Indira Gandhi
National Open University
School of Social Sciences

MANI-001
Anthropology and Methods of Research

Block

5

STATISTICAL ANALYSIS

Expert Committee

Professor I. J. S. Bansal
Retired, Department of Human Biology
Punjabi University, Patiala

Professor K. K. Misra
Director, Indira Gandhi Rashtriya Manav Sangrahalaya, Bhopal

Professor Ranjana Ray
Retired, Department of Anthropology
Calcutta University, Kolkata

Professor P. Chengal Reddy
Retired, Department of Anthropology
S. V. University, Tirupati

Professor R. K. Pathak
Department of Anthropology
Panjab University, Chandigarh

Professor A. K. Kapoor
Department of Anthropology
University of Delhi, Delhi

Professor V. K. Srivastava
Principal, Hindu College
University of Delhi, Delhi

Professor Sudhakar Rao
Department of Anthropology
University of Hyderabad
Hyderabad

Professor. S. Channa
Department of Anthropology
University of Delhi, Delhi

Professor P. Vijay Prakash
Department of Anthropology
Andhra University
Visakhapatnam

Dr. Nita Mathur
Associate Professor
Faculty of Sociology
SOSS, IGNOU, New Delhi

Dr. S. M. Patnaik
Associate Professor
Department of Anthropology
University of Delhi, Delhi

Dr. Manoj Kumar Singh
Assistant Professor
Department of Anthropology
University of Delhi, Delhi

Faculty of Anthropology SOSS, IGNOU

Dr. Rashmi Sinha, Reader

Dr. Mitoo Das
Assistant Professor

Dr. Rukshana Zaman
Assistant Professor

Dr. P. Venkatramana
Assistant Professor

Dr. K. Anil Kumar
Assistant Professor

Programme Coordinator: Dr. Rashmi Sinha, SOSS, IGNOU, New Delhi

Course Coordinator: Dr. Mitoo Das & Dr. K. Anil Kumar

Content Editor

Dr. Kaustuva Barik
Associate Professor, Discipline of Economics
Indira Gandhi National Open University

Language Editor

Dr. Anita Sinha
GDM College
Magadh University, Patna

Block Preparation

Unit Writers

Dr. Neha Garg (Unit 1)
Assistant Professor
Discipline of Statistics
IGNOU, New Delhi

Dr. Neha Garg (Unit 2)
Assistant Professor
Discipline of Statistics
IGNOU, New Delhi

Dr. Rajesh Kaliraman (Unit 3)
Assistant Professor
Discipline of Statistics
IGNOU, New Delhi

Dr. G.S. Naidu (Unit 4)
Planning Division
IGNOU, New Delhi

Authors are responsible for the academic content of this course as far as the copyright issues are concerned.

Print Production

Mr. Manjit Singh
Section Officer (Pub.), SOSS, IGNOU, New Delhi

Cover Design

Dr. Mitoo Das, Assistant Professor
Anthropology, SOSS, IGNOU

August, 2011

ISBN-978-81-266-5646-2

Further information about the School of Social Sciences and the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by Director, School of Social Sciences.

Laser typeset by Mctronics Printographics, 27/3 Ward No. 1, Opp. Mother Dairy, Mehrauli, New Delhi-30

Printed by :

BLOCK 5 STATISTICAL ANALYSIS

Introduction

Application of statistical tools in data analysis these days has become an integral part of research methods in several disciplines such as economics, sociology, anthropology, psychology and many others. The abiding interest in statistics emanates from the fact that it opens up avenues to abstract summary measures from seemingly intractable datasets - it shows that it is possible to provide summary statistics, interpret them, and use them for policy formulation.

In this block we have presented some basic statistical tools which you will find useful not only in your anthropological research but also in other spheres of activities. Proper understanding of these tools will sharpen your analytical capabilities.

The **first unit** of the block deals with collection and presentation of data. Collection of data could be through several methods such as survey, experimentation and participation. In all these methods a key issue before us is whether to take a census or a sample. Several factors determine our choice - cost, time, personnel, and most important, the objective of our research. Once data is collected, the next step is to present them - in the form of tables and graphs. Various types of sampling, along with their relative merits and demerits, form the core of Unit 1.

The **second unit** takes you a step further. Summary measures such as arithmetic mean, median and mode help us in finding out a representative central value from a large set of data. This central value helps us in forming opinions about a group, or compare across groups, or observe change over time in a group. The spread of the individual observations around mean, which is called dispersion, is also a subject matter of the unit. In the case of bivariate data, indicating data on two variables on each respondent, we can apply some additional statistical tools, viz., correlation and regression. Correlation coefficient shows the degree of association between two variables. Regression shows a cause and effect relationship - the independent variables lead to some effect on the dependent variable. Remember that these are just statistical tools - in order to apply them you need your insight, logic or theory. Statistics cannot, however, point out which variable is the cause and which one is the effect.

