.68$

(iv) $P(X > 4)$, $= p(5) + p(6) = 0.32 (= 1 - F(4))$

(v) $E(X) = 1*(0.08) + 2*(0.18) + 3*(0.24) + 4*(0.18) + 5*(0.25) + 6*(0.07) = 3.55$

$V(X) = (1-3.55)^2*(0.08) + (2-3.55)^2*(0.18) + \ldots + (6-3.55)^2*(0.07) = 2.0075$

(vi) $F(3) = 0.5$; this is the value of x, below and above which probability is 0.5. This is the median of the distribution.

9.6.4 Continuous Random Variables

Suppose you want to select a number between 0 and 1 randomly. You can make use of random number tables. The probability that a digit between 0 and 9 is selected would be 1/10 to each of the 10 possible outcomes. Think of a situation wherein you want to select any number between 0 and 1. Software such as EXCEL, MINITAB, SYSTAT, etc. can do it for you. Think of a spinner wheel that turns freely on its axis and comes to rest. The sample space of this random experiment is the interval of numbers. We can not assign probability to each individual value of r. v. X and can not calculate probability of an event by our earlier approach. The probability that the continuous r. v. X takes a particular value is always zero. This is something similar that you have studied in your school geometry. In the school geometry, you assume that every point on a real line has no length. On a line segment there are infinitely many points. We define and obtain length of a line joining the two points.

Can we assign probabilities to events such as $\{X < 0\}$, $\{0.1 \leq X \leq 0.5\}$ or $\{X \geq 0.90\}$? Answer is yes. We use a new approach of assigning probabilities directly to events as "areas under a density curve". The density curve of any continuous r. v. has area exactly unity below it and it corresponds to total probability. **Probability as area under a density curve is an important way of assigning probabilities to events of a continuous random variable.**

Definition: A continuous random variable X takes all values in an interval of numbers. The probability density function (or probability distribution) of a continuous r.v. X is described by a density curve f(x). Any function f(x) satisfying $f(x) \geq 0$ for all values of x and $\int_R f(x)dx = 1$, where R is the range of X, is called probability density function (probability density function) of X.

Definition: The cumulative distribution function (CDF) of a random variable X is defined for any real x by F(x) = P[$X \leq x$]. The function F(x) is often referred as the distribution function of X. (Some authors also use the notation $F_X(x)$).

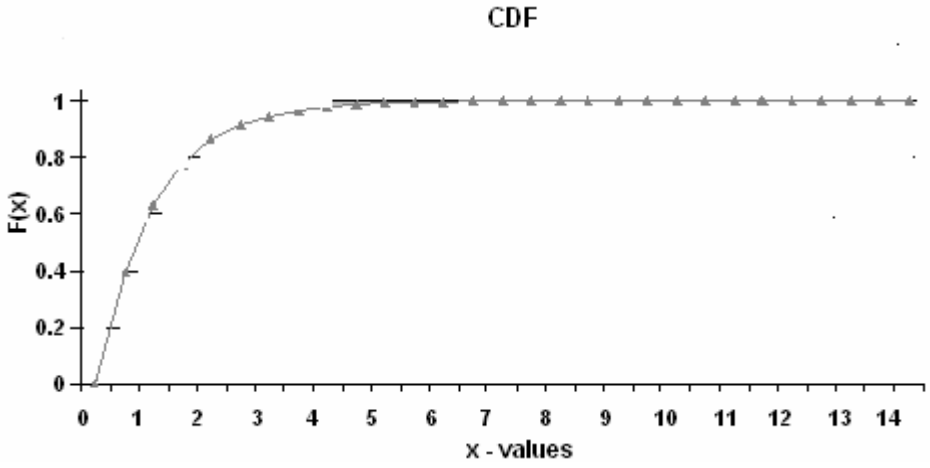
Points to remember:

- All continuous probability distributions, always assign probability 0 to every individual outcome.

- Only intervals of values may have a non-negative probability.
- Since probability at a point is always zero, whether lower and upper limit are included or excluded in the event defined makes no difference for the probability that the r.v. X takes a value in that interval. Thus we have $P(a \le X \le b) = P(a \le X < b) = P(a < X < b)$ and so on.
- $P(a < X < b)$ = Area upto b – area upto a = $\int_{x=a}^{b} f(x)dx$ = F(b) – F(a)
- If $f(x) \ge 0$ for all values of X and $\int_{R} f(x)dx = 1$, then f(x) is a well-defined function of some random variable X. This is useful property and can be used to verify that the given function is probability density function or nor and also to obtain unknown constants, if any in the expression for probability density function.
- Notice that we have stated $f(x) \ge 0$. It must be non-negative, however It can take any non-negative value and so it can certainly exceed 1. This is so because by itself f(x) does not define a probability.

9.6.5 Properties of CDF (Cumulative distribution function) of Continuous r. v. X

1. $F(x) = P(X \le x)$ is defined for every real number x.
2. $0 \le F(x) \le 1$, i.e. F(x) is bounded below and above. .
3. F(x) is a non-decreasing function of x, i.e. $F(x_1) \le F(x_2)$, whenever $x_1 \le x_2$.
4. $F(-\infty) = 0$ and $F(\infty) = 1$ (These are limit values)
5. If a and b are any two real numbers such that $a \le b$, then
   a. $P(a < X \le b) = F(b) - F(a)$
   b. $P(a \le X \le b) = F(b) - F(a)$
   c. $P(a \le X < b) = F(b) - F(a)$
   d. $P(X > a) = 1 - P(X \le a) = 1 - F(a)$
6. CDF of a continuous random variable is a continuous function e.g., study the graph of cdf Vs x shown in the following figure.



Mean and variance

If X is a continuous random variable with pdf f(x), then the mean or expected value of X (also denoted by μ) is defined by $E(X) = \int_R xf(x)dx$, provided it exists. (Remember that existence of E(X), i.e. convergence of integration needs to be established. In this course we will deal with only those variables for which mean and variance exists.).

If X is a continuous random variable with pdf f(x), then the variance of X (also denoted by σ^2) is defined by

$V(X) = \int_R (x-\mu)^2 f(x)dx$ **, where** $\mu = E(X)$

For any random variable X, it can also be shown that $V(X) = E(X^2) - [E(X)]^2$

Example 9.6: Find the value of unknown constant k so that the function f defined below can be considered as probability density function of some r.v. X.

$f(x) = k$ for $a \le x \le b$, $k > 0$
$= 0$ otherwise

Suppose a = 0 and b = 1. State probability density function and evaluate probabilities of events

(a) $\{X \le 0.5I\}$, (b) $\{X \le 0.8\}$, (c)$\{0.5 \le X \le 0.8\}$, and

(d) $\{X > 0.8\}$. Also obtain mean and variance of X.

Since k is positive f(x) is always non-negative. We need $\int_a^b f(x)dx = 1,$ It implies

$\int_a^b kdx = k[x]_a^b = k(b-a) = 1$ implies k = 1 / (b – a) for a < X < b. Thus probability density function of X can be stated as f(x) = 1 / (b – a) for $a \le x \le b$,

= 0 otherwise.

By definition of CDF, F(x) = 0 for all x < a,

$= \int_a^x f(x)dx = \int_a^x \frac{1}{b-a}dx = \frac{x-a}{b-a}$, for $a \le x \le b$ and

= 1 for x > b

Since a = 0 and b = 1, we have f(x) = 1 for 0 <x < 1 and F(x) = x for $0 \le x \le 1$

P{X < 0.5} = F(0.5) = 0.5; similarly P[X < 0.8] = F(0.8) = 0.8 and therefore

P[0.5 < X < 0.8] = F(0.8) - F(0.5) = 0.3 and P[X > 0.8] = 1 – P[X < 0.8] = 0.2

$E(X) = \int_R xf(x)dx = \int_0^1 xdx = \frac{1}{2}$, and $V(X) = \int_R (x-\mu)^2 f(x)dx = \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \frac{1}{12}$

Thus the random variable X has mean ½ and variance 1/12.

Example 9.7: The probability density function of a continuous r. v X is as given below

$f(x) = 3x^2$ for $0 \le x \le 1$
$= 0$ otherwise

Verify that f(x) is a well-defined probability density function. Find its mean and variance. Sketch the probability density function and cdf of X. Also find P(0.75 < X < o.90)

Notice that f(x) is always nonnegative for all values of x and $\int_a^b f(x)dx = \int_0^1 3x^2 dx = 3[\frac{x^3}{3}]_0^1 = 1,$

so the function f(x) is a probability density function.

E(X) = $\int_a^b xf(x)dx = \int_0^1 3x^3 dx = 3*\left[\frac{x^4}{4}\right]_0^1 = \frac{3}{4},$ and E(X^2) = $\int_a^b x^2 f(x)dx = \int_0^1 3x^4 dx = 3*\left[\frac{x^5}{5}\right]_0^1 = \frac{3}{5},$

and hence V(X) = (3/5) – $(3/4)^2$ = (48 – 45)/ 80 = 3/80

F(x) = P[X ≤ x] = $\int_a^x f(x)dx = \int_0^x 3x^2 dx = 3\left[\frac{x^3}{3}\right]_0^x = x^3$, for 0 ≤ x ≤ 1

= 1 for x > 1

Figure 9.3 shows the graph of pdf and cdf of r. v X.

Figure 9.3: pdf and cdf of r. v X

9.7 Some Special Continuous Probability Distributions

In this part we will study some important continuous distributions. First we will study (i) Uniform (rectangular) distribution (ii) Exponential distribution and (iii) Normal distribution. We will study situations where these distributions can be used as models to describe the random variable under consideration.

9.7.1 Uniform Distribution

The uniform distribution provides a model for selecting a point "at random" from an interval (a, b) on a real line. Suppose that a student has to catch a bus to reach his place of education. Suppose the public transport system is such that the bus promptly arrives at a bus stop at fifteen minutes interval. The student however arrives at a bus stop at a random time between bus arrivals. The waiting time, say X, can be modeled by continuous uniform distribution.

Definition: A continuous random variable X is said to have Uniform distribution on the interval (a, b) if its pdf is of the form

f(x) = 1/(b – a) for a < x < b

= 0; otherwise

We use the notation X ~U(a, b).

You can write computer programs to generate random numbers. The random number generators are functions in the computer language and are based on programs that use uniform distribution defined on the interval (0, 1).

CDF

Since F(x) is defined for all real numbers x,

F(x) = 0 for x < a.

$F(x) = \int_a^x f(x)dx = \int_a^x \frac{1}{b-a}dx = \frac{x-a}{b-a}$, for a < x < b,

F(x) = 1, for x > b,

The graphs of pdf and CDF are given in figure 9.4.

Figure 9.4: graphs of pdf and CDF

Mean and Variance

$$\text{Mean} = E(X) = \int_R xf(x)dx = \int_a^b xf(x)dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b+a}{2},$$

$$\text{Similarly } E(X^2) = \int_R x^2 f(x)dx = \int_a^b x^2 f(x)dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{b^2+ab+a^2}{3},$$

$$\text{Hence } V(X) = E(X^2) - \{E(X)\}^2 = = \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12}$$

9.7.2 Normal Distribution

The normal distribution is the single most important distribution in probability and statistics. It plays a crucial role in statistics. This distribution is extensively used for modeling several random phenomena in almost all branches of science, social science, biology and medical science etc. Under certain assumptions this distribution can be used as an approximation to several other distributions. It is also known as the Gaussian distribution. The distribution is mathematically tractable and has beautiful perfect bell shape and symmetric nature. We will study normal distribution in detail.

Definition: A random variable X follows the normal distribution with mean μ and variance σ^2 if it has the pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$, for $-\infty < x < \infty$, $-\infty < \mu < \infty$ and $\sigma > 0$

Notation X ~N(μ, σ^2). If mean = 0 and variance = 1, then the random variable is called as SNV (standard normal variate). It is customary to use Z as a symbol for SNV. It is easy to verify that the normal pdf integrates to 1 and the values of the parameters μ and σ^2 are indeed the mean and variance of random variable X. We will do it later in this chapter. We sketch below the pdf of the normal probability curve with $\mu = 0$ and $\sigma = 1$:

Figure 9.5: Standard Normal Distribution Curve

Some salient features of the normal distribution are listed below:

- All normal distributions have the same overall shape. The curve extends from $-\infty$ to $+\infty$
- The exact density curve for a particular normal distribution is specified by giving its mean and standard deviation.
- The mean is exactly at the center of the symmetric curve.
- Changing of mean without changing standard deviation shifts the normal curve along the horizontal axis without changing its spread.
- The spread of the normal curve is controlled by its standard deviation. More the spread of the normal curve more is its standard deviation.
- It has a symmetric bell-shaped density curve. (but there are other distributions that also have bell-shaped density curves but are not normal.
- The height of the curve at any point x is given by $\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
- For normal distribution mean = median = mode;
- Index of skewness $\beta_1 = \mu_3 / \mu_2^3 = 0$ and index of kurtosis = $\beta_2 = \mu_4 / \mu_2^2 = 3$

Standard Normal curve

The 68-95-99.7 rule:

- 68% of the observations fall within 1* S.D. of the mean

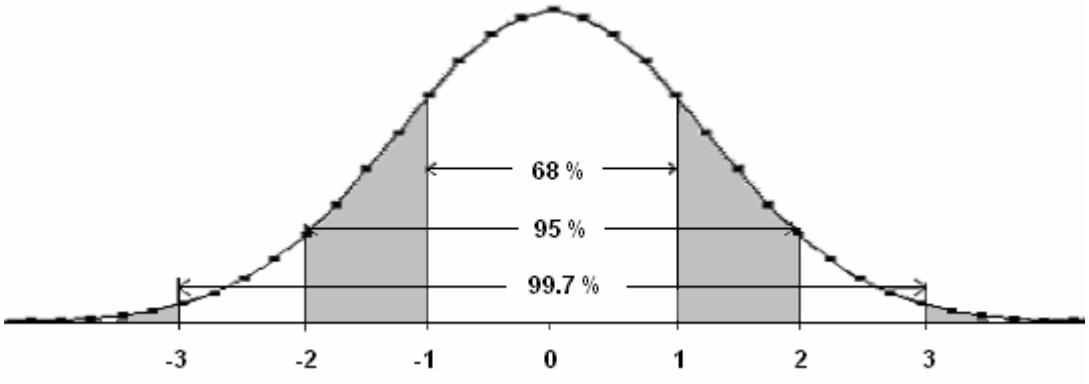


Figure 9.6: 68-95-99.7 rule

- 95% of the observations fall within 2* S. D. of the mean;
- 99.7% of the observations fall within 3 * S. D. of the mean.

Table 9.1 : Cumulative probabilities for standard normal distribution are reported below.

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.40	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.30	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.20	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.10	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.00	**0.0228**	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	**0.0475**	0.0465	0.0455
-1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.20	0.1151	0.1131	0.1112	0.1093	0.1075	**0.1056**	0.1038	0.1020	0.1003	0.0985
-1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.00	**0.1587**	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.60	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.00	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	**0.5000**	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	**0.8413**	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	**0.9772**	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	**0.9987**	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Standardizing observations:

Normal distribution is a good description for some distributions of real data. For example scores in CET or MBA entrance examinations, characteristics of human populations such as height, weight, body surface area, head circumference and other anthropometric measurements, yield per hectare of paddy, lengths of individuals of caecilians, etc. are often close to normal distribution. Normal distribution is also a good approximation of the results that originate from chance outcomes such as tossing of a coin or a die many times or noting of a colour of a pebble drawn from an urn containing pebbles of different colours.

In general a normal distribution can have any mean μ (any real number) and standard deviation σ (> 0). It is many times necessary to standardize these distributions on a common platform in order to compare the individual distributions and observations. *In order to standardize a value subtract the mean of the distribution and then divide by the standard deviation. A standardize value is called a* ***z-score***. Thus if x is an observation from a distribution that has mean μ and standard deviation σ, then the standardized value of x is $z = (x - \mu) / \sigma$. It is referred as z-score (of x). z-score indicated how many standard deviations the original observation falls apart from the mean and in its direction. If z-score is positive (negative) then it means that original x observation is above (below) mean and vice versa. It can be shown that if X has normal distribution with mean μ and variance σ^2 then $Z = (x - \mu) / \sigma$ is a standard normal variate.

Example 9.8: Rama scored 75 marks in HSC examination. Mean and s.d of the marks was 65 and 8 respectively. Z-score of Rama's mark would be $z = (75 - 65) / 8 = 1.25$. Krishna scored 86 marks in CBSE examination. In CBSE examination, mean and s.d. are 82 and 4 respectively. Z-score of Krishna would be $z = (86 - 82) / 4 = 1$. Since 1.25 > 1, it can be argued that Rama's performance is better than Krishna's performance.

Problems based on area under Standard Normal Curve.

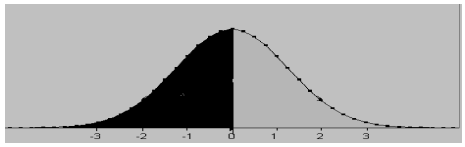
The probability that a continuous variable Z takes values in the interval (a, b) can be obtained using the following integral

$$P(a < Z < b) = \int_a^b f(z)dz = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\{-\frac{z^2}{2}\}dz = F(b) - F(a)$$

Since the evaluation of this integral is not easy, statisticians have prepared tables. One such table (Table A) is given above. Using it you can obtain probabilities of various events for various values of b and a. In Table A, area from $-\infty$ to b is tabulated You can make use of symmetry property and $P(A') = 1 - P(A)$, to evaluate probabilities of events of interest. For example, for $X \sim N(\mu, \sigma^2)$. Assume that mean = 60 and s.d. = 10. To illustrate use of Table A to evaluate probabilities we will find probabilities of various events:

Define $Z = (X - \mu) / \sigma$. Z follows N(0, 1).

$P(X < \mu) = P(Z < [\mu - \mu] / \sigma) = P(Z < 0) = 0.5$

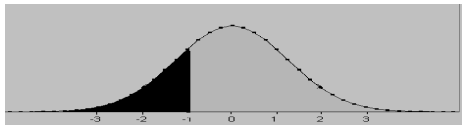


The required probability is read from Table A and corresponding area (black shaded area) is shown in the picture. Remember that in Table A, area from $-\infty$ to z (between -3.48 to 3.49, at the interval of 0.01) is computed.