The **third unit** deals with testing of a hypothesis and some elementary design of experiment, particularly ANOVA. The major objective of sample survey is to make prediction and draw inference about the population from which the sample is selected. Here you would be exposed to the rudimentary rules of t-test and chi-square. A few examples are provided to familiarize you with their applications. The third unit also discusses some basic principles of ANOVA.

The **fourth unit**, the last one of the block, deals with application of software, particularly SPSS, in statistical analysis. It assumes that you do not have any knowledge about SPSS and thus begins from a scratch. Opening a new file, defining variables, entering data, and carrying out simple data analysis are explained in the unit. The statistical tools which you have read in the first three units of the block can be applied through SPSS. Given a dataset the SPSS can prepare a graph, compute mean and standard deviation, compute correlation coefficient between two variables or estimate a regression line.

UNIT 1 COLLECTION AND PRESENTATION OF DATA

Contents

Learning Objectives

After going through this unit you will be in a position to:

- explain the importance of statistical techniques in research;
- determine when to use sampling instead of a census;
- distinguish between probability and non-probability sampling;
- define what is sampling and classify sampling methods;
- distinguish between qualitative and quantitative characters;
- prepare a frequency distribution; and
- illustrate data through graphs.

1.1 INTRODUCTION

In order to overcome the challenges of life, human beings continually explore nature, doing some innovation in an attempt to understand and control it. New problems lead to new questions, and our ability to solve these questions and make correct decisions determines our future. Superstition and guesswork have often given way to scientific methods where problems are solved, questions are answered, decisions are made, and actions are taken based on information. We continuously receive

data from many sources – books, magazines, newspapers, journals, computer, internet, radio, television, telephone, movies, etc. Then a question comes to our mind: "what shall we do with all this data?"

Although we spend enormous time, money and energy in collecting data, it will be of no use unless that data can be processed, interpreted and properly used for making decisions. Any research requires decisions or actions based on data, and statistical methods are needed to accomplish the set objectives. Statistical techniques can be used to help in the following:

- Data collection (either through census or sample survey)
- Data processing, which includes organizing, displaying, summarizing and analyzing data
- Data interpretation.

In the beginning of this unit, we will throw light on the significance of statistical techniques in research. We will then highlight different methods of sampling used in collection of data, different types of data and the ways of arranging them. Finally, we will discuss about the different methods available for its graphical presentation of data.

1.2 RELEVANCE OF STATISTICAL TECHNIQUES IN RESEARCH

Statistical techniques are employed not only to help us in making simple decisions every day, but also answering profound questions that society face. For example, a statistical experiment can be conducted to test for the effectiveness of a newly developed vaccine of a dangerous disease. From this experiment we can ascertain whether the vaccine is indeed effective in preventing that disease.

Fig. 1.1 given below shows the general sequence of steps in a research project. Statistical thinking can contribute to every stage; although the major steps like design, analysis and interpretation are of prime focus.

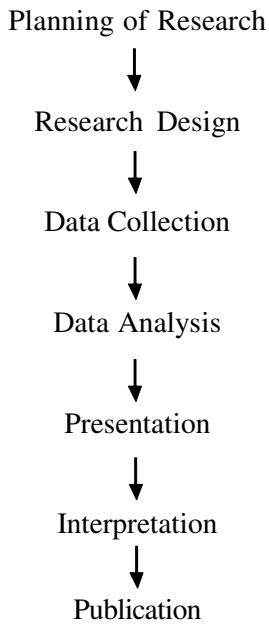


Fig. 1.1

If we wish to investigate, for example, the relationship between weight and diabetes (sugar level) of women, it is not feasible to study all women having diabetes. Then we must study a sample of women having diabetes. The aim of this research would be to extrapolate the findings from this sample to all diabetes. For this inference to be reasonable, it is necessary for the sample of women to be representative of all women having diabetes. In theory we can obtain a truly representative sample only by choosing women at random but even then the sample would be specific to a time period and geographical area. In practice, samples are nearly always chosen in a scientific manner and the subjects' characteristics are described so that their representativeness can be judged. The study just proposed would probably be carried out by taking all women registering at one or more specific hospitals in a set time period.

All the issues described above come under the broad heading of research design. We can say that a correct research design is an essential part of good research.

Apart from the research design, analysis of data plays a major role in research. There are various methods of data analysis in Statistics which are useful in research. We will see that statistical methods of analysis are based on the key idea that we use data from a sample to draw inference about a population. Anthropologists frequently use Statistics to analyze their results. Statistics can help understand a phenomenon by confirming or rejecting a hypothesis. Statistics is often vital to change scientific theories. Then, researchers may apply different Statistical methods to analyze and understand the data better and more accurately. It is generally observed that researches collect data according to their plan of research, objectives and limitations, and then try to find out which statistical tools would suit the data and give a rich interpretation. It is essential to take into account a statistical design before collecting data, especially with respect to sampling. The interpretation of results of statistical analysis always plays an important role. It becomes simpler when the study has a clear objective.

1.2.1 Role of Statistics in Research

Data collected through social survey, laboratory experiment or clinical trial requires statistical analysis before arriving at valid conclusions. Research requirements today call for the use of statistical techniques in every field of knowledge. The findings of any research have to be justified in the light of statistical logic. Simple statistics like averages, percentages and standard deviation would reveal great information in many observational studies. With the availability of standard computer software, it is now an easy job to 'compute' all the statistical parameters. Commonly available software such as MS office can do a lot of data analysis although specialised software are available exclusively for statistical work. Many of these specialised software are freely available on the Internet. The researcher has only to decide which statistical tool to use for particular analysis. Graphical aids from computer would also help in proper interpretation of the results.

Many researchers treat 'Statistics' as a tool that represent their research findings in the form of tables, graphs and summary statistics such as mean, variance, correlation, etc. But statistics has something more than this. It is an inferential science. It is a science of decision-making which helps to find out the truth from the available data. It is the only way out to take decision in the face of uncertainty. The problems posed by researchers in various field such as biology, agricultural research, anthropological studies, economics, etc. to handle quantitative and qualitative data can be solved by using suitable statistical methods.

The importance of Statistics in today's research cannot be undermined. Statistics provides a platform for research as to

- how to go about your research, either to consider a sample or the whole population,
- the techniques to use in data collection and observation,
- how to go about the data description (using measures of central tendency and dispersion).

To wrap it up, statistics as a science of data collection, analysis, interpretation, explanation and presentation will guide you in research for proper characterization, summarization, presentation and interpretation of your research result.

The growing importance of statistical education in anthropology can be easily seen. However, anthropological data are not usually collected as a result of an experimental test which can be replicated. In fact, anthropologists cannot provide for true repeatability of observations in most cases (how can we replicate a historical event, a culture?). For example the researcher is interested in studying the presence of property ownership in married females in hunting and gathering groups.

In anthropology where both qualitative and quantitative data are used, Statistics helps the researchers to present the data in a comprehensive way to explain and predict the patterns of behavior/trend. In cases where the characteristics of the population being studied are normally distributed, the best and statistically important decisions about variables being investigated are possible by using either parametric or non-parametric statistics to explain the pattern of activities. Physical anthropologists, for example, are interested in how humans differ from, and are similar to, other animals, especially nonhuman primates. For example, unlike other primates, we are bipedal. But what are the major anatomical components of bipedal locomotion, and how do they differ from, say, those in quadrupedal ape? To answer these questions, anthropologists have for years studied human locomotion and compared the anatomy of our spine, pelvis, legs, and feet with that of various nonhuman primates.

In the following section, we examine some aspects of data collection for a research study.

1.3 SAMPLING TECHNIQUES

Sampling is widely used as a mean of gathering useful information about a population. Data are gathered from samples and conclusions are drawn about the population. Before performing any sampling, it is important to define clearly the population of interest.

1.3.1 Basic Concepts

a) **Population**

When we come across to the term *population*, we usually think of people in our town, region, state or country and their respective characteristics such as gender, age, marital status, ethnic membership, religion and so forth. In a research design the term 'population' takes on a slightly different meaning. The collection of all units of a specified type in a given region at a particular point or period of time is termed as a 'population' or 'universe'. We can say that the totality or aggregate of all individuals

with the specified characteristic is a population. We may consider a population of persons, families, farms, cattle in a region, or a population of trees or birds in a forest, or a population of fish in a tank depending on the nature of data required. Population may be of two types, finite population, and infinite population.

- **Finite Population**

  A finite population contains a finite number of members or population units. The numbers of citizens in a country, the number of students in IGNOU etc. are the examples of finite population. The number may be large but not infinite.