1) $P(X < 50) = P(Z < [50 - 60]/10)$

$= P(Z < -1)$

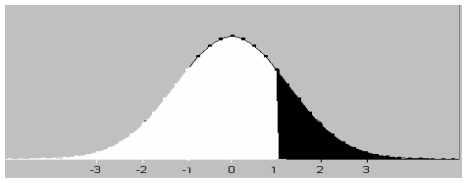
$= 0.1587;$



2) $P(X > 50) = 1 - P(X < 50)$

$= 1 - 0.1587$

$= 0.8413$

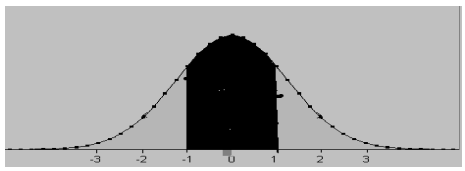


3) $P(50 < X < 70) = P(-1 < Z < 1)$

$= F(1) - F(-1)$

$= 0.8413 - 0.1587$

$= 0.6826$

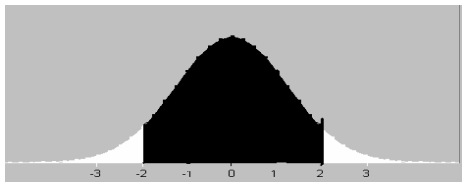


4) $P(40 < X < 80) = P(-2 < Z < 2)$

$= F(2) - F(-2)$

$= 0.9772 - 0.0228$

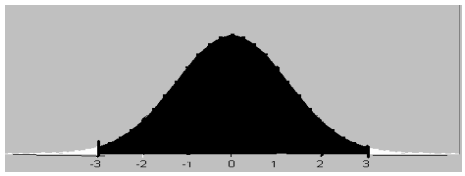
$= 0.9524$



5) $P(30 < X < 90) = P(-3 < Z < 3)$

$= F(3) - F(-3)$

$= 0.9974$



6) Find the value k such that $P(X < k) = 0.80$ where X is N (100, 100).

We are given that $P(X < k) = 0.80$. Suppose $Z = (x - \mu) / \sigma = (x - 100)/10$. Therefore

$P(X < K) = P(Z < [k - 100] /10) = 0.80.$

From Table A, below z = 0.84 the cumulative probability is approximately 0.80, hence we write $(k - 100) /10 = 0.84$. When we solve it for k, it gives k = 108.4.

7) The blood glucose level is important because high or low levels increase the risk of diabetics for an individual subject. Suppose that the distribution of blood glucose level is approximately normal in a large population of ordinary individuals of the same age and sex compositions. Levels above 140 units may require medical check up. Suppose blood glucose level is normally distributed with mean 100 units and standard deviation of 15 units. If 1000 individuals are checked for blood glucose level, how many of the will be

(a) above threshold value of 140 units,

(b) below threshold value of 75 units,

(c) how many individuals are expected to lie in between 80 to 120 units?

We will call the blood glucose level as X. The variable X has N (100, 15^2) distribution. We want the proportion of individuals above threshold value of 140 units. Now we subtract the mean from 140 and then divide it by s. d. = 15. In other words we standardize x values. Thus (X>140) is same as (Z > [140 – 100] /15) = P(X > 2.67) = 1 -0.9962 = 0.0038. It means amongst 1000 adult healthy normal individuals only 38 may be above threshold value of 140 units. Similarly X < 75 is same as Z <([75 – 100] /15 – 1.67) = 1.67; hence P(X < 75) = P(Z < -1.67) = 0.0475. It means about 47.5 (rounded to 48) individual in 1000 normal individuals may have blood glucose level below 75 units. Lastly we need to compute P(80 < X < 120) which is same as P(-1.25 < Z < 1.25) = F(1.25) – F(-1.25) = 0.8945 – 0.1056 = 0.7889; it means approximately 789 of 1000 adult normal individuals would have blood glucose level in between 80 to 120 units.

9.7.3 Some more Results from Normal Distribution (without proof)

1. Distribution of aX + b

Suppose X ~ N(μ, σ^2). Consider a variable Y = aX + b, where a and b are any real numbers. Y is a linear transformed variable, It can be shown that Y is also a normal variable and has mean aμ + b and variance $a^2\sigma^2$.

2. Suppose X ~ N(μ_x, σ_x^2) and Y ~ N(μ_y, σ_y^2). Further assume that X and Y are independent random variables. Then it can be shown that aX + bY + c follows normal distribution with mean $a\mu_x+b\mu_y+$ c and variance $a^2\sigma^2_x + b^2\sigma^2_y$.

(Here a, b and c are any real constants.)

In particular suppose a = b = 1 and c = 0, Then we get X + Y ~N($\mu_x+\mu_y, \sigma^2_x+\sigma^2_y$).

If a = 1, b = -1 and c = 0, then X - Y ~N($\mu_x- \mu_y, \sigma^2_x+\sigma^2_y$).

Example 9.9: Suppose energy intake of an adult individual follows normal distribution with mean intake of 2200 kcals with a standard deviation of 400 kcals. Further assume that his energy expenditure is also normally distributed with mean energy expedite 2000 kcal with a standard deviation of 425 nits. On how many days do you expect that the adult individual will be in a positive balance?

We assume that the energy intake (X) of an adult individual follows N(2200, 400^2) and his energy expenditure (Y) follows N(2000, 425^2). We further assume that X and Y are independent random variables. Define B = X – Y, here B denotes energy balance in kcal. If X > Y, the individual is in positive balance and if X < Y, then he would be in negative balance. It is easy to check that the energy balance, B follows N(200, $400^2 + 425^2$). We want to find P(X > Y). It is same as finding P(B = X - Y > 0) = P(Z > [0 – 200]/583.6309 = -0.34) P = 1 – P(Z < -0.34) = 0.3669; it means in a month of 30 days, the adult individual is expected to be in a positive energy balance on 30 *0.3669 = 11 days on an average.

We state below a general result related with linear combinations of n independent random variables:
Suppose X_i ~ N ($\sum a_iX_i$) = $\sum a_i\mu_i$ for i = 1, …, n and Cov(X, X_j) = 0_j for i≠j (this is because variables are independent)

Define $X = a_1X_1 + a_2X_2 + \ldots + a_nX_n$ where a_i's (I = 1, 2…, n) are constants, then X has a normal distribution. Its mean and variance is given by

$\mu = E(X) = E(\sum a_iX_i) = \sum a_i\mu_i$ and

$\sigma^2 = Var(X) = Var(\sum a_iX_i) = \sum a_i^2V(X_i)$

If in particular if X_i's independent identically distributed $N(\mu, \sigma^2)$, then the sample mean would follow $N(\mu, \sigma^2/n)$ distribution. This result is important and is useful. Its proof of it is left as an exercise for students.

9.7 Sampling Distributions

In this section we will introduce the concept of sampling distributions and illustrate it. We also discuss its application in tests of hypothesis. The purpose of statistical data analysis is to make sample based inference about the population e.g., when a researcher tests the efficacy of a new drug over the earlier drug, he wants to infer about the performance of a new drug. Whenever a statement is made such as "a random sample X_i (i = 1, 2 …n) is drawn from a certain population", it means the observations X''s in the sample are independent and moreover these observations arise from the identical distribution. We then proceed to define a new random variable based upon these observations (and some known constants). It does not depend on any unknown parameters. This new random variable is called a "statistic". The statistic so defined is a real valued function of the sample observations. For example sample mean, sample variance or sample proportion are examples of a statistic. It is important to realize that statistic is also a random variable. The statistic can take different values with different probabilities. In other words, the statistic such as a sample mean will have its own probability distribution. This particular distribution of statistic is called "sampling distribution of the statistic". Study the following example.

Example .10: Suppose you are given a population consisting of N = 5 observations and the population values are {1, 2, 3, 4, 5}. You are asked to draw a random sample of size 2 without replacement. Consider sample mean ($\bar{x}$) and sample mean square (s^2) as two statistics. We use sample mean as an estimator of population mean and sample mean square as an estimator of population variance. The following table gives list of all possible simple random samples (without replacement), values of sample means and sample mean squares.

Table 9.2 : List of all possible samples

sample no	sample values		Sample mean	Sample mean square
1	1	2	1.50	0.50
2	1	3	2.00	2.00
3	1	4	2.50	4.50
4	1	5	3.00	8.00
5	2	3	2.50	0.50
6	2	4	3.00	2.00
7	2	5	3.50	4.50
8	3	4	3.50	0.50
9	3	5	4.00	2.00
10	4	5	4.50	0.50
Total			30.00	25.00
Mean "of all samples"			**3.00**	**2.50**
Population mean			**3.00**	
population mean square			**2.50**	

Notice that mean of all sample means is exactly equal to the population mean (= 3.00) and mean of sample mean square equals population mean square. In statistical jargon we say that sample mean is an unbiased estimator of population mean, and also sample mean square is an unbiased estimator of population mean square. Therefore whenever population mean is unknown, we estimate it by sample mean, which is an unbiased estimator of population mean. Remember that the sample should be random, then only above results hold.

Suppose for time being that population mean is unknown. We can prepare sampling distribution of sample mean as follows:

Table 9.3 : Sampling distribution of sample mean

Sample mean	1.5	2.0	2.5	3.0	3.5	4	4.5
P(sample mean)	1/10	1/10	2/10	2/10	2/10	1/10	1/10

We have defined E(X) = $\sum_x xP(x)$

The mean of all possible samples is therefore given by

E(X) = (1.5)(1/10) + (2)(1/10) + (2.5)(2/10) + (3)(2/10) + (3.5)(2/10) + (4)(1/10) + (4.5)(1/10) = 30/10 = 3

In this illustrative example we have considered SRSWOR (simple random sample without replacement); you may repeat the above exercise for SRSWR (simple random sampling with replacement) procedure. Verify that sample mean remains an unbiased estimator of population mean.

In this part we will not derive any proof related with sampling distribution of a statistic. Instead we will state some important results and properties of statistic and its distribution. We will make use of these results in the study of test of hypothesis.

1. **Sampling from normal distribution:** The sampling distribution of $\bar{x}$ is N (μ, σ^2/n)
2. **Standard error of $\bar{x}$: S.E. (standard error) of $\bar{x}$ is equal to** population standard deviation/√n, where n is the sample size drawn from the population under consideration.
   word about the term standard error (SE): SE of the mean is synonym for SD of the sampling distribution of mean.
   he Normal distribution plays a key role in sampling. If X_i (i = 1, 2, …n) is a random sample from N(μ, σ^2) and if $\bar{x}$ is the sample mean then the standardized variable Z = $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is always distributed as N(0,1). In case of large sample test (also referred as Z-test) we use Z as a test statistic if σ is known. You may wonder what happens if the random sample is not drawn from Normal population? It can be shown that even when the sample does not come from Normal distribution, under certain conditions, such as large sample size (n > 30) we can make use of the concept of CLT (central limit theorem) and claim that the sample mean follows normal distribution.
3. **Sampling distribution for a proportion:** Suppose that the population consisting of N individuals (units) can be classified into one of the two characteristics such as place of residence (urban/rural), gender (male/female), education (below HSC/HSC or above), eating habits (vegetarian/non vegetarian). The parameter of interest in such situation is proportion (say P) of units possessing the characteristic under consideration. Clearly P is the ratio of number of units possessing characteristic (say A), to the population size (N). Thus P = A/N, but it is usually unknown. Therefore we wish to estimate it. Suppose we draw a random sample of size n from the population and observe that 'a' units in the sample posses the characteristic. So the sample proportion p = a/n is a natural candidate to estimate population proportion P. In other words sample proportion is a statistic for estimating population proportion.
   It can be shown that E(sample proportion) = population proportion and the standard error of this estimator is $\sqrt{\frac{P(1-P)}{n}}$. But since P is unknown, and we estimate it by p so we get estimated standard error as $\sqrt{\frac{p(1-p)}{n}}$.
4. **Student's t- distribution:** This distribution is also a sampling distribution and is used in tests of hypothesis dealing with specified value of mean (one sample problem) or equality of two population means (two sample problem). A t-distribution is completely specified by its parameter. This parameter is called degrees of freedom (usually denoted by n). We have mentioned that sample variance S^2 can be used to make inferences about the population variance σ^2 in a normal distribution. Similarly sample mean $\bar{X}$ is useful when we estimate population mean μ; but the distribution of sample mean depends on σ^2; it may happen that population variance is unknown and the sample size is small. In such a situation you can not make use of Z-test. If you estimate population standard deviation by corresponding sample

quantity then the test statistic $\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ is no more N(0,1), but it does not depend on σ^2. William Gosset (1908) who wrote in the pseudo name of "Student" derived the distribution of the above test statistic. (It is the ratio of standard normal distribution to the square root of Chi-square distribution adjusted for its d. f.). This distribution is referred as Student's t distribution and is denoted as t_n. The percentage points of t-distribution for various values of d. f. are available in standard text books. These tables usually are available up to n = 30. The t-distribution is symmetric about zero and its general shape is similar to N(0,1). Compared to normal it is flatter with thicker tails. For large values of n > 30, t-distribution approaches normal distribution. Mean and variance of t distribution with n d. f. is respectively 0 and $\frac{n}{n-2}$. Notice that the variance of t-distribution exists only for n > 2. and is approximately 1 for large n.

5. **Chi-square distribution:** This is also a continuous distribution. If X_i's are n i. i. d. N(0,1) variables then $\sum_{i=1}^{n} X_i^2$ is distributed as a Chi-square (χ^2) random variable with n d. f. For small values of n, the chi-square distribution is skewed to the right. For large n, it may be approximated to normal distribution. Mean and variance of Chi-square distribution with n d. f. are respectively n and 2n. The tables of χ^2 distribution for various values of n are available. It can be shown that when the sample comes from Normal distribution $(n-1)\ S^2 / \sigma^2$ is a Chi-square random variable with n-1 d. f.

   Here we have mentioned the concept of sampling distribution of the test statistic. We make use of it in tests of hypothesis. For more study of it we refer students to books on distribution theory and statistical inference.

9.8 Self Test

1. **Identify whether the following variables are discrete or continuous. Also state the possible values of the variable or its range.**
   1. 1 A coin is tossed n times and the outcomes are recorded in a linear array. Number of single heads in the string is counted.
   2. A person visits a family having 4 children. He records the proportion of male children in the family.
   3. Number of accidents per day on Gangapur road leading to YCMOU is recorded for a week.
   4. Number of suicides in a region is recorded for a period of 12 months.
   5. Petal length and petal width of a species is recorded for 20 flowers.
   6. Litter size of different animal species is recorded.
   7. Blood pressure of patients admitted to a hospital is recorded.
   8. A researcher records the data on head circumference, chest circumference, height, weight and sex of children in a public school. Classify each variable.
   9. A social scientist studies the data on percent salary income spent on food by a household.
   10 A person tosses a coin till he observes k (= 3, say) heads successively.
   11. A student records waiting time for the arrival of a bus in an interval of (9, 9.30) a.m on a working day. Assume that he reaches the bus stand randomly in the specified time-interval. Buses arrive at stop after every 30 minutes.

12. I. Q of adult healthy individuals is a study variable.
13. Number of peas in a pod, length of a pod and weight of a pod is noted.
14. Residual life time of an electric component is being studied by an electronic engineer.
15. The number of defectives in a lot consisting of 100 units is reported.
16. A vehicle is tested for its mileage per gallon of petrol.

2. Check whether the following functions are wel-defined pmf (or pdf) of some random variable. If answer is yes, find its mean and variance.

2.1 $f(x) = x^2/30$ for x = 1, 2, 3 and 4
2.2 $f(x) = (-1)^x \log(x)$ for x = 1, 2, ….
2.3 f(x) = 1/x for x = 2, 3, 4, …
2.4 f(x) = 1/2 for x = 2, 3, 4.
2.5 $f(x) = {}^{10}C_x(1/2)^{10}$ for x = 0, 1, …10
= 0, otherwise
2.6 f(x) = 1/6 for x = 1, 2, …6 ;
= 0, otherwise.
2.7 f(x) = cx, for c > 0, x in (0, 2)
2.8 $f(x) = e^{-x}$ for x > 0
2.9 $f(x) = (1/2)e^{-1/2x}$ for x > 0
2.10 f(x) = 2*exp(-0.5x) for x > 0
2.11 f(x) = 1/100 for x in (0, 100)
2.12 f(x) = 1/(x(x-1)) for x = 2, 3, …
2.13 f(x) = ½ for x in[2,4]

3. State properties of CDF of a discrete random variable. Illustrate them with suitable example.

4. Give example of a discrete uniform distribution. State its pmf, hence obtain its CDF, E(X) and V(X).

5. State salient features of N(0, 1) variate. Draw its pdf curvre.

6. Assume that X ~N(0,1). Using normal probability integral table, find probabilities of following events:
X > 0; X > -1 , X > -2, X > -3; X < 1, X < 2, X < 3, -1 < X < 1; -2 < X < 2,-3 < X < 3
|X -1| < 1;

7. X ~ N(0,1). Find x if
(1) P(X < x) = 0.5 ; (2) P(X < x) = .95 (3) P(X < x) = 0.05
(4) P(-x< X < x) = 0.95 (5) P(X < 2x) = .25 (6) P(X < -x) + P(X > x) = 0.05

8. Assume that energy intake (X) and energy expenditure (Y) both are normally distributed respectively with means 2200, 2000 and s.d. 250 and 400 measured in appropriate units. Assume that X and Y are independent random variables.
Find (i) P(X > 2500), (ii) P(Y < 2500), (iii) P(X-Y < 0), (iv) P(X – Y > 200) (v) mean and variance of X-Y.

9. It is reported that yield of a crop is $N(500, 10^2)$. If the data on yield of crops is available from 1000 agricultural plots, what is the lowest yield of top 10% plots?