- **Infinite Population**

  A population that contains an infinitely large number of members is called an infinite population. Number of stars in the sky, number of fishes in the sea, all possible height within the range 150 cm-160 cm (an infinite number of values of heights can be spotted in this range) are examples of infinite population.

b) **Target Population**

The target population is that complete group whose relevant characteristics are to be determined through sampling. A target population may be, for example, all faculty members in the School of Social Sciences, IGNOU, all housewives in Delhi, all students in IGNOU, and all medical doctors in Delhi.

c) **Sample**

Sample is a part of the population. Sampling is the process of drawing a sample from a given population. The results obtained from a sample will be of interest only if they convey something about the population. Thus it has to be an appropriate representative of the population. Remember that census study involves collection of data from the whole population.

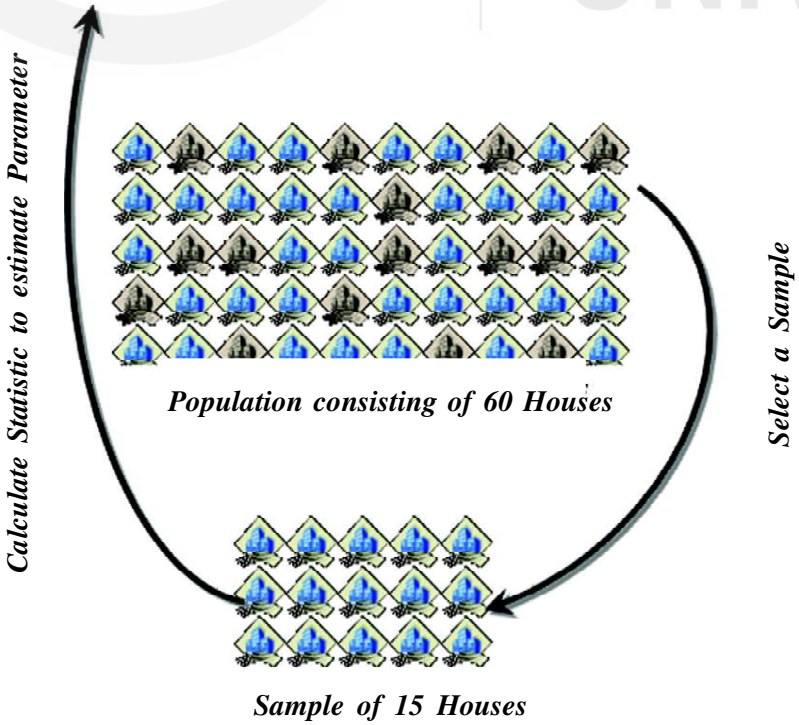


Fig. 1.2

d) **Sampling Units**

The individual elements in the population of interest are called sampling units. The sampling unit may be a single element or group of elements depending upon our research objectives. A sampling unit may be a person, a family, a city, an object or anything else that is the unit of analysis in a population. For example,

- Suppose that we wish to determine the average amount of soft-drink consumed each week by 15 to 17 years olds living in Delhi. In this case, the sampling units would be teenagers between 15 and 17 years residing in Delhi at a particular time.
- Every student at IGNOU
- All child passengers under 18 years of age who are traveling in a train from destination X to destination Y
- All jewelry shops in South Extension market, New Delhi

e) **Sampling Frame**

The sampling frame is a list of all the sampling units belonging to the population that will be used to select the sample. Examples of sampling frames are a student telephone directory (for the student population), the list of companies on the stock exchange, the directory of medical doctors and specialists, the yellow pages (for businesses), list of farms and a list of suitable area segments like villages in India, etc.

f) **Parameters**

Any function of the values of all the population units (or of all the observations constituting a population) is known as a population parameter or simply a parameter. Some of the important parameters usually required to be estimated in surveys are population mean and population variance.

g) **Statistic, Estimator and Estimate**

Suppose a sample of 'n' units is selected from a population of 'N' units. Any function of these sample values is called a 'statistic'. An 'estimator' is a statistic obtained by a specified procedure (for example a formula). The estimator is used for estimating a population parameter. The value of the estimator differs from sample to sample. The particular value, which the estimator takes for a given sample, is known as an estimate.

h) **Sampling and Non-sampling errors**

As mentioned earlier, we draw inferences about the population on the basis of information contained in the sample. Since we do not study the whole population we are likely to commit certain error in the process. The error arising due to drawing inferences about the population on the basis of sample is termed sampling error. The sampling error is non-existent in a complete enumeration census.

Non sampling errors are due to wrong reporting, recording or entry of data. Thus Non sampling errors can occur both in sample surveys and census. However, the extent of Non sampling errors is less in a sample survey as we may employ qualified field staff in data collection gives proper training and scrutinise the data effectively.

i) Sample or Census

In a census or 'complete enumeration' we obtain data from every member of the population. It is usually difficult to carry out a true census as it involves a lot of money, manpower and time. Further 100 percent survey is sometimes impossible.

A census becomes essential in some cases where the units to be inspected are a few and the nature of the problem requires data from each such unit. For example:

Suppose there are only 10 sugar factories in a region. If the study requires some vital information about this industry, it is necessary to include all of them. Similarly when a university wants to test the knowledge of its students, it examines all students belonging to its various affiliated colleges. In such cases, census is the only way.

In order to know the blood glucose level in a person, one should be satisfied with the estimate based on a few drops (sample) of blood. One cannot think of extracting all the blood (population) from the body. Similarly, when the quality control officer wants to test the length of life of electric bulbs produced in a factory he only takes a sample of bulbs, burns them out, then puts his opinion about the whole lot of bulbs on the basis of these few bulbs. For this purpose, he cannot burn out all the bulbs. Thus, in certain situations, we cannot think of studying the whole population. In these situations the only way is to take a sample and to make an inference about the whole population. A good sample carefully selected and studied can give satisfactory information on the characteristics of the population. Whenever sampling is resorted to it is important that the sample should reflect all the qualities found in

Non- sampling errors arising through non-response, incompleteness and inaccuracy of response are likely to be more wide-spread and important in a census than in a sample survey.

Sampling error usually decreases with increase in sample size (number of units selected in the sample) while the non-sampling error is likely to increase with increase in sample size.

As regards the non-sampling error, it is likely to be more in the case of a complete enumeration survey than in the case of a sample survey since it is possible to reduce the non-sampling error to a great extent by using better organization and suitably trained personnel at the field and tabulation stages in the latter than in the former.

j) Advantages of Sampling

Sampling is used extensively for many reasons. In many situations a sample produces information about the population more accurately than a census. Additionally, in obtaining a sample, fewer interviewers are required and it is likely that they will be better trained than the huge team of interviewers required to perform a census.

Sampling offers several advantages over census survey:

- Sampling can save money (cost of data collection and processing)
- Sampling can save time as data on fewer units need to be collected
- For given resources, sampling can broaden the scope of the study
- If accessing the population is impossible, sampling is the only option.

1.3.2 Sampling and its Types

The principal objective of sampling is to get maximum information about the population with minimum effort or with limited resources. There are various procedures to obtain a representative sample (to select sampling units). All the procedures have their advantages and disadvantages. Sampling procedures may broadly be of two categories:

- Probability Sampling
- Non-probability Sampling

Under each category there are several procedures of sample selection.

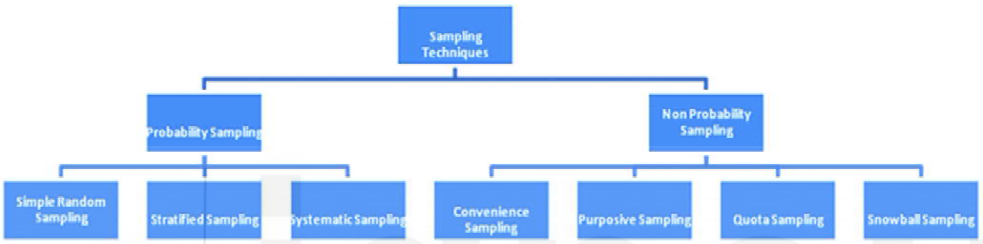


Fig. 1.3

a) Probability Sampling

Probability sampling is a scientific method of selecting sampling units. In this method, sampling units are selected according to certain laws of probability. Each unit of the population has some definite (non-zero) probability of being included in the sample. In this procedure, the units are selected by using some random mechanism so this is also called random sampling. Some of the important probability sampling methods are simple random sampling, stratified random sampling and systematic random sampling.

i) Simple Random Sampling

Simple random sampling is the most elementary technique of probability sampling. In simple random sampling, each and every unit of population has an equal opportunity of being selected in the sample. The item that gets included in the sample is just a matter of chance and the selection is not influenced by personal bias of the investigator. Simple random sampling is of two types:

- Simple random sampling without replacement (SRSWOR)
- Simple random sampling with replacement (SRSWR)

Simple random sampling without replacement

If the selected unit is not replaced in the population and the next draw is made from the remaining units then the selection procedure is called simple random sampling without replacement. Here there is no chance of a sampling unit being replaced.

Simple random sampling with replacement

If the unit selected at one draw is replaced in the population before the next draw then the procedure is called simple random sampling with replacement. Here the same sampling unit can be selected a second time in another draw.

There are two methods of selecting a simple random sample.