10. Explain with illustration concept of a sampling distribution of sample mean.

Unit 10 : Test of Hypothesis

10.1 Overview

In scientific or social research, we develop theories or propound new conjectures and then we conduct either experiments or conduct appropriate sample surveys to test them. It is rare that we achieve perfect control in any designed experiment, some unknown causes called chance causes always creep in and variation results. Also there is a limit to the number of times we get an opportunity to collect similar data. It is true that our measurements or observations are just a sample of reality. Like all other statistical methods, statistical test of a hypothesis uses available data as evidence and arrives at a conclusion in a decision making process. In this unit you will learn basic principles and statistical techniques that are used in tests of hypothesis and large sample tests.

10.2 Objectives

After studying this unit you will be able to

1. State a null hypothesis and alternative hypothesis.
2. Decide critical region and acceptance region for the statistical test.
3. Explain the concept of Type I error, Type II error and level of significance.
4. Perform a statistical test for testing theories about population mean μ and population proportion P.
5. Explain large sample tests. .
6. Identify the situations where large sample test is appropriate
7. Perform test for testing specified population mean and also test for equality of two population means (independent samples)
8. Solve simple problems that deal with framing and testing of appropriate null hypothesis against its alternative hypothesis

10.3 Introduction

In different disciplines such as biological and social sciences we are required to answer the questions about the validity of theories or hypotheses related to them. Statistical methodology is extensively employed in this process. Like all other statistical methods statistical test of a hypothesis uses available data as evidence and arrives at a conclusion. Performing a statistical test to arrive at a conclusion is a decision making process. Also it is a way of summarizing the conclusions from the data in a single statement. A statistical test is a scientific and objective process because the conclusions drawn are supported by the available data. Whenever you perform a statistical test it is your job to assure that these data have been collected in a scientific way. In this chapter we shall first study what statistical hypotheses are and how statistical tests work. We will discuss what kinds of errors are possible. We will then study some important statistical tests.

In order to understand the concept of a hypothesis, consider the following statements:

1. A coin is fair.
2. A die is balanced.
3. Small is beautiful but big is wonderful.
4. A rust resistant wheat variety has lower average yield than a nonresistant variety.
5. Educational material prepared by a researcher is better than the traditional material.
6. There are infinitely many prime numbers.
7. I am a good student.
8. 60% of my students are successful.
9. Managers in small scale factories experience more stress than the managers in large scale industries.

Some of the above statements can be verified, confirmed or discarded. For example, consider the statement (3) above. It can not be verified by collecting data. It is the opinion of an individual. Consider the statement (6). No data are required to prove the statement. It is a mathematical hypothesis and can be proved by using mathematical induction. Statement numbers (1), (2), (4), (5) (8) and (9) may be verified by designing appropriate experiments, collecting and analyzing relevant data using appropriate statistical methods. We will call such unproved statement as a statistical hypothesis. Simply speaking, hypothesis is an unproved statement. We will restrict only to statistical hypothesis.

10.4 Statistical Hypothesis

A statistical hypothesis is an unproved statement about the distribution and or parameters of the distribution of random variable(s) under consideration. In other words a statistical hypothesis is an assertion about the unknown distribution and/or parameter of one or more random variables. For example, suppose that a pharmaceutical company claims that its new drug brought in the market is effective in curing the disease. To test the claim, an experiment is planned and the drug is administered to 20 patients. Suppose X denotes the number of patients cured and p is the probability that a patient is cured. We assume p to be the same for all patients. Here X is a binomial random variable with parameters n = 20 and p. Suppose that company's claim is that the drug has 90% success rate. This claim needs to be verified by conducting appropriate clinical trial. In statistical language the drug manufacturer's claim is p = 0.90. It is a statistical hypothesis.

If a statistical hypothesis completely specifies the distribution and/or its parameters, then it is called a **simple hypothesis.** If it does not completely specify the distribution and/or its parameters then it is called **composite hypothesis.** In the above example the number of patients cured, X, follows binomial distribution with n = 20 and p = 0.9. It means that the distribution of X is completely specified, hence p = 0.90 is a simple hypothesis. Contrary to it, the claim was that the drug has success rate above 70%, then $p \geq 0.70$ is a composite hypothesis.

10. 5 Null Hypothesis and Alternative Hypothesis

In the above example p = 0.90 is equivalent to the statement that p-0.90 = 0. That is the difference between parameter value and the specified value is zero (i.e., null). Null hypothesis is usually denoted by Ho. In statistical procedures we always test a null hypothesis. In words of Sir Ronald Fisher "Null hypothesis is the hypothesis which is tested for the possible rejection under the assumption that it is true". We can now formally define "null hypothesis".

Definition: A null hypothesis is a statistical hypothesis. It is a relation between a function of the parameter(s) and its possible value(s).

Null hypothesis is a hypothesis which the researcher would like to claim to be correct unless the data gives strong evidence against it (for support of some alternative hypothesis). It is usually denoted by H_o.

Examples:

1. A coin is fair, i.e., H_o: $p = 0.5$
2. A six faced die is balanced, i.e., H_o: $p_i = 1/6$ for $i = 1, 2,\ldots 6$.
3. Average effect before treatment (μ_B) and after treatment (μ_A) is same. Here H_o: $\mu_B = \mu_A$

If a null hypothesis is rejected then the hypothesis that may be accepted as a consequence is termed as the alternative hypothesis**.** It is denoted either by H_1 or H_A. For example in case of the drug manufacturers claim a medical practitioner may set alternative hypotheses as H_1: $p < 0.90$ whereas in case of random experiment of tossing of a coin H_1: $p \neq 0.5$ and in case of tossing of a die it would be H_1: $p_i \neq 1/6$ for at least one $I = 1, 2,\ldots 6$.

Thus a statistical hypothesis can be classified into two attributes: (i) simple or composite, and (ii) null or alternative. In practice every statistical hypothesis is a combination of these two attributes.

The null and alternative hypotheses may be one of the following three types:

1. Ho: $\mu \leq \mu_o$ or Ho: $\mu = \mu_o$ Vs H_1 : $\mu > \mu_o$
2. Ho: $\mu \geq \mu_o$ or Ho: $\mu = \mu_o$ Vs H_1 : $\mu < \mu_o$
3. Ho: $\mu = \mu_o$ or Ho: $\mu = \mu_o$ Vs H_1 : $\mu \neq \mu_o$

The hypothesis in (1) and (2) above are one-sided and that in (3) is two-sided test. The corresponding tests will therefore be termed as one sided and two sided (tailed) tests accordingly.

10.6 Test of a Statistical Hypothesis

A test of a statistical hypothesis is a rule, which specifies for each possible observed data whether to accept or reject the null hypothesis under consideration.

10.7 Test Statistic

In order to test H_o, we perform an experiment. The experiment generates sample data. The sample data is then summarized into a single number. This number, which is a function of sample values, is called a test statistic. The formula for computation of the test statistic depends on H_o. The conclusion (acceptance/rejection) of the null hypothesis depends on the sample data only through the test statistic. We usually denote the test statistic by T. Note that this test statistic is strictly a function of sample values and known constants, if any. Hence we write $T = T(X)$. We compare observed value of the test statistic with standard (or reference) value to arrive at a conclusion.

Before we proceed further let us consider the drug manufacturer's problem. Suppose we want to test Ho: $p \geq 0.7$ against H_1: $p < 0.7$.

The drug is tested upon 100 patients and X the number of successfully treated patients is observed. . We can use $T(X) = X$ itself as a test statistic. Common sense suggests that we should accept Ho if $X \geq 70$ and reject if $X < 70$ is observed. Is this a good procedure? Note that even when $p = 0.70$ is correct, around 50% of the times a value of $X < 70$ will be observed. Thus a value like 69, does it form strong evidence against Ho? The generally accepted answer of this

question would be 'No'. What if X = 40 is observed? Again generally most people would say that X = 40 is strong evidence against Ho or is a strong evidence in favour of the alternative H_1. p < 0.70. Thus there should be a critical value x_o of X (x_o < 70) such that if X ≥ x_o is observed we would accept Ho and if X < x_o is observed then we would reject Ho in favour of H_1. Hpw to find such a critical value x_o? We will discuss this next.

10.8 Critical Region and Acceptance Region

When we make use of a statistical test procedure, it will involve a test statistic say T. Values of T will vary from sample to sample. The possible values of T will always belong to sample space S of possible values of T.. Some values of T lead to acceptance of H_o, whereas the others lead to rejection of H_o. The set of values of T that are associated with rejection of H_o is called the critical region (C). All the other values of T will lead to acceptance of H_o. It forms acceptance region (A). Note that C and A are mutually exclusive and exhaustive subsets of S. By this we mean that there is nothing common in rejection and acceptance region, i.e., $C \cap A = \phi$ and moreover $C \cup A = S$. In the earlier example of drug manufacturer, we can set critical region as C = {0, 1, …59} and acceptance region as A = {60, 61 …100}.

10.9 Type I Error and Type II Error

When we apply **appropriate** statistical test to verify our claim, we have to either accept H_o or reject H_o. If we reject H_o, then as its consequence we will be favouring acceptance of H1. Our decision is based on whether the value of the test statistic belongs to acceptance region or critical region. Can we say that our decisions will be always correct? Answer is no. Notice that we come across following situations.

(1) Accepting H_o, when it is true.

(2) Accepting H_o, when it is false.

(3) Rejecting H_o, when it is true.

(4) Rejecting H_o, when it is false.

Thus two kinds of errors are possible: we may reject H_o when it is true or we may accept H_o when it is false. These errors are respectively referred as Type I and Type II errors. We can summarize this discussion in table form as follows:

Conclusion	Reality	
	H_o is true	H_o is false
Accept H_o	No error	Type II error
Reject H_o	Type I error	No error

Since our conclusion is based upon the data generated in the random experiment (or based upon data available from random sample), these errors may occur with certain probabilities.

We denote P(Type I error) = α; i.e. P(Reject H_o|H_o is true) = α = P(T ε C |Ho is true)

And P(Type II error) = β, i. e P(Accept H_o| H_o is false) = β = P(T ε C |H_1 is true)

When a statistical test is constructed our aim is to reduce the possibilities of both type of errors and take a correct decision. Remember that both the errors cannot be minimized simultaneously. Therefore the proposed solution is to consider only that procedure for which P(Type I error) ≤ α (a fixed number). This is so because wrong rejection of Ho is considered to be more serious error as opposed to wrong rejection of H_1

Suppose you are given a drug. H_o is that the drug is poisonous; H_1 is, it is OK as a medicine. When you accept H_o it means you conclude that drug is poisonous. So you decide not to use it. If you reject H_o it means you believe that it is OK for use. Suppose you commit type I error. It means you intake poison. If you commit type II error, it means you neglect good medicine. Which error is serious? Certainly type I error. So we must minimize error that is more serious. However remember that whenever you apply a statistical test to a dataset and you arrive at a decision of rejecting Ho, either the decision is right or you have committed a type I error. Similarly if you have accepted Ho, either the decision is right or you have committed a type II error. You can not state what has exactly happened.

10.10 Level of Significance

The probability of committing type I error is called as the size of the test. The level of significance (l. o. s.) is the maximum probability of type I error we are willing to tolerate. It is usually denoted by α. It is many times expressed in percentage form. It is customary that in most of statistical work, α is taken as 5% or 1%, but it is not a sacrosanct value. The appropriate question would be, "How to decide the l. o. s. and who should decide it? ". The answer to the first question is not trivial. It is a job of subject expert who plans and executes the experiment. Now-a-days computer soft ware provides p-value of the test. The statistical soft ware can compute the probability of getting a result as extreme as (or more extreme than) the observed result under H_o. Such a probability is known as p-value of the test. With these preliminaries now we will discuss some large sample tests.

10.11 Large Sample Tests

In this part we shall consider some approximate tests, which are valid for sufficiently large sample only. The requirements of large samples may be considered as a drawback however large sample tests have wider applicability. These tests are applicable for all populations, not necessarily normal. Also these tests are comparatively easier to use in practice from computational point of view. Of course this is no more a merit as personal computers are now readily available. In dealing with sample mean or proportion the approximation is good if the sample size is 30 or more. The large sample tests make use of the properties of sample mean. These properties are as follows:

1. The expected value of sample mean $\bar{x}$ is always equal to population mean μ.
2. If population size N is large relative to the sample size n, then s. d.($\bar{x}$) = $\sigma / \sqrt{n}$, where σ is population s. d.
3. When n is large, $\bar{x}$ will be approximately normally distributed random variable as per the Central Limit Theorem (CLT) (where mean and variance are finite numbers).

10.11.1 Test for Specified Population Mean

Suppose X is a random variable with population mean μ and variance s^2, both unknown. Our interest is to test H_o: $\mu = \mu_o$ (specified). Suppose $X_1, X_2, \ldots, X_n$ is a random sample of size n from the population of X, then it can be shown that the distribution of sample mean, for large n, will follow Normal distribution with mean μ and variance s^2/n. it can be shown that for sufficiently large values of n, the test statistic Z, under H_o, follows N(0,1), where the test statistic Z is given by

$$Z = (\bar{x} - \mu) / (s / \sqrt{n}) \qquad \text{where} \quad \bar{x} = \sum x_i / n \quad \text{and} \quad s^2 = \sum (x_i - \bar{x})^2 / n$$

Test procedure: First decide null and alternative hypothesis.

Case 1: H_o: $\mu = \mu_o$ vs H_1: $\mu \neq \mu_o$

- This is a two-sided alternative hypothesis.
- Compute test statistic $Z_0 = (\bar{x} - \mu_0) / (s / \sqrt{n})$, based on the sample values.
- Decide level of significance α, usually $\alpha = 5\%$ or 1%.
- Reject H_o if $|Z_o| > Z_{\alpha/2}$; accept otherwise.
- Use standard normal probability integral tables and verify that

  for $\alpha = 5\%$, $Z_{0.025} = 1.96$ and for $\alpha = 1\%$, $Z_{0.005} = 2.58$.

Thus the test procedure is simple for two sided test for testing H_o: $\mu = \mu_o$ vs H_1: $\mu \neq \mu_o$:

If the observed value of the test statistic $Z_0 = (\bar{x} - \mu_0) / (s / \sqrt{n})$ exceeds critical value 1.96 at 5% l. o. s., then reject H_o and conclude accordingly; otherwise conclude that present evidence "fails to reject H_o".

Case 2: H_o: $\mu = \mu_o$ vs H_1: $\mu > \mu_o$

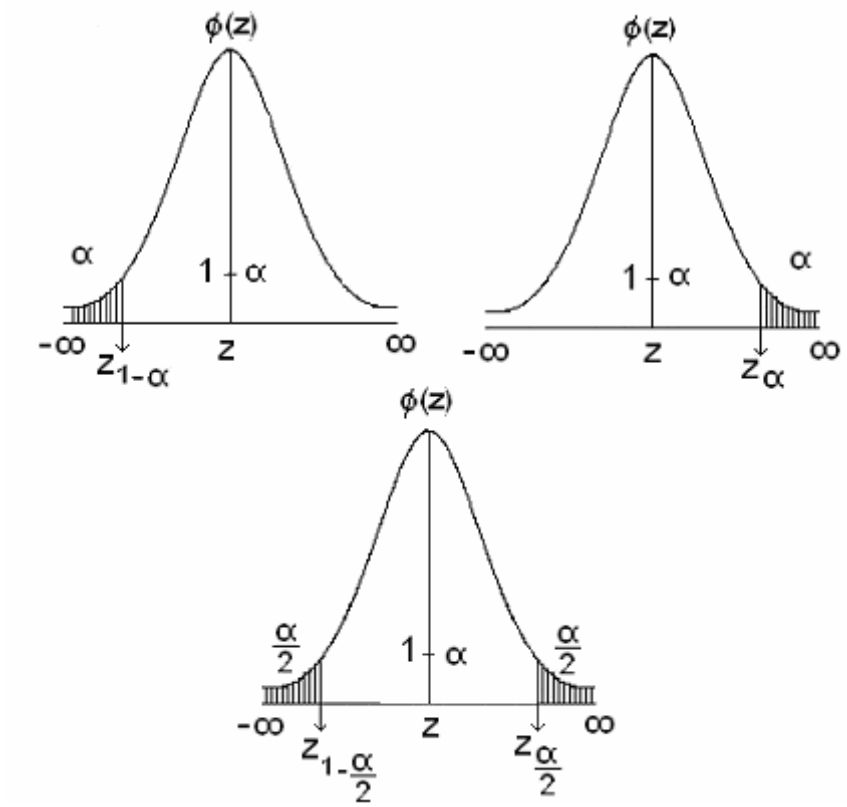
It is a one sided test. Here large values of $\bar{x}$ will favour the alternative H_1. So critical region will be to the right of the distribution. Here the test statistic is same as in case 1. The test procedure is simple. Reject H_o in favour of H_1 if the value of the test statistic Z_o exceeds Z_α.

Typically for $\alpha = 5\%$, $Z_{0.05} = 1.65$ and for $\alpha = 1\%$, $Z_{0.01} = 2.33$.

Case 3: H_o: $\mu = \mu_o$ vs H_1: $\mu < \mu_o$

This is one sided test; Here small values of $\bar{x}$ will favour the alternative H_1. So critical region will be on the left side of the distribution. Again the test statistic is same as above. The test procedure is : Reject H_o in favour of H_1 if the value of the test statistic Z_o is less than $Z_{1-\alpha}$. (or if $Z_o < -Z_\alpha$).

Typically for $\alpha = 5\%$, $Z_{0.05} = 1.65$ and for $\alpha = 1\%$, $Z_{0.01} = 2.33$



Example: A manager claims that the average daily yield of a product is 1000 metric tons per day. The sample based upon 50 observations yielded sample mean = 991 and s = 20 tons. Can you support the manager's claim?

Here we use two sided test: first we will state null and alternative hypothesis as H_o: $\mu = 1000$ vs H_1: $\mu \neq 1000$; now we compute value of the test statistic Z_0 :

$Z_0 = (\bar{x} - \mu_0) / (s / \sqrt{n}) = (991-1000) / (20/\sqrt{50}) = -9 / 2.8284 = -3.18$

Suppose l. o. s. α is 5%. Since $|Z_o| = 3.18 > Z_{.025} = 1.96$ (this is available from statistical tables), we reject H_o. The calculated value falls in the critical region. It appears that mean yield is less than 1000 tons per day. The present data fails to support the manager's claim.