– Lottery Method

In this method, all units (say N) of the population are numbered or named on separate slips of paper of identical size and shape. These slips are folded and mixed up in a bowl and one of the slips is drawn after thoroughly mixing the slips. For example, suppose we have a population of 100 students and we wish to draw a random sample of 10 students. We can number the units of the population (in this case, students) in serially from 1 to 100. We take 100 identical slips of paper, write numbers 1 to 100 on them, put them in a bowl, mix them thoroughly and pick out 10 slips, one by one without looking. This gives us a random sample of 10 students. In simple random sampling with replacement procedure, the slip is replaced before the next draw while in case of without replacement it is not replaced. Sampling is continued till desired number of units are selected. A disadvantage with the lottery method is that it is quite time consuming and cumbersome to use if the population is sufficiently large.

– Random Number Table (RNT)

An alternative method of random selection is the use of random number table (RNT) which is an arrangement of digits 1 to 9. A RNT is so constructed that each of the digit 0, 1, 2, ……..,9 has an equal chance of selection at any draw and independent of each other. Numbers in the list are arranged so that each digit has no predictable relationship to the digits that preceded it or to the digits that followed it. In short, the digits are arranged randomly. The random numbers are prepared by using certain randomizing machines and then arranged in rows and columns.

These tables can be produced by a computer. Some published random number tables in common use are prepared by Tippett (1927), Fisher and Yates (1938), and Kendal and Smith (1939).

A part of random number table (computer generated) is produced in Appendix A. If the size of the population under consideration is less than 10, then you have to consider single digits in the RNT. On the other hand, if your population size is upto 100 you can consider two digits in the RNT (if between 100 and 999, then three digits). A practical method of selecting a random sample is to choose units one-by-one with the help of a table of random numbers. For example, by considering two-digit numbers, we can obtain numbers from 00 to 99, all having the same chance. The simplest way of selecting a sample of the required size from the population of size N is to select a random number from 1 to N and then taking the unit bearing that number. For example, suppose we have 500 students in MA Anthropology and we have to select a simple random sample of 20 students to measure their height and weight. So for selecting a simple random sample without replacement, first of all we assign serial numbers 1 to 500 to all the students in an order. Now from the random number table (see **Appendix A** at the end of this unit), we will arbitrarily select any number and start with this number. Let us assume that we start from the number at the intersection of 5th row from the top and 2nd column from the left (25277). Since our population size is 500, so we need to take three digit numbers between 001 and 500. In this situation we read all three digit numbers starting from this number. We can read numbers horizontally, vertically or diagonally. Suppose we are reading numbers vertically, then the numbers are 252, 478, 318, 735, 902, 916, 884, 180, 897, 815, 778, 478, 158, 789, 099, 280, 229, 897, 436, 880, 224, 917, 001, 561, 508, 754, 635, 396, 497, 518, 317, 429, 722, 012, 938, 889, 864, 516,

640, 508, 026, 341, 379. Drop all the number which are greater than 500 and take only 20 (sample size) numbers which are not greater than 500. If we are selecting SRSWOR, we also drop the numbers which are repeating. So we will select the following numbered students in our sample:

252, 478, 318, 180, 478, 158, 099, 280, 229, 436, 224, 001, 396, 497, 317, 429, 012, 026, 341, 379.

ii) **Stratified Random Sampling**

When the population is heterogeneous (non- homogeneous), it is appropriate to use stratified random sampling. In this sampling scheme, we divide the heterogeneous population into different non-overlapping homogeneous groups called strata. Then we select a separate sample from each stratum through simple random sampling. Some of the commonly used stratifying factors are age, sex, educational or income level, geographical area, economic status and so on. For example:

- If we are interested in studying the consumption pattern of households of Delhi, the city of Delhi may be divided into various strata (such as zones or wards).
- To estimate the average income of Delhi city, we can divide the population in to low, middle and high income groups; and select a sample from each group.
- Suppose we conduct a national survey. We might divide the population into four strata based on geography, viz., north, east, south and west. Then, within each stratum, we might randomly select survey respondents.

The purpose of stratification is to increase the efficiency of sampling by dividing a heterogeneous population into homogeneous groups in such a way that there should be as much homogeneity as possible within each stratum and heterogeneity between the strata. Stratified sampling gives a more representative sample than simple random sample.

iii) **Systematic Random Sampling**

Systematic random sampling is a commonly used technique if the complete and up-to-date list of the sampling unit (sampling frame) is available. The list may be prepared in alphabetical, geographical, numerical or some other order. This consists in selecting only the first unit at random and the rest being automatically selected according to some predetermined pattern involving regular spacing of units.