10.11.2 Test for Specified Population Proportion

Let P be the proportion of members of a population possessing a characteristic A, where P is unknown. Suppose p is the proportion of the members in the random sample of size n possessing the characteristic A. our interest is to test H_o: $P = P_o$ (specified)

For sufficiently large n, the test statistic $Z = (\hat{P} - P) / \sqrt{P(1-P)/n}$ follows approximately N(0, 1). Therefore, under H_o, $Var(\hat{P}) = P_0(1 - P_0) / n$. We use it in the denominator of Z. So we compute the test statistic $Z_0 = (\hat{P} - P_o) / \sqrt{P_o(1-P_o)/n}$ where $\hat{P}$ =x / n, sample proportion.

(Comment: While computing the denominator of the test statistic, in place of P_o, some authors recommend use of $\hat{P}$)

Test procedure:

- To test H_o, compute the test statistic Z_o.
- The alternative hypothesis may be either one sided or two sided.
- We will reject H_o in favour of
    1. H_1: $P > P_o$ at α% l. o. s., if $Z_o > Z_\alpha$; accept H_o otherwise.
    2. H_1: $P < P_o$ at α% l. o. s., if $Z_o < -Z_\alpha$; accept H_o otherwise.
    3. H_1: $P \neq P_o$ at α% l. o. s., if $|Z_o| > Z_{\alpha/2}$; accept H_o otherwise.
- Depending upon our decision we will be have to conclude accordingly.

Example: A medical practitioner opines that the proportion of low birth weight babies have changed in the last decade. Ten years back the proportion of low weight birth babies was 22%. The recent survey report reveals that there were 425 low birth weight babies in 2000 babies born in last decade. Is this evidence strong enough to make a claim that birth weights have improved in the current decade?

Here we want to test H_o: P = 0.22 against H_1**:** P < 0.22

The value of the test statistic is statistic $Z_0 = (\hat{P} - P_0) / \sqrt{P_o(1-P_o)/n} = -0.8097$

[Verify that $Z_0 = (0.2125 - 0.22) / \sqrt{0.22(0.78)/2000} = -0.0075 / \sqrt{0.1716/2000}$

$= -0.8097$]

At 5% l. o. s., $-Z_{0.05} = -1.96$; since $-0.8097 > -1.96$, we again conclude that we do not reject H_o. in other words we can not support the claim that proportion of low birth weight babies have decreased significantly from earlier proportion of 22%.

10.11.3 Test for Comparing Two Means

There are many occasions when we need to compare the responses in two groups. For example, during a clinical trial one group of subjects receives a treatment and a control group receives placebo. A psychologist develops a test to test the emotional quotient of male and female students wants to check research hypothesis that emotional quotient of females is higher than the males. A mutual fund company wishes to know which of the two incentive plans are preferred by its customers. In such problems each group is considered to be a sample from a distinct population. It is further assumed that the responses in each group are independent of those in other groups.

Suppose $(X_1, X_2, \ldots X_m)$ and $(Y_1, Y_2, \ldots Y_n)$ are two independent random samples of sizes m and n respectively drawn from the populations of X and Y. We assume that X and Y are two random variables with means μ_X, μ_Y and variances σ_1^2, σ_2^2 respectively. We assume that these population parameters are unknown.

We wish to test H_o: $\mu_X = \mu_Y$ based on the above samples. For sufficiently large m and n, compute the test statistic $Z = (\bar{X} - \bar{Y}) / \sqrt{(s_1^2/m) + (s_2^2/n)}$.

It follows N (0, 1) distribution under H_o.

Test procedure: First decide null and alternative hypothesis.

Case 1 (Two sided test): $H_o: \mu_x = \mu_y$ vs $H_1: \mu_x \neq \mu_y$

This is a two-sided alternative hypothesis.

The test statistic is $Z_o = (\bar{X} - \bar{Y}) / \sqrt{(s_1^2/m) + (s_2^2/n)}$.

Suppose level of significance is α, usually we consider α = 5% or 1%.

Reject H_o if $|Z_o| > Z_{\alpha/2}$; accept it otherwise.

e.g., for α = 5%, $Z_{0.025} = 1.96$ and for α = 1%, $Z_{0.005} = 2.58$.

Case 2 (One sided test): $H_o: \mu_x = \mu_y$ vs $H_1: \mu_x > \mu_y$

It is a one sided test, critical region will be to the right of the distribution. Here the test statistic is same as in case 1.

Test procedure: Reject H_o in favour of H_1 if the value of the test statistic Z_o exceeds Z_α. Typically for α = 5%, $Z_{0.05} = 1.65$ and for α = 1%, $Z_{0.01} = 2.33$.

Case 3 (One sided test): $H_o: \mu_x = \mu_y$ vs $H_1: \mu_x < \mu_y$

Here note that the critical region will be on the left side of the distribution. Again the test statistic is same as above.

Test procedure: Reject H_o in favour of H_1 if the value of the test statistic Z_o is less than $Z_{1-\alpha}$. (Or if $Z_o < -Z_\alpha$). Typically for α = 5%, $Z_{0.05} = 1.65$ and for α = 1%, $Z_{0.01} = 2.33$

Let us study following example and understand above test procedure.

Example: Does early pregnancy in the age-group 16-19 years cause their babies to have low birth weight? To study this question, birth weights of babies of women of ages 16-19 years (first pregnancy) were compared with birth weights of babies of women in the age groups 20+ years. All these women were of residing in similar locality and were "similar" in all respects. The summary statistics is as follows:

Group (age in years)	n	mean	s^2
16- 19	50	2.25	0.16
20+	40	2.10	0.40

Here our interest is to test H_o: $\mu_x = \mu_y$ against appropriate alternative hypothesis. For illustration sake we will consider all the three cases.

Suppose X is birth weight of a baby of age-group 16-19 years and Y weight of a baby of age group 20+ years.

Case 1: H_o: $\mu_x = \mu_y$

Average birth weight of babies in both the groups do not differ

vs H_1: $\mu_x \neq \mu_y$: Average birth weight of babies in two groups differ.

Using summary statistics given above we first compute the value of the test statistic Z.

$$Z_o = (\bar{X} - \bar{Y}) / \sqrt{(s_1^2/m) + (s_2^2/n)}$$

$$= (2.25 - 2.10) / \sqrt{(0.16/50) + (0.40/40)} = 0.15 / 0.1149 = 1.3055$$

At 5% l. o. s., the critical value is $Z_{0.025} = 1.96$; since observed value of the test statistic is 1.3055 which is less than the critical value 1.96, we conclude that the data fails to reject null hypothesis. The birth weights of babies in the two groups do not differ significantly.

Case 2: H_o: $\mu_x = \mu_y$ vs H_1: $\mu_x > \mu_y$. Here the alternative hypothesis is average birth weight of babies born to mothers of age-group 16-19 is more than the other group. Here the test statistic is the same and its value is $Z_o = 1.3055$ as before; however value of Z_0 favour H_1 So at 5% l. o. s. the critical value will be 1.65 which is more that 1.3055. Thus here again we conclude that the birth weights of babies in the two groups do not differ significantly.

Case 3 : H_o: $\mu_x = \mu_y$ vs H_1: $\mu_x < \mu_y$ is left as an exercise for students.

10.11.4 Test for Comparing Two Proportions

In social sciences comparative studies are common. There are several occasions when we wish to compare the proportions of two groups such as rural and urban, smokers and non-smokers, treated and control, men and women, etc. that have common characteristics. We will call each groups being compared as population 1 and population 2. Suppose the two population proportions are P_1 and P_2 (and these are unknown). We will use following notations:

Population	Population proportion	Sample size	Successes (in sample)	Sample proprtion
1	P_1	n	X_1	$p_1 = X_1 / n$
2	P_2	m	X_2	$p_2 = X_2 / m$

Our interest is to test H_o: $P_1 = P_2$ (two population proportions do not differ).

The test statistic is $Z = (p_1 - p_2) / \sqrt{p(1-p)[(1/n) + (1/m)]}$ where

p = pooled estimate of population proportion = $(X_1 + X_2) / (n + m)$

Test procedure:

Case 1 (Two sided): H_1: $P_1 \neq P_2$

Reject H_o in favour of H_1 if $|Z_o| > Z_{\alpha/2}$; accept it otherwise.

Case 2 (One sided): H_1: $P_1 > P_2$

Reject H_o in favour of H_1 if $Z_o > Z_\alpha$; accept it otherwise.

Case 3 (One sided): H_1: $P_1 < P_2$

Reject H_o in favour of H_1 if $Z_o < Z_{1-\alpha}$, which is same as $|Z_o| > Z_\alpha$; accept it otherwise.

10. 12 Self Test

1. Which of the following statements can be termed as statistical hypothesis?

a. Washing powder "A" is better than another washing powder "B".

b. There are infinitely many even numbers.

c. Fertility rate of women has changed during the last 100 years.

d. Voltage of an electric current is directly proportional to the resistance.

e. Cattle depredation is more frequent during rainy season.

f. A rust resistant wheat variety has lower average yield compared to traditional wheat variety.

g. A coin is biased.

h. Students from urban area perform better than students from rural area.

i. Flowers at the top have more seeds than flowers at the bottom of a plant.

j. Smoking and cancer are associated.

k. Children fed on vegetarian diet weigh less than their counterparts fed on non-vegetarian diet.

2. State whether following statements are true or false

(i) A statistical test of a theory gives absolute guarantee that the conclusions arrived is always true.

(ii) An experiment is planned perfectly. The data are collected. Using appropriate statistical test decision is taken to accept Ho at 5% l. o. s.

In this case you may say

a) Significant results by chance 5% of the time

b) Non-significant results 95% of the time.

c) Non-significant results even when the null hypothesis is false 5% of the time.

d) There is no strong evidence in the data to reject Ho.

(iii) Type I error means rejecting the null hypothesis when it is true.

(iv) Type I error means accepting the null hypothesis when it is false.

(v) Type II error means accepting the null hypothesis when the null hypothesis is false.

(vi) Type II error means accepting the alternative hypothesis when the null hypothesis is true.

(vii) Suppose that a researcher's theory is true and P(type II error) is large. It means the chances of proving the theory with a statistical test are poor.

(viii) Suppose that a researcher's theory is true and P(type II error) is large. It means the chances of proving the theory with a statistical test are excellent.

3. A random sample of 180 children in village Kirkatwadi showed that 22 were undernourished. Another random sample of 210 children in neighboring village Donje showed that 38 were undernourished. It was argued that proportion of undernourished children is more in village Kirkatwadi. Test appropriate hypothesis and comment on your results.

4. B. S. G. N. (Biodiversity study Group of Nasik) conducted a study in the Trimbakeshwar and Torangan area of Western ghats. In the study report, it was mentioned that 240 bird species were seen and of which 113 species were local. Can it be said that population proportion of local birds species in the study area is at least 50%?

5. A student is investigating price differences between city malls and Kirana shops in the state of Maharashtra. He determined the total price of a basket of commodities in 40 city malls and 60 retailer shops. The average price of the basket was respectively 560 Rs and 545 Rs respectively from malls and retailer shops and corresponding sample standard deviations were 40 and 50 Rs. Can you consider these two independent samples? If yes, do the average prices really differ in malls and retailer shops? Use 5% l. o. s. and comment on the results.

6. A tube manufacturer wished to compare the proportion of second quality tubes that are manufactured in different shifts. In the morning shift 25 out of 400 tubes produced are of second quality, while in 2nd and third shift this number is respectively 35 out of 500 and 20 out of 300 tubes. The quality control engineers claims that not more than 5% tubes manufactures are of second quality? Test appropriate null hypothesis separately for each shift. Also make comparison between different shifts w. r. t proportion of second quality tubes produced? Use 1% l. o. s.

Unit 11 : Small Sample Tests

11.1 Overview

The purpose of this unit is to present small sample statistical tests and estimation procedure for population means and variances. In the earlier chapter we have studied large sample tests. These tests rely on the CLT (Central limit theorem) to justify normality of the estimators and test statistics. These are valid only when sample sizes are large enough. If sample sizes are small the statistical techniques differ and therefore these techniques are studied separately. Here we therefore take a brief review of tests that are based on Student's t-statistic. In particular we study one sample and two sample problems. We also study tests based on chi-square distribution. It includes tests for goodness of fit and test for independence of two attributes.

11.1 Objectives

After studying this unit you will be able to

1. Identify the situations where small sample test will be appropriate
2. Perform test for testing specified population mean and also test for equality of two population means (independent samples)
3. Recognize when you have dependent samples and then apply paired 't' test
4. Distinguish between paired t test and test for independent samples
5. Apply appropriate test for testing null hypothesis of equality of population means.
6. Apply test for specified value of population variance and for equality of two population variances.
7. Understand the role of chi-square distribution and chi-square test statistic in testing goodness of fit
8. Perform the test of independence of two attributes.
9. Solve simple problems that deal with framing and then testing appropriate null hypothesis against its alternative hypothesis.

11.2 Introduction

Large sample tests for making inferences concerning population means and proportions were discussed in the last unit. For some reasons such as cost, expertise, availability of time and technical resources the size of the sample may be restricted. In such situations tests based upon large sample sizes are of little use. We therefore study small sample inferential tests. They almost look like closely related to large sample tests. But there is a difference.

We have stated earlier that for large n, according to CLT, $Z = (\bar{x} - \mu) / (\sigma/\sqrt{n})$ follows N (0, 1) distribution. When the sample size is large we can replace σ by s its estimate from the sample and CLT continues to hold. But when the sample size is small, the distribution of the

quantity $t = (\bar{x} - \mu) / (s/\sqrt{n})$ for samples drawn from normally distributed population is not normal. W. S. Gosset in the year 1908 derived its distribution. He published the result under the pen name of Student and since then it is known as Student's t. The distribution of the statistic t in repeated sampling is again bell shaped and perfectly symmetrical about t = 0. It is much more variable as compared to Z. Its distribution has thicker tails compared to normal distribution. The variability in t is contributed by sample mean and also by sample s. d. for small values of n. As n increases and becomes infinitely large, the distribution of t and Z become almost identical. We reproduce below critical values of t. Verify that critical values of t are always larger than the corresponding critical values of Z for a specified level of significance α. We will require following table of t-distribution.

Table 11.1: Student's t-table for some selected values of α

Degree of freedom (d.f.)	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0..010}$	$t_{0..005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	2.977
11	1.363	1.796	2.201	2.718	2.947
12	1.356	1.782	2.179	2.681	2.921
13	1.350	1.771	2.160	2.650	2.898
14	1.345	1.761	2.145	2.625	2.878
15	1.341	1.753	2.131	2.602	2.861
16	1.337	1.746	2.120	2.583	2.845
17	1.333	1.740	2.110	2.567	2.831
18	1.328	1.734	2.101	2.552	2.819
19	1.325	1.729	2.093	2.539	2.807
20	1.323	1.725	2.086	2.528	2.797
25	1.316	1.708	2.060	2.485	2.787
30	1.311	1.697	2.042	2.457	2.750
Infinity	1.282	1.345	1.960	2.326	2.576
Note : The critical values of 't' for various d.f and l.os; $[t_n > t_{n;\alpha}] = \alpha$					

Source: *"M. B. Kulkarni, S. B. Ghatpande, S. D. Gore (!998) "Common Statistical Tests", Satyajeet Prakashan, Pune*

11.3 Test for a Population Mean

There are many real life situations where the experimenter wishes to test whether the sample evidence can support or reject the claim that it comes from a population with a specified mean value. For example, a manufacturer of small electric motors claims that on the average they will not draw more than 0.8 amperes under normal load conditions. A sample of 16 motors was tested. It showed that the sample mean current is 0.96 amperes with s. d. of 0.32 amperes. In this case we wish to test $H_o: \mu = \mu_o = 0.80$

In general we have following set up for such a test.

Null hypothesis, H_o: $\mu = \mu_o$

Alternative hypothesis, H_1: The investigator depending upon the alternative values of the population mean he conjectures must specify this.

Case 1: H_1: $\mu \neq \mu_0$: This is a two sided alternative hypothesis.

Compute test statistic $t_0 = (\bar{x} - \mu_0) / (s/\sqrt{n})$, based on the sample values.

Decide level of significance α. From table of t-distribution read critical values at n-1 d. f and at α% level of significance.

Reject H_0 if $|t_0| > t_{n-1,\, \alpha/2}$; accept otherwise.

Thus the test procedure is simple:

For two sided test for testing H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$

Reject H_0 if value $t_0 = (\bar{x} - \mu_0) / (s/\sqrt{n}) > t_{n-1,\, \alpha/2}$, and conclude accordingly; otherwise conclude that present evidence "fails to reject H_o".

Case 2: H_1: $\mu > \mu_0$: This is a one sided alternative hypothesis.

Compute test statistic $t_0 = (\bar{x} - \mu_0) / (s/\sqrt{n})$, based on the sample values.

Decide level of significance α, usually α = 5% or 1%. From table of t-distribution read critical values at n-1 d. f. at α% level of significance.

Reject H_0 if $t_0 > t_{n-1,\, \alpha}$; accept otherwise.

Thus the test procedure is simple:

For two sided test for testing H_0: $\mu = \mu_0$ vs H_1: $\mu > \mu_0$

Reject H_0 if $t_0 = (\bar{x} - \mu_0) / (s/\sqrt{n}) > t_{n-1,\, \alpha}$, and conclude accordingly; otherwise conclude that present evidence "fails to reject H_o".

Case 3: H_1: $\mu < \mu_0$: This is a one sided alternative hypothesis.

Compute test statistic $t_0 = (\bar{x} - \mu_0) / (s/\sqrt{n})$, based on the sample values.

Decide level of significance α. From table of t-distribution read critical values at n-1 d.f. at α% level of significance.

Reject H_0 if $t_0 = (\bar{x} - \mu_0) / (s/\sqrt{n}) < t_{n-1,1-\alpha}$, (this is same as $-t_0 > t_{n-1,\, \alpha}$ by symmetry of t distribution) and conclude accordingly; otherwise conclude that present evidence "fails to reject H_0".

Example 11.1 : A refilling machine is set at 1000 gm to fill the iodized salt into plastic bags. A consumer activist visits a mall and records the weight of eight randomly selected bags. The weights of these 8 bags are 1005, 995, 985, 995, 998, 1005, 985, 990. Do the data support the claim that average weight of a bag is 1000 gm? (Use 5% l. o. s)

Here H_0: $\mu = 1000$ gm vs H_1: $\mu \neq 1000$ gm

We are given 5% l. o.s., therefore we place 0.025 probability in each tail of the t-distribution. Critical value of t at 7 d. f is 2.365. It means we will reject H_0 if $t_0 > 2.365$ or $t_0 < -2.365$.

Here $t_0 = (\bar{x} - \mu_0) / (s/\sqrt{n}) = (994.75 - 1000) / (7.87 / \sqrt{8}) = -1.89$. Since the observed value of t is in the acceptance region, it means we fail to reject H_0. We therefore favor the claim that average weight of bags is 1000 gm.

The consumer activist may test H_0: $\mu = 1000$ gm vs H_1: $\mu < 1000$ gm

Notice that the test statistic and its value remain unchanged, but critical region changes. Now critical region consists of values of the test statistic that are below -1.86. Since $t_0 = -1.89 < t_{8, 0.95} = -1.86$, we reject H_0 at 5% l. o. s. We conclude that there is no evidence to support H_0. It means average weight of refilled bags is less than 1000 gm.

11.4 Test for Equality of Two Population Means

Here we will study statistical tests which we can use to investigate the differences between means of two sets of observations. You will learn about two scenarios. The first one is the case in which you have data on two samples selected at random from two different populations. The other case is in which observations appear in pairs. For examples, you may have situations "before" and "after" treatment.

11.4.1 Test for Independent Samples

Here we assume that we have independent random samples of sizes m and n. These samples are drawn from normal populations which possess means variances μ_1, σ_1^2 and μ_2, σ_2^2. We further assume that the variances of these two normal populations are equal to σ^2, even though the common value of σ^2 is not known.

Here we test **H_0:** $\mu_1 = \mu_2$, against **H_1:** $\mu_1 \neq \mu_2$

The test statistic is $t = (\bar{X} - \bar{Y}) / \sqrt{s^2[(1/m) + (1/n)]}$

where s^2 = pooled estimator of σ^2 and is given by

$$s^2 = [(m-1)\, s_1^2 + (n-1)s_2^2] / (m+n-2) = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2}{m+n-2}$$

The (m+ n -2) is "number of degrees of freedom (d. f.) associated with s^2.

Rejection region: Reject H_0 if $|t| > t_{\alpha/2}$ where $t_{\alpha/2}$ is based on m+n-2 d. f. (see statistical table of t-distribution such as Table 11.1); otherwise accept H_0 and conclude accordingly.

In case of one-sided test, the rejection region will be only on left tail or right tail as per alternative hypothesis.

For example if **H_1:** $\mu_1 > \mu_2$, then we will reject H_0: $\mu_1 = \mu_2$.

if $t_0 > t_{m+n-2, \alpha}$ and conclude accordingly; otherwise we conclude that the present evidence fails to reject H_0 at α% level of significance.

If **H_1:** $\mu_1 < \mu_2$, then the test procedure is:

Reject H_0: $\mu_1 = \mu_2$ if $t_0 < t_{m+n-2, 1-\alpha}$ (which is same as $-t_0 > \mathbf{t_{n-1, \alpha}}$) and conclude accordingly; otherwise we say that the present data fails to reject H_0 at α% level of significance.

Example 11.2 : It is common practice to develop communication kit that is suited for training of workers who work at grass root levels. A researcher from the field of preventive medicine developed a new teaching communication kit for training Aanganwadi workers. On one such occasion, 15 Aanganwadi workers attended a workshop. For testing the efficacy of the new teaching communication kit these workers were randomly placed into two groups. For workers in group 1, training was given as per standard practice and for workers placed in group 2 the new communication kit was experimented. The workers were trained for 40 hours. After a gap of two weeks, a test was conducted to compare the efficacy of new communication kit. All the workers received scores out of 100. These scores are given below. Do the data present evidence that the two methods differ? Use $\alpha = 5\%$).

Table 11.2: Data on scores

Standard procedure	64	65	62	63	82	83	72	
New procedure	62	68	70	58	80	70	86	66

Suppose μ_1 and μ_2 denote the mean score of standard and new procedure respectively. The summary statistics is as follows

Table 11.3: Summary statistics

Group	sample size	Sample mean	sample mean square (s^2)
1	7	68.57	51.9524
2	8	71.38	57.4107

Pooled estimate of common variance is $s^2 = 54.8915$

Here H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 < \mu_2$

i.e. we test the null hypothesis that there is no difference in mean effect between the two procedures against the alternative that new method is better than the standard procedure.

Here $t_0 = (\bar{X} - \bar{Y}) / \sqrt{s^2[(1/m) + (1/n)]} = -2.81 / 3.8345 = -0.7328;$

Table value of $t_{13, 0.95} = -1.77$ (by symmetry of t- distribution).

Since $t_0 = -0.7328 > t_{13, 0.95} = -1.771$, we conclude that the available evidence fails to reject H_0. In other words, the new teaching kit and traditional method of teaching have same effect on training of Aanganwadi workers.

11.4.2 Test for dependent Samples (Paired t-test)

Here we will first study an illustrative example. A student took readings on heights of his five randomly selected friends. He was curious and was eager to verify whether there is significant difference between height of an individual when measured with shoes and without shoes. He collected following data:

Table 11.4: Data on heights of students

Student	X: Height (with shoes) (in cm)	Y: Height (without shoes) (in cm)
Ashish	166	163
Neeraj	168	164
Chinmay	158	156
Harshad	165	162
Hitesh	169	167

H_0: Here H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$

Test statistic is $t_0 = (\bar{X} - \bar{Y}) / \sqrt{s^2[(1/m) + (1/n)]}$

$= (165.2 - 162.4) / \sqrt{17.5[(1/5) + (1/5)]} = 2.8 / 2.6457 = 1.058.$

Since $t_0 < t_{8,.025} = 2.306$, he accepted H_0. He remarked that data supports the claim that wearing of shoes or otherwise has no effect on average height of students. However he grumbled, "something is wrong".

Question: Are you also puzzled or surprised by the student's conclusion? The commonsense tells us that two groups must differ in average height. The height of an individual has to be more when he wears the shoes than he does not. We can also think of alternative approach. Suppose we reanalyze the data as follows:

Table 11.5: Data on differences

Student	Ashish	Neeraj	Chinmay	Harshad	Hitesh
d = x – y	3	4	2	3	2

H_0: $\mu_D = 0$, i.e., mean difference is zero

H_1: $\mu_D \neq 0$, i. e mean difference is not zero.

Here we are making use of one sample t-test.

Our test statistic is $t = \bar{d} / (s_d / \sqrt{n}) = 7.48$ (verify it). Since it exceeds the critical value of $t_{4,.0.025} = 2.776$, we reject H_o and conclude that average height of a student wearing shoes is more than his height when he is without shoes. This is not at all surprising. It is trivial.

The statistical test that we have performed just above is referred as a paired t-test. Remember that the pairing was inherently there when the experiment was performed. It was not artificially prepared after the data were collected.

The paired t-test is equivalent to H_0: $\mu_D = 0$ where $\mu_D = \mu_1 - \mu_2$. (Refer section 11.3). Here the observations on (X, Y) are likely to be correlated. So remember that unlike t-test for independent samples, paired t-test does not require the assumption of equality of variances of two populations. In paired t-test the concept of "pairing" of individuals is important. Usually pairs of similar individuals or units are selected. Identical twins are natural pairs in biological experiments. In fact while using paired t-test we are using the distribution of D = x – Y and test the hypothesis that the population mean of D is zero. If the measurement at the end of experiment is the subject's ability to perform some task, then subjects similar in natural ability and previous training to some extent need to be paired. Here ability means response to a treatment. It is therefore necessary that researchers justify pairing of individuals on scientific grounds and then perform paired t-test.

We now summarize the procedure of paired t-test:

Step 1: Set H_0: $\mu_D = 0$

Step 2: Using paired data obtain differences D = X- Y

Step 3: Find the value of the test statistic $t = = \bar{d} / (s_d / \sqrt{n})$

Step 4: Compare the observed value of the test statistic with critical value at α% l. o. s. at n-1 degrees of freedom.

Case 1: H_1: $\mu_D = \mu_1 - \mu_2 > 0$, i. e, $\mu_1 > \mu_2$

Reject H_0 if $t_0 > t_{n-1, \alpha}$ at 100 α% l. o. s.; accept H_0 otherwise.

Case 2: H_1: $\mu_D = \mu_1 - \mu_2 < 0$ i. e, $\mu_1 < \mu_2$

Reject H_0 if $t_0 < t_{n-1,1-\alpha}$ at 100 α% l. o. s.; accept H_0 otherwise.

(Because of symmetry of t-distribution we can also reject H_0 if $-t_0 > t_{n-1, \alpha}$).

Case 3: Here H_1: $\mu_D \neq 0$, i.e., $\mu_1 \neq \mu_2$. In this case the test procedure is

Reject H_0 if $|t_0| > t_{n-1, \alpha/2}$ at 100 α% l. o. s.; accept it otherwise.

11.5 Tests for Variances

We have studied large sample tests and small sample tests for population mean. In these tests while computing a test statistic whenever population variance σ^2 is unknown we estimate it by sample quantities. There are many situations when the subject experts may be interested not in comparing the means of the populations but comparing variability between two populations or verifying whether the population variance is some specified value. In this section we present tests to verify null hypothesis regarding population variance.

Recall that in comparing the means of two normal populations (independent samples) we have assumed that these populations have equal variances. Here we will first discuss test for population variance and then it will be followed by test for equality of population variances.

11.5.1 Test for Population Variance

There are many situations where population variance is the basic objective of investigations. Many times researchers are aware about the mean levels but they seek information on inherent variability in the population. It is possible that the researchers working in their fields of expertise may have some conjectures about specific values of σ^2. We have seen that sample mean square defined as $s^2 = \sum(x_i - \bar{x})^2 / (n - 1)$ serves as a good estimate of σ^2 under certain conditions. Since s^2 is also a random variable it will also have some probability distribution and its minimum possible value is zero. We assume that s^2 is based upon a random sample of size n from normal population. The quantity $\boldsymbol{\chi^2 = (n - 1)s^2 / \sigma^2}$ is called a chi-square variable and it follows chi-square distribution with (n –1) degrees of freedom (which is the parameter of the distribution). The statistical table of chi-square distribution is available (included in this book). We will study test for testing a specific value of population variance based on use of chi-square test statistic. To begin with we assume that population mean μ **is unknown**.

Here H_0: $\sigma^2 = \sigma_0^2$ and H_1: $\sigma^2 \neq \sigma_0^2$ (a two- tailed statistical test)

Test statistic: $\chi^2 = (n - 1)s^2 / \sigma_0^2$

Test procedure: Reject H_0 if $\chi^2 < \chi^2_{1-\alpha/2,\, n-1}$ or $\chi^2 > \chi^2_{\alpha/2,\, n-1}$ at α% l. o. s.

[For alternatives which are one-sided, you must use one tailed test; the test statistic is the same, you have to consider α either in upper tail or in the right tail according as alternative $\sigma^2 > \sigma_0^2$ or $\sigma^2 < \sigma_0^2$ respectively.]

Note: If the population mean is known, then while computing $\mathbf{s^2}$ use μ in place of sample mean $\bar{x}$. In this case the test statistic is $\chi^2 = (n - 1)s^2 / \sigma_0^2$, but it follows chi-square distribution with n degrees of freedom, accordingly we should find critical values of chi square distribution with n degrees of freedom at α% l. o. s and employ test procedure. For brevity of space we do not include details here once again.

Example 11.3: A tube manufacturer claims that tubes manufactured in his industry would possess a better strength and would lie within a range of 50 units (measures in appropriate scale). The quality control inspector took 10 readings on randomly selected manufactured items. These observations are given below. Do these data support the manufacturer's claim?.

Data: 97.6, 94.1, 93.0, 92.3, 110.3, 94.1, 94.1, 114.0, 95.7, 109.5

Here we wish to test H_0: $\sigma^2 = \sigma_0^2 = 156.25$ (Since 4σ = range = 50) against

H_1: $\sigma^2 \neq \sigma_0^2$ (a two- tailed statistical test)

We first obtain mean = 99.47 and hence $s^2 = \sum(x_i - \bar{x})^2 / (n - 1) = 69.61$

The value of test statistic is $\chi^2 = (n - 1)s^2 / \sigma_0^2$, = 4.01 (approx).

Since $\chi^2_{0.975,9} = 2.700$ and $\chi^2_{0.025,9} = 19.023$, we observe that computed value of test statistic = 4.01 lies between the acceptance region and therefore we support the claim of the tube manufacturer at 5% l. o. s.

11.5.2 Test for the Equality of the Variances

Here we assume that X ~ $N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. We also assume that independent random samples from the distribution of X and Y are available. Here parameters of the normal distribution are unknown. Here we wish to test $H_0: \sigma_1^2 = \sigma_2^2$ against $H_1: \sigma_1^2 \neq \sigma_2^2$.

To test the H_o, we are given sample of size m from the distribution of X and another sample of size n from the distribution of Y. Using these sample values we first compute the value of the test statistic $F = s_1^2 / s_2^2$, where $s_1^2 = \sum(x_i - \bar{x})^2 / (m-1)$ and $s_2^2 = \sum(y_i - \bar{y})^2 / (n-1)$

The test procedure is simple.

Reject H_0: $\sigma_1^2 = \sigma_2^2$ if (i) $F_0 < F_{(m-1, n-1);\ 1-\alpha/2}$ or (ii) $F_0 > F_{(m-1, n-1);\ \alpha/2}$

At 100 α% l. o. s.; accept H_o otherwise.

Note: In many text books, it is advised that while computing test statistic take greater mean square in numerator. It is not at all necessary. By doing so we restrict F distribution to the interval [1,∞), which is not the case. We know that the range of F is [0,∞). Appropriate statistical tables for the F-distribution are now available. These are included at the end of this book in the appendix.

Example 11.4 : The variability in the type of defects present in different batches during a shift depends upon waiting time between the two successive operations. Some technical changes were made with a hope to reduce the variability as well as the mean number of defects. Samples of sizes 15 and 10 measurements from two batches of two different shops of the company yield following information:

Table 11.6: Summary information

Batch no	Sample size	Sample mean	Sample mean square
1	15	12.5	2.50
2	10	12.1	2.00

Do the data support the claim that the process variability is less for batch 2?

Here the managers of the company believe that the mean number of defects in different batches is almost same (infact this also can be tested), but they are interested in knowing whether variability is really different in two batches or not. A more variability is a risk. The products may not be acceptable in an international market. So the problem has large implications.

Here $H_0: \sigma_1^2 = \sigma_2^2$ Vs $H_1: \sigma_1^2 \neq \sigma_2^2$

Suppose α=0.05, the value of the test statistic is *F is*

$F = s_1^2 / s_2^2 = 2.50 / 2.00 = 1.25$

Note that $F_{14,9,0.975} = 0.312$ and $F_{14,9,0.025} = 3.798$

As per the test procedure, since $F_0 = 1.25$ lies between above two critical values i. e, since $0.312 < F_o = 1.25 < 3.798$, we conclude that the test fails to reject null hypothesis. In other words we conclude that at 5% l. o. s. variability in the two batches is not significantly different. It means technical changes have not changed variability in the process. Since null hypothesis of equality of variances is accepted, we can proceed to test equality of means. To do this you will have to perform test for equality of population means (independent samples). This is left as an exercise.

11.6 Tests Based on Chi-Square Distribution

In social science researchers carry our sample surveys. These surveys result in count data. For example you may classify people according to their income group, food habits, interest in music, suffering from gastrointestinal diseases, etc.

This can be seen as a multinomial experiment. It is an experiment in which more than 2 outcomes are possible. In such experiment we assume that

1. The experiment consists of n identical trials.
2. The outcome of each trial falls into k distinct classes.
3. The trials are independent. P(outcome falls in i^{th} cell) = P_i for i = 1, 2,… k. and $\sum_{i=1}^{k} p_i = 1$.

   P_i's are called cell probabilities.
4. Suppose n_i denotes the number of trials in which the outcome belongs to cell i. Then $\sum_{i=1}^{k} n_i = n$

In the tests that we will be studying in this part we want to make inferences about these cell probabilities. It was Karl Pearson who proposed test statistic for testing hypothesis dealing with these cell probabilities.

11.6.1 A Test of an Hypothesis Concerning Specified Cell Probabilities

Suppose you toss a six faced die and observe following frequencies. Can you conclude that the die you have tossed is fair?

Table 11.7: Results of experiment of tossing of a die

Outcome on a die →	1	2	3	4	5	6	Total
Observed frequency	24	28	23	23	26	26	150

Here H_0: $P_i = 1/6$ for i = 1, 2, …6 against H_1: $P_i \neq 1/6$ for atleast one i

Here the test statistic is $\chi^2 = \sum_{i=1}^{k}\left(\frac{(O_i - E_i)^2}{E_i}\right)$, where $E_i = nP_i$

Here n = 150, hence E_i = 150 x (1/6) = 25 for each i. Therefore the value of the test statistic is χ^2 = 0.04 + 0.36 + 0.16 + 0.16 + 0.04 + 0.04 =0.80

Test procedure: We will reject if observed value of χ^2 exceeds table value of χ^2 at k-1 d. f seen at 100 α% l. o. s. Suppose α is 0.05,. then here k = 6 which gives $\chi^2_{5,0.05}$= 11.0705

Since Observed value is less than the table value we accept the null hypothesis at 5% l. o. s. and conclude that the die is a fair one.

Note: In this test we have demonstrated that how to fit a discrete uniform distribution to observed data. The important thing is you must obtain expected frequencies. To do it you must know the probability model that you are imposing on the observed data and you must estimate of parameters of the distribution (if unknown) and then use it to obtain expected frequencies. If observed and expected frequencies are given to you then the task is simple. You obtain test statistic, use test procedure and conclude. The d. f are usually equal to k-1-p where p denotes number of parameters that are estimated using data.