The first unit is selected at random generally by following the lottery method or using random number table. Subsequent units are selected by taking every kth unit from the list where k refers to the sampling interval or sampling ratio, i.e., the ratio of population size to the sample size. Symbolically,

$$k = \frac{N}{n}$$

where k= sampling interval, N= population size, n= sample size

Example: In a class, there are 86 students with roll numbers 1 to 86. It is desired to take a sample of 10 students using the systematic random sampling method

To determine the sample, we first decide 'k'.

$$k = \frac{N}{n} = \frac{86}{10} = 8.6 \text{ or } 9$$

From 1 to 86 roll nos., the first student should be between 1 and k, i.e., 1 and 9. We select this student at random by lottery method. Subsequently, we take every ninth student. Suppose the random number between 1 and 9 comes out to be 5. The sample would then consist of the following roll numbers.

5,14,23,32,41,50,59,68,77,86

b) **Non-probability Sampling**

Sampling procedures, which are based not on random procedures, but on subjective judgment or convenience of the sampler, are known as non-random sampling or non-probability sampling. This method of sampling includes convenience sampling, judgment (or purposive) sampling, quota sampling, and snowball sampling.

Random sampling is preferred over non-random sampling for variety of reasons. Besides eliminating the subjectivity in selection, it provides a measure of reliability associated with the estimates developed from the samples. Thus, we can make inferences from random samples with a known level of confidence. As stated earlier, in random sampling procedure, every unit in the population is assigned certain probability of selection. The randomness associated with the sampling procedure is the key to make valid inference from the sample.

i) **Convenience Sampling**

In this sampling technique, we select those sampling units which are most conveniently available at a certain point or over a period of time. Convenience samples are obtained by choosing the easiest objects available. Suppose you want to study the length of stay and food habits of cancer patients in a hospital. Then choose a hospital close to your office or home and interview some people over there. In this case you are following the convenience sampling method. Convenience samples are prone to bias by selecting sampling units which are convenient to choose. Hence the results obtained by convenience sampling may not be representative of the population. Major advantage of convenience sampling is that it is less time consuming, convenient and economical; a major disadvantage is that the sample may not be representative of the population. Convenience sampling is best used for the purpose of exploratory research. For example, to include the first ten people to walk out of a store in your sample.

ii) **Judgment (Purposive) Sampling**

This is a sampling technique in which the researcher selects the sample based on judgment about some appropriate characteristic of the sampling units. This is usually an extension of convenience sampling. Personal bias may be prominent in purposive sampling.

- **Example 1:** The Consumer Price Index (CPI) is based on a judgment sample of market-based items, housing costs, and other selected goods and services which are representative for most of the population in terms of their consumption.
- **Example 2:** Suppose a researcher wants to give the picture that the standard of living has increased in New Delhi. One may take individuals in the sample from the posh localities and ignore the localities where low or middle class families live.
- **Example 3:** A researcher may decide to draw the entire sample from one representative village; even through the population may be distributed over a number of villages. In this method, the researcher feels that the chosen village is representative of the entire population.

iii) **Quota Sampling**

In quota sampling, the population is first segmented into mutually exclusive sub-groups, just as in stratified random sampling. Then judgment sampling is used to select the units from each strata or segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60. This means that researcher can put a demand on who he/she wants to sample (targeting). In other words, quota sampling is a method of stratified sampling in which the selection within strata is non-random. Thus it is a restricted type of purposive or judgment sampling. In quota sampling quotas are fixed according to some specified characteristic such as income level, sex, age, occupation or religious affiliation. This is a sampling technique in which the researcher ensures that certain characteristics of a population are represented in the sample to an extent he or she desires. Each interviewer is then told to interview a certain number of persons, which constitute his quota. Within the quota, the selection of sample items depends on personal judgment. For example, in a TV viewership survey each interviewer may be told to interview 100 persons, out of which 50 are to be housewives, 30 to be servicewomen and 20 to be children under the age of 15. Within these quotas the interviewer is free to select the people to be interviewed.

iv) **Snowball Sampling**

Snowball sampling is a special non-probability method used when the desired sample characteristic is rare. It is also called a network or chain referral sampling. Snowball sampling is based on an analogy to a snowball which begins with small but becomes larger as it rolls on wet snow and picks up additional snow. It may be extremely difficult or cost may be prohibitive to locate respondents in these situations. Snowball sampling relies on referrals from initial subjects to generate additional subjects. In other words, this is a sampling technique in which individuals or organizations are selected first, and then additional respondents are identified based on information provided by the first group of respondents. For example:

- Through a sample of 500 individuals, 20 scuba-diving enthusiasts are identified which, in turn, identify a number of other scuba-divers.
- An investigator finds a rare genetic trait in a person, and starts tracing his pedigree to understand the origin, inheritance and etiology of the disease.