11.6.2 Test for Goodness of Fit

Example 11.5: (Goodness of fit: Normal distribution)

It was reported that a random sample of size 200 was drawn from N(50, 225). As a part of practical, students used the sample data to estimate parameters of the distribution and then they obtained expected frequencies. The results are shown below. Do you agree with the claim that the normal distribution is a good fit to these data? Use 5% l. o. s.

Table 11.8: Distribution of 200 random numbers

Class	Observed frequency (O_i)	Expected frequency (E_i)
13.20 – 20.90	2	3.5728
20.90 – 28.60	10	8.3032
28.60 – 36.30	16	18.8960
36.30 – 44.00	37	32.3500
44.00 --51. 70	43	41.6620
51.70—59.40	39	40.3660
59.40 –67.10	23	29.4220
67.10—74.80	13	16.1322
74.80—82.50	6	6.6540
82.50—90.20	5	2.0638
Above 90.20	0	0.5780
Total	200	200

Here null hypothesis is H_0: Normal fit is good.

It is easy to calculate observed value of the test statistic is $\chi^2 = \sum_{i=1}^{k}\left(\frac{(O_i - E_i)^2}{E_i}\right)$, (left as exercise). Verify that it equals 7.53 (approximately). Here at 5% l. o. s. and d. f = 11 – 1- 2 = 8, critical value of chi square distribution is 15. 507 which is larger than 7.53 and therefore we conclude that normal fit is appropriate for this data set.

Note: Students should learn computational techniques or software with which goodness of fit tests can be carried out on a given data set. For more illustrative examples and step by step procedure refer books on mathematical statistics such as Common Statistical Tests by Kulkarni, Ghatpande and Gore (1999).

11.6.3 Test for Independence of Attributes (Contingency tables)

In the analysis of count data, we come across situations where we have to answer the query – "Whether the two attributes under consideration are independent?" For example, the units produced in a day's production may be classified as confirming or not confirming to specifications. This production can also be classified according to 1st, 2nd or 3rd shift. The company manager may wish to investigate whether proportions of units not confirming to specifications has any association between shift? If the two classifications, say A_i and B_j are independent, then $P(A_i \cap B_j) = P(A_i)P(B_j)$ for all values of i and j, using this rule of independence of two events, expected frequencies are obtained. The difference between observed and expected frequencies is then compared using chi-square statistic. We first discuss below case 1 in which we prepare a 2 x 2 contingency table. In case 2, we will discuss r x s contingency table.

Case 1: A 2 x 2 contingency table

Suppose we have data for two attributes A and B recorded on N individuals. Our interest is to test the null hypothesis H_0: Attributes A and B are independent against its negation. For these data we can prepare a 2 x 2 contingency table as below:

Table 11.9 : Table of observed frequencies

Attribute A → Attribute B↓	Present	Absent	Row total
Present	a	b	$R_1 = a + b$
Absent	c	d	$R_2 = c + d$
Column total	$C_1 = a + c$	$C_2 = b + d$	$N = a+ b + c + d$

Under H_0, i.e. under the assumption of independence we can obtain expected frequencies of each cell above. These are shown below

Table 11.10 : Table of expected frequencies

Attribute A → Attribute B↓	Present	Absent	Row total
Present	(R1 x C1)/N	(R1 x C2)/N	R1
Absent	(R2 x C1)/N	(R2 x C2)/N	R2
Column total	C1	C2	N

To test H_0: Attributes A and B are independent, a test statistic based on chi-square (χ^2) distribution is suggested. It can be shown that it reduces to following formula:

Test statistic is $\chi^2 = (ad - bc)^2 * N / (a+b)(c + d)(a + c)(b + d)$ ---------- (1)

This statistic under H_o has a χ^2 distribution with (2-1)*(2-1) =1 degrees of freedom.

The test procedure is simple. First state the null hypothesis of independence, prepare a table of observed frequencies and use the formula in (1). You will get observed value of χ^2 test statistic. If the observed value is greater than the table value of χ^2 at 1 degree of freedom seen at 100 α% l.o.s, then reject H_o and conclude that attributes A and B are not independent; accept H_o otherwise and conclude accordingly.

Note: Chi-square test is a large sample test. It assumes that all frequencies are "large". If they are not large, then groups should be combined. If any expected frequency is less than 5, then pooling of groups is suggested. However in case of 2 x 2 contingency table use of pooling method for validity of χ^2 test results in χ^2 with zero d. f. - which is meaningless. Therefore in year 1934 Yates suggested correction for continuity. It consists of adding 0.5 to the cell frequency which is less than 5 and then adjusting for the remaining cell frequencies accordingly.

Example: Following are the data on preference for a colour of 52 students and the sex of the student. Can it be concluded that the colour preference is independent of sex ? Use 5 % l. o. s.

Table 11.11: Observed frequencies

Colour preference → Sex ↓	Pink	Other than Pink	Total
Female	17	3	20
Male	9	23	32
Total	26	26	52

Here H_0: Color preference and sex are independent

$\chi^2 = (ad - bc)^2 * N / (a+b)(c + d)(a + c)(b + d)$

$= [(17 * 23) - (3 * 9)]^2 * 52 /(20 * 32 * 26 * 26) = 13.73$

From statistical tables, $\chi^2_{1,0.05} = 3.841 < 13.73$; we therefore reject H_0 and conclude that colour preferences and sex are not independent. Here it appears that 85% of sampled females prefer pink colour as compared to only 28% of males preferring pink colour. Thus the colour preference can be used to predict the sex of a student.

Note that there are only 3 females who have preferred other than pink color. So let us apply Yates correction

Table 11.12: Observed frequencies (Yates' correction applied)

Colour preference → Sex↓	Pink	Other than Pink	Total
Female	17-0.5 = 16.5	3 +.5 = 3.5	20
Male	9 +0.5 = 9.5	23 - 0.5 = 22.5	32
Total	26	26	52

Now $\chi^2 = (ad - bc)^2 * N / (a+b)(c+d)(a+c)(b+d)$ still remains 13.73 (This of course will not happen always). Hence our conclusion remains unchanged even after applying Yates' correction.

Note: The determination of number of degrees of freedom associated with the test statistic is important. The general rule is - 'The degrees of freedom associated with a contingency table possessing r rows and c columns will always equal (r-1)*(c-1). For example if table has 2 rows and 3 columns then degrees of freedom will be (2-1)*(3-1) = 2; if table has 3 rows and 4 columns, then degrees of freedom will be 6 and so on. We will illustrate the procedure for r rows and c columns with the help of an illustrative example

Case 2: r x s contingency table

Example11.7: A survey was conducted in tribal belt to evaluate the effectiveness of DDT. A survey of 450 tribal hamlets revealed the following data (Table 11.13). Use the data to verify the claim that spraying of DDT enhances the chances of malaria cases.

Table 11.13: Observed frequencies

DDT→	Not used at all	Used in huts only	Used in Cow sheds only	Total
Malaria + ve	80	100	160	275
No Malaria	70	20	20	175
Total	150	120	180	450

H_o: Incidence of malaria and spraying of DDT are independent.

Table 11.14: Expected frequencies

DDT→	Not used at all	Used in huts only	Used in Cow sheds only	Total
Malaria + ve	113.33	90.67	136	275
No Malaria	36.67	29.33	44	175
Total	150	120	180	450

Here the test statistic is $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right)$ and it is equal to

$$\chi^2 = (80 - 113.33)^2 / 113.33 + (100 - 90.67)^2 / 90.67 + \ldots + (20 - 44)^2 / 44 = 61.36$$

Here d. f = (2-1)*(3-1) = (no of rows -1) * (no. of columns -1) = 2 and here we use 5% l. o. s. See the table and note that $\chi^2_{2,0.05}$ = 5.992. It is less than observed value of chi-square statistic (= 61.36). We therefore reject H_0 at 5% l. o. s. and conclude that incidence of malaria in the tribal belt and spraying of DDT in these localities is associated. Is this result surprising? If DDT is not sprayed at all then about 53% of tribal people have suffered from malarial infection; if DDT is used in huts then about 83% cases from this tribal group are recorded and lastly if DDT is sprayed in cow sheds only then this proportion reaches the score of 88% of malarial cases. DDT is meant for protecting against malaria, but the results are contrary. Why this should happen? Think about it. Rejection or acceptance of null hypothesis is not the end of the analysis. It opens a new set of problems and researcher has to look for further details and investigate for possible reasons that led to the conclusion.

11.7 Self Test

1. Multiple choice questions:

(i) What is the difference between a parameter and a statistic?

a. We estimate a statistic but not a parameter

b. Parameter can have a sampling distribution while statistic can not.

c. Parameter is a (possibly unknown) constant while statistic is a random variable.

d. Parameter can have a standard error while statistic cannot.

(ii) For a sample of 5 observations $\Sigma Y = 10$ and $\Sigma Y^2 = 100$. Hence sample variance is ---

a. 1 **b.** 16 **c.** 20 **d.** 98

(iii) Significance level of a test of hypothesis is ---

a. Selected before data are analyzed

b. Usually 95%

c. A function of data 20

d. Probability of accepting a hypothesis when in fact it is false

(iv) I want to test whether males are always taller than females. I have a sample data on 40 males and females from the given population. The appropriate statistical test to test the null hypothesis would be ---

a. z- test **b.** Test for equality of proportions

c. Paired t test **d.** t-test based in independent samples

2. In **Q. 1 (iv)** above, if you are told that these are the data collected on married couples, would you like to reanalyze the data? Justify your answer.

3. Following are the yields of 10 plants in a uniformity trial. Using these data verify whether the mean yield is 200 gm.

X (weight in gm)	225	181	190	230	176	235	318	234	200

4. A tyre manufacturer claims that the tyre has an average life of 60,000 km. In a sample study of 16 cars using the branded tyres it was revealed that sample mean is 58,116 km with a sample s. d. of 8000 km. Can you support the claim? (use 5 % l. o. s.)

5. The serum albumin content in the blood was estimated using two different methods in the same laboratory. For 8 paired sample units the observed differences (in gm per 100 ml) were as follows:

0.61 0.73 0.82 0.91 0.30 0.55 0.50 -0.51

Test H_o that the population mean of these differences is zero.

6. Following are the data on number of items produced per hour by machines M_1 and M_2.

Table 11.15: Data on # items/machine

M_1	48	46	56	56	49	48	57	55
M_2	52	50	52	54	53	51	52	55

Check whether two machines have same variability? Also test whether machines differ in mean production?

7. It is common experience that students are unhappy with the present examination system. A small survey was conducted. I the questionnaire following two questions were asked:

Q.1: Has student passed the examination in the first attempt?

Q.2: Is he (she) satisfied with the present system?

Table 11.16: Table of observed frequencies

	Boys		Girls	
Q2→ / Q1↓	Yes	No	Yes	No
Yes	22	13	30	20
No	10	15	35	22

Examine whether the responses to these questions are associated? Carry out appropriate test separately for boys, girls and for pooled data (boys and girls together).

8. Psychologists believe that level of education and MAS (Marriage adjustment score) are highly associated. Using following data test the claim.

Table 11.17: Marriage adjustment score

Level of Education	MAS (Marriage Adjustment Score)				
	Very low	Low	Average	High	Very high
Primary l	24	32	35	40	20
H. S. C.	29	30	38	42	30
Graduate or above	30	10	25	20	10

9. A person tossed a die 100 times and recorded following data. Using it verify whether the die can be regarded as a fair?

Table 11.18: Experiment of tossing of a die

Outcome	1	2	3	4	5	6
Frequency	16	18	20	12	18	16

10. In a socio-economic study a researcher prepared a questionnaire and examined whether the respondent is bias towards either sex in his daily behavior. Accordingly she awarded scores, called as bias-score to the respondent. Smaller the score, less is the bias of the individual. Following are the data on "bias score" of individuals as revealed from a sample study. Are these scores normally distributed?

Table 11.19: Frequency distribution of Bias-score

Bias score	0 - 10	10 – 20	20 - 30	30 - 40	40 - 50	50 - 60
Observed frequency	15	25	66	72	30	12

(**Hint:** first obtain mean and sample variance of the above data, use them to estimate population parameters, obtain expected frequencies, use chi-square test for goodness of fit and then conclude).

Also test the hypothesis that the population mean of the Bias- score is 30 against the alternative that it is more than 30. Use 5% l. o. s.

11. Following are the data on murders in a district in 12 consecutive months.

Table 11.20: Monthwise murders in a district

Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
8	19	8	6	3	8	11	12	16	8	7	5

Test the null hypothesis that the probability of a murder is the same in each month.

12. Following are the data on time required (in seconds) for the PC to restart the machine:

100 90 95 97 85 95 99 89

Test Ho: Population variance = 25 units. Also test that population mean is 90 sec.

13. A student performs the experiment of tossing a coin 100 times. He tosses the coin with left hand and observes 65 heads. He tossed the same coin with right hand and observed 56 heads. Can you say that the coin he used was a fair one? The student claims that he always gets more heads when he tosses the coin with right hand than left. Can you support his claim?

14. There are five different roads that lead to university administration building from the entrance of university gate. On a particular day a student recorded the number of vehicles on these roads during peak hours.

Table 11.21: Number of vehicles on different roads

Road no	A	B	C	D	E
# vehicles	230	280	270	245	275

Can you support the claim that certain roads are preferred more compared to others? Use 1% l. o. s..

15. In the university students opt for university canteen for their food. In order to maintain the quality of food it is a practice to conduct small sample surveys and get primary data from students themselves. In one such survey, students were asked if service by the university canteen was satisfactory or not. Then, there was dispute between students and the contractor. After this incidence the students were interviewed again. Can you test the hypothesis that the opinion remains unchanged in the same population.

Table 11.22: Results of opinion poll

	Satisfied	Unsatisfied	Total
Before dispute	80	20	100
After dispute	72	28	100

Hint: you can't use this if you are asking the same people! Not independent! Better way to arrange:

	Satisfied after	Not satisfied	Total
Originally satisfied	70	10	80
Unsatisfied	2	18	20
Total	72	28	100

16. It is claimed that average (BMR) basal metabolic rate (per kg of body weight), of males is usually more than females of the same age-group. Using following data check the claim at 5% l. o. s.

Table 11.23 : Sex-wise BMR (age-group 30-35 years)

Males	40	42	46	44	40	48	55
Females	42	38	44	50	46	38	

17. Describe statistical test for testing equality of population means for

(i) Independent samples and (ii) Dependent samples.

18. How you will perform chi-square test for testing that normal distribution is appropriate model for a given data set? Describe test procedure with the help of an example.

19. A chemical balance connected to PC is guaranteed to report the weight accurately within 2 units. Observations are recorded on the same unit by 2 different technicians. These are as follows:

Technician 1:	300.25	300.65	300.45	300.15
Technician 2:	300.65	300.45	300.10	300.20

Test the hypothesis that population variance is unity. You may pool the data for both technicians together. Can you say that technicians do not differ on an average and are consistent? Use 1% l. o. s.

Table 1 : Cumulative Standard Normal Distribution*

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.40	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.30	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.20	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.10	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.00	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.20	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.00	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.60	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.00	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

* $p[z \leq z]$ = area in the table = $\phi(z)$

Table 2 : Tables of Binomial Probability Sums $P[x \leq k] = \sum_{x=0}^{K} b(x; N, P)$

N = 2

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01
1	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

N = 3

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.729	0.512	0.343	0.216	0.125	0.064	0.027	0.008	0.001
1	0.972	0.896	0.784	0.648	0.500	0.352	0.216	0.104	0.028
2	0.999	0.992	0.973	0.936	0.875	0.784	0.657	0.488	0.271
3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

N = 4

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.6561	0.4096	0.2401	0.1296	0.0625	0.0256	0.0081	0.0016	0.0001
1	0.9477	0.8192	0.6517	0.4752	0.3125	0.1792	0.0837	0.0272	0.0037
2	0.9963	0.9728	0.9163	0.8208	0.6875	0.5248	0.3483	0.1808	0.0523
3	0.9999	0.9984	0.9919	0.9744	0.9375	0.8704	0.7599	0.5904	0.3439
4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

N = 5

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.59049	0.32768	0.16807	0.07776	0.03125	0.01024	0.00243	0.00032	0.00001
1	0.91854	0.73728	0.52822	0.33696	0.18750	0.08704	0.03078	0.00672	0.00046
2	0.99144	0.94208	0.83692	0.68256	0.50000	0.31744	0.16308	0.05792	0.00856
3	0.99954	0.99328	0.96922	0.91296	0.81250	0.66304	0.47178	0.26272	0.08146
4	0.99999	0.99968	0.99757	0.98976	0.96875	0.92224	0.83193	0.67232	0.40951
5	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N = 6

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.53144	0.26214	0.11765	0.04666	0.01562	0.00410	0.00073	0.00006	0.00000
1	0.88574	0.65536	0.42018	0.23328	0.10938	0.04096	0.01094	0.00160	0.00006
2	0.98415	0.90112	0.74431	0.54432	0.34375	0.17920	0.07047	0.01696	0.00127
3	0.99873	0.98304	0.92953	0.82080	0.65625	0.45568	0.25569	0.09888	0.01585
4	0.99994	0.99840	0.98906	0.95904	0.89062	0.76672	0.57982	0.34464	0.11426
5	1.00000	0.99994	0.99927	0.99590	0.98438	0.95334	0.88235	0.73786	0.46856
6	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N = 7

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.47830	0.20972	0.08235	0.02799	0.00781	0.00164	0.00022	0.00001	0.00000
1	0.85031	0.57672	0.32942	0.15863	0.06250	0.01884	0.00379	0.00037	0.00001
2	0.97431	0.85197	0.64707	0.41990	0.22656	0.09626	0.02880	0.00467	0.00018
3	0.99727	0.96666	0.87396	0.71021	0.50000	0.28979	0.12604	0.03334	0.00273
4	0.99982	0.99533	0.97120	0.90374	0.77344	0.58010	0.35293	0.14803	0.02569
5	0.99999	0.99963	0.99621	0.98116	0.93750	0.84137	0.67058	0.42328	0.14969
6	1.00000	0.99999	0.99978	0.99836	0.99219	0.97201	0.91765	0.79028	0.52170
7	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N = 8

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.43047	0.16777	0.05765	0.01680	0.00391	0.00066	0.00007	0.00000	0.00000
1	0.81310	0.50332	0.25530	0.10638	0.03516	0.00852	0.00129	0.00008	0.00000
2	0.96191	0.79692	0.55177	0.31539	0.14453	0.04981	0.01129	0.00123	0.00002
3	0.99498	0.94372	0.80590	0.59409	0.36328	0.17367	0.05797	0.01041	0.00043
4	0.99957	0.98959	0.94203	0.82633	0.63672	0.40591	0.19410	0.05628	0.00502
5	0.99998	0.99877	0.98871	0.95019	0.85547	0.68461	0.44823	0.20308	0.03809
6	1.00000	0.99992	0.99871	0.99148	0.96484	0.89362	0.74470	0.49668	0.18690
7	1.00000	1.00000	0.99993	0.99934	0.99609	0.98320	0.94235	0.83223	0.56953
8	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N = 9

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.38742	0.13422	0.04035	0.01008	0.00195	0.00026	0.00002	0.00000	0.00000
1	0.77484	0.43621	0.19600	0.07054	0.01953	0.00380	0.00043	0.00002	0.00000
2	0.94703	0.73820	0.46283	0.23179	0.08984	0.02503	0.00429	0.00031	0.00000
3	0.99167	0.91436	0.72966	0.48261	0.25391	0.09935	0.02529	0.00307	0.00006
4	0.99911	0.98042	0.90119	0.73343	0.50000	0.26657	0.09881	0.01958	0.00089
5	0.99994	0.99693	0.97471	0.90065	0.74609	0.51739	0.27034	0.08564	0.00833
6	1.00000	0.99969	0.99571	0.97497	0.91016	0.76821	0.53717	0.26180	0.05297
7	1.00000	0.99998	0.99957	0.99620	0.98047	0.92946	0.80400	0.56379	0.22516
8	1.00000	1.00000	0.99998	0.99974	0.99805	0.98992	0.95965	0.86578	0.61258
9	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =10

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.34868	0.10737	0.02825	0.00605	0.00098	0.00010	0.00001	0.00000	0.00000
1	0.73610	0.37581	0.14931	0.04636	0.01074	0.00168	0.00014	0.00000	0.00000
2	0.92981	0.67780	0.38278	0.16729	0.05469	0.01229	0.00159	0.00008	0.00000
3	0.98720	0.87913	0.64961	0.38228	0.17188	0.05476	0.01059	0.00086	0.00001
4	0.99837	0.96721	0.84973	0.63310	0.37695	0.16624	0.04735	0.00637	0.00015
5	0.99985	0.99363	0.95265	0.83376	0.62305	0.36690	0.15027	0.03279	0.00163
6	0.99999	0.99914	0.98941	0.94524	0.82812	0.61772	0.35039	0.12087	0.01280
7	1.00000	0.99992	0.99841	0.98771	0.94531	0.83271	0.61722	0.32220	0.07019
8	1.00000	1.00000	0.99986	0.99832	0.98926	0.95364	0.85069	0.62419	0.26390
9	1.00000	1.00000	0.99999	0.99990	0.99902	0.99395	0.97175	0.89263	0.65132
10	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =11

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.31381	0.08590	0.01977	0.00363	0.00049	0.00004	0.00000	0.00000	0.00000
1	0.69736	0.32212	0.11299	0.03023	0.00586	0.00073	0.00005	0.00000	0.00000
2	0.91044	0.61740	0.31274	0.11892	0.03271	0.00592	0.00058	0.00002	0.00000
3	0.98147	0.83886	0.56956	0.29628	0.11328	0.02928	0.00429	0.00024	0.00000
4	0.99725	0.94959	0.78970	0.53277	0.27441	0.09935	0.02162	0.00197	0.00002
5	0.99970	0.98835	0.92178	0.75350	0.50000	0.24650	0.07822	0.01165	0.00030
6	0.99998	0.99803	0.97838	0.90065	0.72559	0.46723	0.21030	0.05041	0.00275
7	1.00000	0.99976	0.99571	0.97072	0.88672	0.70372	0.43044	0.16114	0.01853
8	1.00000	0.99998	0.99942	0.99408	0.96729	0.88108	0.68726	0.38260	0.08956
9	1.00000	1.00000	0.99995	0.99927	0.99414	0.96977	0.88701	0.67788	0.30264
10	1.00000	1.00000	1.00000	0.99996	0.99951	0.99637	0.98023	0.91410	0.68619
11	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =12

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.28243	0.06872	0.01384	0.00218	0.00024	0.00002	0.00000	0.00000	0.00000
1	0.65900	0.27488	0.08503	0.01959	0.00317	0.00032	0.00002	0.00000	0.00000
2	0.88913	0.55835	0.25282	0.08344	0.01929	0.00281	0.00021	0.00000	0.00000
3	0.97436	0.79457	0.49252	0.22534	0.07300	0.01527	0.00169	0.00006	0.00000
4	0.99567	0.92744	0.72366	0.43818	0.19385	0.05731	0.00949	0.00058	0.00000
5	0.99946	0.98059	0.88215	0.66521	0.38721	0.15821	0.03860	0.00390	0.00005
6	0.99995	0.99610	0.96140	0.84179	0.61279	0.33479	0.11785	0.01941	0.00054
7	1.00000	0.99942	0.99051	0.94269	0.80615	0.56182	0.27634	0.07256	0.00433
8	1.00000	0.99994	0.99831	0.98473	0.92700	0.77466	0.50748	0.20543	0.02564
9	1.00000	1.00000	0.99979	0.99719	0.98071	0.91656	0.74718	0.44165	0.11087
10	1.00000	1.00000	0.99998	0.99968	0.99683	0.98041	0.91497	0.72512	0.34100
11	1.00000	1.00000	1.00000	0.99998	0.99976	0.99782	0.98616	0.93128	0.71757
12	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =13

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.25419	0.05498	0.00969	0.00131	0.00012	0.00001	0.00000	0.00000	0.00000
1	0.62134	0.23365	0.06367	0.01263	0.00171	0.00014	0.00000	0.00000	0.00000
2	0.86612	0.50165	0.20248	0.05790	0.01123	0.00132	0.00007	0.00000	0.00000
3	0.96584	0.74732	0.42061	0.16858	0.04614	0.00779	0.00065	0.00002	0.00000
4	0.99354	0.90087	0.65431	0.35304	0.13342	0.03208	0.00403	0.00017	0.00000
5	0.99908	0.96996	0.83460	0.57440	0.29053	0.09767	0.01822	0.00125	0.00001
6	0.99990	0.99300	0.93762	0.77116	0.50000	0.22884	0.06238	0.00700	0.00010
7	0.99999	0.99875	0.98178	0.90233	0.70947	0.42560	0.16540	0.03004	0.00092
8	1.00000	0.99983	0.99597	0.96792	0.86658	0.64696	0.34569	0.09913	0.00646
9	1.00000	0.99998	0.99935	0.99221	0.95386	0.83142	0.57939	0.25268	0.03416
10	1.00000	1.00000	0.99993	0.99868	0.98877	0.94210	0.79752	0.49835	0.13388
11	1.00000	1.00000	1.00000	0.99986	0.99829	0.98737	0.93633	0.76635	0.37866
12	1.00000	1.00000	1.00000	0.99999	0.99988	0.99869	0.99031	0.94502	0.74581
13	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =14

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.22877	0.04398	0.00678	0.00078	0.00006	0.00000	0.00000	0.00000	0.00000
1	0.58463	0.19791	0.04748	0.00810	0.00092	0.00006	0.00000	0.00000	0.00000
2	0.84164	0.44805	0.16084	0.03979	0.00647	0.00061	0.00003	0.00000	0.00000
3	0.95587	0.69819	0.35517	0.12431	0.02869	0.00391	0.00025	0.00000	0.00000
4	0.99077	0.87016	0.58420	0.27926	0.08978	0.01751	0.00167	0.00005	0.00000
5	0.99853	0.95615	0.78052	0.48585	0.21198	0.05832	0.00829	0.00038	0.00000
6	0.99982	0.98839	0.90672	0.69245	0.39526	0.15014	0.03147	0.00240	0.00002
7	0.99998	0.99760	0.96853	0.84986	0.60474	0.30755	0.09328	0.01161	0.00018
8	1.00000	0.99962	0.99171	0.94168	0.78802	0.51415	0.21948	0.04385	0.00147
9	1.00000	0.99995	0.99833	0.98249	0.91022	0.72074	0.41580	0.12984	0.00923
10	1.00000	1.00000	0.99975	0.99609	0.97131	0.87569	0.64483	0.30181	0.04413
11	1.00000	1.00000	0.99997	0.99939	0.99353	0.96021	0.83916	0.55195	0.15836
12	1.00000	1.00000	1.00000	0.99994	0.99908	0.99190	0.95252	0.80209	0.41537
13	1.00000	1.00000	1.00000	1.00000	0.99994	0.99922	0.99322	0.95602	0.77123
14	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =15

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.20589	0.03518	0.00475	0.00047	0.00003	0.00000	0.00000	0.00000	0.00000
1	0.54904	0.16713	0.03527	0.00517	0.00049	0.00003	0.00000	0.00000	0.00000
2	0.81594	0.39802	0.12683	0.02711	0.00369	0.00028	0.00001	0.00000	0.00000
3	0.94444	0.64816	0.29687	0.09050	0.01758	0.00193	0.00009	0.00000	0.00000
4	0.98728	0.83577	0.51549	0.21728	0.05923	0.00935	0.00067	0.00001	0.00000
5	0.99775	0.93895	0.72162	0.40322	0.15088	0.03383	0.00365	0.00011	0.00000
6	0.99969	0.98194	0.86886	0.60981	0.30362	0.09505	0.01524	0.00078	0.00000
7	0.99997	0.99576	0.94999	0.78690	0.50000	0.21310	0.05001	0.00424	0.00003
8	1.00000	0.99922	0.98476	0.90495	0.69638	0.39019	0.13114	0.01806	0.00031
9	1.00000	0.99989	0.99635	0.96617	0.84912	0.59678	0.27838	0.06105	0.00225
10	1.00000	0.99999	0.99933	0.99065	0.94077	0.78272	0.48451	0.16423	0.01272
11	1.00000	1.00000	0.99991	0.99807	0.98242	0.90950	0.70313	0.35184	0.05556
12	1.00000	1.00000	0.99999	0.99972	0.99631	0.97289	0.87317	0.60198	0.18406
13	1.00000	1.00000	1.00000	0.99997	0.99951	0.99483	0.96473	0.83287	0.45096
14	1.00000	1.00000	1.00000	1.00000	0.99997	0.99953	0.99525	0.96482	0.79411
15	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =16

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.18530	0.02815	0.00332	0.00028	0.00002	0.00000	0.00000	0.00000	0.00000
1	0.51473	0.14074	0.02611	0.00329	0.00026	0.00001	0.00000	0.00000	0.00000
2	0.78925	0.35184	0.09936	0.01834	0.00209	0.00013	0.00000	0.00000	0.00000
3	0.93159	0.59813	0.24586	0.06515	0.01064	0.00094	0.00003	0.00000	0.00000
4	0.98300	0.79825	0.44990	0.16657	0.03841	0.00490	0.00027	0.00000	0.00000
5	0.99670	0.91831	0.65978	0.32884	0.10506	0.01914	0.00157	0.00003	0.00000
6	0.99950	0.97334	0.82469	0.52717	0.22725	0.05832	0.00713	0.00025	0.00000
7	0.99994	0.99300	0.92565	0.71606	0.40181	0.14227	0.02567	0.00148	0.00001
8	0.99999	0.99852	0.97433	0.85773	0.59819	0.28394	0.07435	0.00700	0.00006
9	1.00000	0.99975	0.99287	0.94168	0.77275	0.47283	0.17531	0.02666	0.00050
10	1.00000	0.99997	0.99843	0.98086	0.89494	0.67116	0.34022	0.08169	0.00330
11	1.00000	1.00000	0.99973	0.99510	0.96159	0.83343	0.55010	0.20175	0.01700
12	1.00000	1.00000	0.99997	0.99906	0.98936	0.93485	0.75414	0.40187	0.06841
13	1.00000	1.00000	1.00000	0.99987	0.99791	0.98166	0.90064	0.64816	0.21075
14	1.00000	1.00000	1.00000	0.99999	0.99974	0.99671	0.97389	0.85926	0.48527
15	1.00000	1.00000	1.00000	1.00000	0.99998	0.99972	0.99668	0.97185	0.81470
16	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =17

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.16677	0.02252	0.00233	0.00017	0.00001	0.00000	0.00000	0.00000	0.00000
1	0.48179	0.11822	0.01928	0.00209	0.00014	0.00000	0.00000	0.00000	0.00000
2	0.76180	0.30962	0.07739	0.01232	0.00117	0.00006	0.00000	0.00000	0.00000
3	0.91736	0.54888	0.20191	0.04642	0.00636	0.00045	0.00001	0.00000	0.00000
4	0.97786	0.75822	0.38869	0.12600	0.02452	0.00252	0.00010	0.00000	0.00000
5	0.99533	0.89430	0.59682	0.26393	0.07173	0.01059	0.00066	0.00001	0.00000
6	0.99922	0.96234	0.77522	0.44784	0.16615	0.03481	0.00324	0.00008	0.00000
7	0.99989	0.98907	0.89536	0.64051	0.31453	0.09190	0.01269	0.00049	0.00000
8	0.99999	0.99742	0.95972	0.80106	0.50000	0.19894	0.04028	0.00258	0.00001
9	1.00000	0.99951	0.98731	0.90810	0.68547	0.35949	0.10464	0.01093	0.00011
10	1.00000	0.99992	0.99676	0.96519	0.83385	0.55216	0.22478	0.03766	0.00078
11	1.00000	0.99999	0.99934	0.98941	0.92827	0.73607	0.40318	0.10570	0.00467
12	1.00000	1.00000	0.99990	0.99748	0.97548	0.87400	0.61131	0.24178	0.02214
13	1.00000	1.00000	0.99999	0.99955	0.99364	0.95358	0.79809	0.45112	0.08264
14	1.00000	1.00000	1.00000	0.99994	0.99883	0.98768	0.92261	0.69038	0.23820
15	1.00000	1.00000	1.00000	1.00000	0.99986	0.99791	0.98072	0.88178	0.51821
16	1.00000	1.00000	1.00000	1.00000	0.99999	0.99983	0.99767	0.97748	0.83323
17	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =18

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.15009	0.01801	0.00163	0.00010	0.00000	0.00000	0.00000	0.00000	0.00000
1	0.45028	0.09908	0.01419	0.00132	0.00007	0.00000	0.00000	0.00000	0.00000
2	0.73380	0.27134	0.05995	0.00823	0.00066	0.00003	0.00000	0.00000	0.00000
3	0.90180	0.50103	0.16455	0.03278	0.00377	0.00021	0.00000	0.00000	0.00000
4	0.97181	0.71635	0.33265	0.09417	0.01544	0.00128	0.00004	0.00000	0.00000
5	0.99358	0.86708	0.53438	0.20876	0.04813	0.00575	0.00027	0.00000	0.00000
6	0.99883	0.94873	0.72170	0.37428	0.11894	0.02028	0.00143	0.00002	0.00000
7	0.99983	0.98372	0.85932	0.56344	0.24034	0.05765	0.00607	0.00016	0.00000
8	0.99998	0.99575	0.94041	0.73684	0.40726	0.13471	0.02097	0.00091	0.00000
9	1.00000	0.99909	0.97903	0.86529	0.59274	0.26316	0.05959	0.00425	0.00002
10	1.00000	0.99984	0.99393	0.94235	0.75966	0.43656	0.14068	0.01628	0.00017
11	1.00000	0.99998	0.99857	0.97972	0.88106	0.62572	0.27830	0.05127	0.00117
12	1.00000	1.00000	0.99973	0.99425	0.95187	0.79124	0.46562	0.13292	0.00642
13	1.00000	1.00000	0.99996	0.99872	0.98456	0.90583	0.66735	0.28365	0.02819
14	1.00000	1.00000	1.00000	0.99979	0.99623	0.96722	0.83545	0.49897	0.09820
15	1.00000	1.00000	1.00000	0.99997	0.99934	0.99177	0.94005	0.72866	0.26620
16	1.00000	1.00000	1.00000	1.00000	0.99993	0.99868	0.98581	0.90092	0.54972
17	1.00000	1.00000	1.00000	1.00000	1.00000	0.99990	0.99837	0.98199	0.84991
18	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =19

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.13509	0.01441	0.00114	0.00006	0.00000	0.00000	0.00000	0.00000	0.00000
1	0.42026	0.08287	0.01042	0.00083	0.00004	0.00000	0.00000	0.00000	0.00000
2	0.70544	0.23689	0.04622	0.00546	0.00036	0.00001	0.00000	0.00000	0.00000
3	0.88500	0.45509	0.13317	0.02296	0.00221	0.00010	0.00000	0.00000	0.00000
4	0.96481	0.67329	0.28222	0.06961	0.00961	0.00064	0.00001	0.00000	0.00000
5	0.99141	0.83694	0.47386	0.16292	0.03178	0.00307	0.00011	0.00000	0.00000
6	0.99830	0.93240	0.66550	0.30807	0.08353	0.01156	0.00062	0.00001	0.00000
7	0.99973	0.97672	0.81803	0.48778	0.17964	0.03523	0.00282	0.00005	0.00000
8	0.99996	0.99334	0.91608	0.66748	0.32380	0.08847	0.01054	0.00031	0.00000
9	1.00000	0.99842	0.96745	0.81391	0.50000	0.18609	0.03255	0.00158	0.00000
10	1.00000	0.99969	0.98946	0.91153	0.67620	0.33252	0.08392	0.00666	0.00004
11	1.00000	0.99995	0.99718	0.96477	0.82036	0.51222	0.18197	0.02328	0.00027
12	1.00000	0.99999	0.99938	0.98844	0.91647	0.69193	0.33450	0.06760	0.00170
13	1.00000	1.00000	0.99989	0.99693	0.96822	0.83708	0.52614	0.16306	0.00859
14	1.00000	1.00000	0.99999	0.99936	0.99039	0.93039	0.71778	0.32671	0.03519
15	1.00000	1.00000	1.00000	0.99990	0.99779	0.97704	0.86683	0.54491	0.11500
16	1.00000	1.00000	1.00000	0.99999	0.99964	0.99454	0.95378	0.76311	0.29456
17	1.00000	1.00000	1.00000	1.00000	0.99996	0.99917	0.98958	0.91713	0.57974
18	1.00000	1.00000	1.00000	1.00000	1.00000	0.99994	0.99886	0.98559	0.86491
19	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

N =20

K \ P=	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0.12158	0.01153	0.00080	0.00004	0.00000	0.00000	0.00000	0.00000	0.00000
1	0.39175	0.06918	0.00764	0.00052	0.00002	0.00000	0.00000	0.00000	0.00000
2	0.67693	0.20608	0.03548	0.00361	0.00020	0.00001	0.00000	0.00000	0.00000
3	0.86705	0.41145	0.10709	0.01596	0.00129	0.00005	0.00000	0.00000	0.00000
4	0.95683	0.62965	0.23751	0.05095	0.00591	0.00032	0.00001	0.00000	0.00000
5	0.98875	0.80421	0.41637	0.12560	0.02069	0.00161	0.00004	0.00000	0.00000
6	0.99761	0.91331	0.60801	0.25001	0.05766	0.00647	0.00026	0.00000	0.00000
7	0.99958	0.96786	0.77227	0.41589	0.13159	0.02103	0.00128	0.00002	0.00000
8	0.99994	0.99002	0.88667	0.59560	0.25172	0.05653	0.00514	0.00010	0.00000
9	0.99999	0.99741	0.95204	0.75534	0.41190	0.12752	0.01714	0.00056	0.00000
10	1.00000	0.99944	0.98286	0.87248	0.58810	0.24466	0.04796	0.00259	0.00001
11	1.00000	0.99990	0.99486	0.94347	0.74828	0.40440	0.11333	0.00998	0.00006
12	1.00000	0.99998	0.99872	0.97897	0.86841	0.58411	0.22773	0.03214	0.00042
13	1.00000	1.00000	0.99974	0.99353	0.94234	0.74999	0.39199	0.08669	0.00239
14	1.00000	1.00000	0.99996	0.99839	0.97931	0.87440	0.58363	0.19579	0.01125
15	1.00000	1.00000	0.99999	0.99968	0.99409	0.94905	0.76249	0.37035	0.04317
16	1.00000	1.00000	1.00000	0.99995	0.99871	0.98404	0.89291	0.58855	0.13295
17	1.00000	1.00000	1.00000	0.99999	0.99980	0.99639	0.96452	0.79392	0.32307
18	1.00000	1.00000	1.00000	1.00000	0.99998	0.99948	0.99236	0.93082	0.60825
19	1.00000	1.00000	1.00000	1.00000	1.00000	0.99996	0.99920	0.98847	0.87842
20	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Table 3 : Cumulative Paisson Distribution Tables p[x ≤ x] = Mean

X	0.01	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.990	0.951	0.905	0.819	0.741	0.670	0.607	0.549	0.497	0.449	0.407
1	1.000	0.999	0.995	0.982	0.963	0.938	0.910	0.878	0.844	0.809	0.772
2		1.000	1.000	0.999	0.996	0.992	0.986	0.977	0.966	0.953	0.937
3				1.000	1.000	0.999	0.998	0.997	0.994	0.991	0.987
4						1.000	1.000	1.000	0.999	0.999	0.998
5									1.000	1.000	1.000
X	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.368	0.333	0.301	0.273	0.247	0.223	0.202	0.183	0.165	0.150	0.135
1	0.736	0.699	0.663	0.627	0.592	0.558	0.525	0.493	0.463	0.434	0.406
2	0.920	0.900	0.879	0.857	0.833	0.809	0.783	0.757	0.731	0.704	0.677
3	0.981	0.974	0.966	0.957	0.946	0.934	0.921	0.907	0.891	0.875	0.857
4	0.996	0.995	0.992	0.989	0.986	0.981	0.976	0.970	0.964	0.956	0.947
5	0.999	0.999	0.998	0.998	0.997	0.996	0.994	0.992	0.990	0.987	0.983
6	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.997	0.995
7					1.000	1.000	1.000	1.000	0.999	0.999	0.999
8									1.000	1.000	1.000
X	2.2	2.4	2.6	2.8	3.0	3.5	4.0	4.5	5.0	5.5	6.0
0	0.111	0.091	0.074	0.061	0.050	0.030	0.018	0.011	0.007	0.004	0.002
1	0.355	0.308	0.267	0.231	0.199	0.136	0.092	0.061	0.040	0.027	0.017
2	0.623	0.570	0.518	0.469	0.423	0.321	0.238	0.174	0.125	0.088	0.062
3	0.819	0.779	0.736	0.692	0.647	0.537	0.433	0.342	0.265	0.202	0.151
4	0.928	0.904	0.877	0.848	0.815	0.725	0.629	0.532	0.440	0.358	0.285
5	0.975	0.964	0.951	0.935	0.916	0.858	0.785	0.703	0.616	0.529	0.446
6	0.993	0.988	0.983	0.976	0.966	0.935	0.889	0.831	0.762	0.686	0.606
7	0.998	0.997	0.995	0.992	0.988	0.973	0.949	0.913	0.867	0.809	0.744
8	1.000	0.999	0.999	0.998	0.996	0.990	0.979	0.960	0.932	0.894	0.847
9		1.000	1.000	0.999	0.999	0.997	0.992	0.983	0.968	0.946	0.916
10				1.000	1.000	0.999	0.997	0.993	0.986	0.975	0.957
11						1.000	0.999	0.998	0.995	0.989	0.980
12							1.000	0.999	0.998	0.996	0.991
13								1.000	0.999	0.998	0.996
14									1.000	0.999	0.999
15										1.000	0.999
16											1.000

X	6.5	7.0	7.5	8.0	9.0	10.0	12.0	14.0	16.0	18.0	20.0
0	0.002	0.001	0.001	0.000							
1	0.011	0.007	0.005	0.003	0.001						
2	0.043	0.030	0.020	0.014	0.006	0.003	0.001				
3	0.112	0.082	0.059	0.042	0.021	0.010	0.002				
4	0.224	0.173	0.132	0.100	0.055	0.029	0.008	0.002			
5	0.369	0.301	0.241	0.191	0.116	0.067	0.020	0.006	0.001		

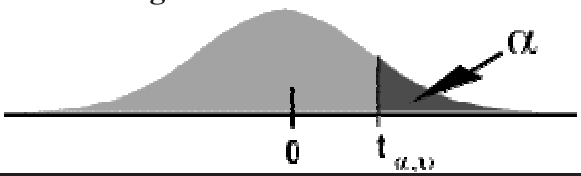
X	6.5	7.0	7.5	8.0	9.0	10.0	12.0	14.0	16.0	18.0	20.0
6	0.527	0.450	0.378	0.313	0.207	0.130	0.046	0.014	0.004	0.001	
7	0.673	0.599	0.525	0.453	0.324	0.220	0.090	0.032	0.010	0.003	0.001
8	0.792	0.729	0.662	0.593	0.456	0.333	0.155	0.062	0.022	0.007	0.002
9	0.877	0.830	0.776	0.717	0.587	0.458	0.242	0.109	0.043	0.015	0.005
10	0.933	0.901	0.862	0.816	0.706	0.583	0.347	0.176	0.077	0.030	0.011
11	0.966	0.947	0.921	0.888	0.803	0.697	0.462	0.260	0.127	0.055	0.021
12	0.984	0.973	0.957	0.936	0.876	0.792	0.576	0.358	0.193	0.092	0.039
13	0.993	0.987	0.978	0.966	0.926	0.864	0.682	0.464	0.275	0.143	0.066
14	0.997	0.994	0.990	0.983	0.959	0.917	0.772	0.570	0.368	0.208	0.105
15	0.999	0.998	0.995	0.992	0.978	0.951	0.844	0.669	0.467	0.287	0.157
16	1.000	0.999	0.998	0.996	0.989	0.973	0.899	0.756	0.566	0.375	0.221
17		1.000	0.999	0.998	0.995	0.986	0.937	0.827	0.659	0.469	0.297
18			1.000	0.999	0.998	0.993	0.963	0.883	0.742	0.562	0.381
19				1.000	0.999	0.997	0.979	0.923	0.812	0.651	0.470
20					1.000	0.998	0.988	0.952	0.868	0.731	0.559
21						0.999	0.994	0.971	0.911	0.799	0.644
22						1.000	0.997	0.983	0.942	0.855	0.721
23							0.999	0.991	0.963	0.899	0.787
24							0.999	0.995	0.978	0.932	0.843
25							1.000	0.997	0.987	0.955	0.888
26								0.999	0.993	0.972	0.922
27								0.999	0.996	0.983	0.948
28								1.000	0.998	0.990	0.966
29									0.999	0.994	0.978
30									0.999	0.997	0.987
31									1.000	0.998	0.992
32										0.999	0.995
33										1.000	0.997
34											0.999
35											0.999
36											1.000

Table 4 : Chi-Squared Values for a Specified Right Tail Area (α = Right Hand Tail Area)

ν	0.999	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005	0.001
1	0.00	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88	10.83
2	0.00	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60	13.82
3	0.02	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75	20.51
6	0.38	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55	22.46
7	0.60	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95	26.12
9	1.15	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59	27.88
10	1.48	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19	29.59
11	1.83	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76	31.26
12	2.21	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30	32.91
13	2.62	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82	34.53
14	3.04	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32	36.12
15	3.48	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80	37.70
16	3.94	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27	39.25
17	4.42	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72	40.79
18	4.90	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16	42.31
19	5.41	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58	43.82
20	5.92	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00	45.31
21	6.45	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40	46.80
22	6.98	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80	48.27
23	7.53	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18	49.73
24	8.08	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56	51.18
25	8.65	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93	52.62
26	9.22	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29	54.05
27	9.80	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.65	55.48
28	10.39	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99	56.89
29	10.99	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34	58.30
30	11.59	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67	59.70
32	12.81	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49	56.33	62.49
34	14.06	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06	58.96	65.25
36	15.32	17.89	19.23	21.34	23.27	25.64	47.21	51.00	54.44	58.62	61.58	67.98
38	16.61	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16	64.18	70.70
40	17.92	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77	73.40
42	19.24	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21	69.34	76.08
44	20.58	23.58	25.15	27.57	29.79	32.49	56.37	60.48	64.20	68.71	71.89	78.75
46	21.93	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20	74.44	81.40
48	23.29	26.51	28.18	30.75	33.10	35.95	60.91	65.17	69.02	73.68	76.97	84.04
50	24.67	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49	86.66
55	28.17	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29	85.75	93.17
60	31.74	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95	99.61
65	35.36	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42	98.10	105.99
70	39.04	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	104.21	112.32
75	42.76	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39	110.29	118.60
80	46.52	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32	124.84
85	50.32	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24	122.32	131.04
90	54.16	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12	128.30	137.21
95	58.02	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97	134.25	143.34
100	61.92	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17	149.45

Table 5 : Values of t for a Specified Right Tail Area

Percentage Points of the t Distribution



	Level of Significance α									
ν	**0.4**	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.656	127.321	318.289	636.578
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.328	31.600
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.894	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.689
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.660
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Table 6 A : Values of F For a Specified Right Tail Area $F_{0.01\ v1,v2}$

(ν_2 DOF for Denominator)

	Degrees of Freedom for Numerator (ν_1)																		
(ν_2)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6157	6209	6234	6260	6286	6313	6340	6366
2	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Table 6 B : Values of F For a Specified Right Tail Area $F_{0.025\ v1,v2}$

(ν_2 DOF for Denominator)

	Degrees of Freedom for Numerator (ν_1)																		
(ν_2)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	648	799	864	900	922	937	948	957	963	969	977	985	993	997	1001	1006	1010	1014	1018
2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5	39.5
3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.2	14.1	14.1	14.0	14.0	13.9	13.9
4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.73
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

Table 6 C : Values of F For a Specified Right Tail Area $F_{0.05\ v1,v2}$

(ν_2 DOF for Denominator)

	Degrees of Freedom for Numerator (ν_1)																		
(ν_2)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	161	199	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Table 6 D : Values of F For a Specified Right Tail Area $F_{0.10\ v1,v2}$

(ν_2 DOF for Denominator)

	Degrees of Freedom for Numerator (ν_1)																		
(ν_2)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.2	61.7	62.0	62.3	62.5	62.8	63.1	63.3
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Table 6 E : Values of F For a Specified Right Tail Area $F_{0.25\ v1,v2}$

(ν_2 DOF for Denominator)

	Degrees of Freedom for Numerator (ν_1)																		
(ν_2)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.85
2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.48
3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47
4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87
6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74
7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69	1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65
8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.53	1.53
10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.48
11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45
12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.42
13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38
15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36
16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33
18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32
19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.30
20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.29
21	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28
22	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37	1.36	1.34	1.33	1.32	1.31	1.30	1.29	1.28
23	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.27
24	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.26
25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.25
26	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26	1.25
27	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.24
28	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24
29	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.23
30	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23
40	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.19
60	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.15
120	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.10
∞	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.00

Table 7 : Random Numbers

39634	62349	74088	65564	16379	19713	39153	69459	17986	24537
14595	35050	40469	27478	44526	67331	93365	54526	22356	93208
30734	71571	83722	79712	25775	65178	07763	82928	31131	30196
64628	89126	91254	24090	25752	03091	39411	73146	06089	15630
42831	95113	43511	42082	15140	34733	68076	18292	69486	80468
80583	70361	41047	26792	78466	03395	17635	09697	82447	31405
00209	90404	99457	72570	42194	49043	24330	14939	09865	45906
05409	20830	01911	60767	55248	79253	12317	84120	77772	50103
95836	22530	91785	80210	34361	52228	33869	94332	83868	61672
65358	70469	87149	89509	72176	18103	55169	79954	72002	20582
72249	04037	36192	40221	14918	53437	60571	40995	55006	10694
41692	40581	93050	48734	34652	41577	04631	49184	39295	81776
61885	50796	96822	82002	07973	52925	75467	86013	98072	91942
48917	48129	48624	48248	91465	54898	61220	18721	67387	66575
88378	84299	12193	03785	49314	39761	99132	28775	45276	91816
77800	25734	09801	92087	02955	12872	89848	48579	06028	13827
24028	03405	01178	06316	81916	40170	53665	87202	88638	47121
86558	84750	43994	01760	96205	27937	45416	71964	52261	30781
78545	49201	05329	14182	10971	90472	44682	39304	19819	55799
14969	64623	82780	35686	30941	14622	04126	25498	95452	63937
58697	31973	06303	94202	62287	56164	79157	98375	24558	99241
38449	46438	91579	01907	72146	05764	22400	94490	49833	09258
62134	87244	73348	80114	78490	64735	31010	66975	28652	36166
72749	13347	65030	26128	49067	27904	49953	74674	94617	13317
81638	36566	42709	33717	59943	12027	46547	61303	46699	76243
46574	79670	10342	89543	75030	23428	29541	32501	89422	87474
11873	57196	32209	67663	07990	12288	59245	83638	23642	61715
13862	72778	09949	23096	01791	19472	14634	31690	36602	62943
08312	27886	82321	28666	72998	22514	51054	22940	31842	54245
11071	44430	94664	91294	35163	05494	32882	23904	41340	61185
82509	11842	86963	50307	07510	32545	90717	46856	86079	13769
07426	67341	80314	58910	93948	85738	69444	09370	58194	28207
57696	25592	91221	95386	15857	84645	89659	80535	93233	82798
08074	89810	48521	90740	02687	83117	74920	25954	99629	78978
20128	53721	01518	40699	20849	04710	38989	91322	56057	58573
00190	27157	83208	79446	92987	61357	38752	55424	94518	45205
23798	55425	32454	34611	39605	39981	74691	40836	30812	38563
85306	57995	68222	39055	43890	36956	84861	63624	04961	55439
99719	36036	74274	53901	34643	06157	89500	57514	93977	42403
95970	81452	48873	00784	58347	40269	11880	43395	28249	38743
56651	91460	92462	98566	72062	18556	55052	47614	80044	60015
71499	80220	35750	67337	47556	55272	55249	79100	34014	17037
66660	78443	47545	70736	65419	77489	70831	73237	14970	23129
35483	84563	79956	88618	54619	24853	59783	47537	88822	47227
09262	25041	57862	19203	86103	02800	23198	70639	43757	52064

Table 8 : Logarithms at Base 10

N	0	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4645	4669	4683	4698	4713	4728	4742	4757
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627

N	0	1	2	3	4	5	6	7	8	9
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996
N	0	1	2	3	4	5	6	7	8	9

References

1. Bhat B. R. (1981) *Modern Probability Theory* Wiliey eastern Pvt. Ltd., New Delhi.
2. Chung K. L. (1984) *Elementary Probability Theory with Stochastic Processes,* Springer International Student Edition , New York.
3. Feller W. (1972) *An Introduction to Probability Theory and Its Application,* Vo. 1, 3rd ed. Wiley Eastern Pvt. Ltd., New Delhi.
4. Gore A. P., Paranjpe S. A. and Kulkarni M. B. (2006) *100 Datasets for Statistics Education* www.stats.unipune.ernet.in
5. Gore A. P., Paranjpe S. A. and Kulkarni M. B. (2006) *Statistics for Everyone* SIPF Academy, Nashik.
6. Kale B. K. (1999) *A First Course in Statistical Inference,* Narosa Publishing House, New Delhi.
7. Kulkarni M. B., Ghatpande S. B. and Gore S. D. (1999) *Common Statistical Tests,* Satyajeet Prakashan, Pune.
8. Kulkarni M. B. and Ghatpande S. B. (2007) *Introducation to Discrete Probability and Probability Distribution,* SIPF Academy, Nashik
9. Ross Sheldon (2003) *A First Course in Probability Theory,* 6th ed, Pearson Education (Singapore) Pvt. Ltd., Indian Branch, New Delhi.