The advantage of snowball sampling is that smaller sample sizes and costs are necessary; a major disadvantage is that the second group of respondents suggested by the first group may be very similar and not representative of the population with that characteristic.

While carrying out a sample survey there are certain steps that you should follow. These are as follows:

- **Specification of objective:** The objective of carrying out sampling is the first and foremost step in sampling procedure. Because all other steps will follow from the objective.
- **Definition of population:** In this step we define the units that should be included in the population. Many times there are certain border cases where proper definition is important. For example, if you want to survey the health workers in an area, whether you should include part-time employees in the population or not.

- **Preparation of 'sampling frame':** Once you have defined the units to be included in the population, the next step is to prepare a list of the units from which the sample is to be drawn. Many times problems come up because the sources from which you want to prepare the sampling frame may be incomplete or obsolete.

- **Identification of sampling procedure:** Sampling procedure refers to the method of selecting the sample. As mentioned in the previous section, there are quite a few sampling procedures available. We should select a method that should be i) in a position to give us a representative sample, ii) feasible to carry out keeping in view our constraints, and iii) cost effective.

- **Determination of sample size:** The next step is the determination of sample size. The factors that influence sample size are i) population size, ii) variance of population units, iii) desired precision level, iv) response rate, and v) availability of resources. The sample size should be relatively larger if population size is larger, variability among units is higher, more precise results are required, and lesser is the response rate. However, we should settle for smaller sample size if there are constraints such as availability of funds or manpower or time. Sample is considered to be large if its size is greater than 30. If sample size is less than or equal to 30 it is considered as small sample. The procedures of drawing inferences for large and small samples are different.

- Selection of sampling units: Once you have decided on the sampling procedure and the size of the sample, the next step is draw the sampling units (that is, units included in the sample) from the sampling frame.

1.4 FREQUENCY DISTRIBUTION

1.4.1 Variable and Data

Let us begin with the concept of a variable. It is a characteristic of the *sample* or the *population* that we intend to measure. Thus age of a person is a variable, so is gender, educational level or mother tongue. All variables are not similar.

Variables can be of two types: qualitative and quantitative. Qualitative variable is one that cannot be expressed in numerical terms. For example, marital status is a qualitative variable. Here we can have two categories: married and single. Of course, if you want a more detailed categorization you can further divide single in to widow/ widower, divorcee and never married. Similarly, gender (male or female), mother tongue (Hindi, Bengali, Oriya, Tamil, Urdu, etc.), subject categories (economics, history, physics, medicine, etc.), religion (buddhism, christianity, hinduism, muslim, etc.) are examples of qualitative variables. Here we study an attribute or quality that cannot be quantified, but can be divided into various categories. Moreover, we cannot say that one category is higher or greater than another category. Such variables are also called 'nominal variables'.

There is another type of qualitative variable where we can divide the observations into various categories and also say that one category is higher or greater than another category. An example could be the educational qualification of the head of a household. Here we can divide their educational qualification into categories such as 'secondary', 'senior secondary', 'graduate' and 'post-graduate'. In this case, obviously, the category 'Senior Secondary' is higher than the category 'Secondary'

in terms of number of years of schooling and expected mental maturity. In this case we can arrange the categories in an ascending or descending order. This sort of variables is called 'ordinal variables'.

In the case of nominal variables we cannot perform any mathematical operations (such as addition, subtraction, multiplication, division,) or logical operations (greater than, less than) across categories. We can simply count the number of observations in each category. In the case of ordinal variables we can say that one category is greater than another category. But we cannot quantify the difference between categories. For example, we cannot express numerically the difference between two categories (say secondary and senior secondary). Also we cannot say that the difference between two categories (say secondary and senior secondary) is the same as the difference between two other categories (say graduate and post-graduate).

A quantitative variable can be expressed in numerical terms. Hence it is also called 'numerical variable'. Examples of numerical variable could be age, income, weight, height, distance travelled, etc. This category of variables can be subjected to various mathematical and logical operations. Thus we can express the monthly income of a doctor in rupees and also say by what percentage it exceeds the salary of a lab assistant.

Numerical variables can be of two types: discrete and continuous. Discrete variable is one where the observations assume values in complete numbers. For example, the number of children in family can only be whole numbers; it cannot be fractions. On the other hand, continuous variables can assume any value in an interval. For example, weight of a person can be measured to any precision and thus can take any value in between two points.

Let us distinguish between variable and data. We obtain data by measuring a variable (qualitative or quantitative) on certain individuals or sampling units. For example, suppose we measure the height of 50 employees in an organisation. Here height is the variable and we obtain 50 observations. These 50 numerical values that we obtain are our data. Thus we have discrete data or continuous data depending upon whether the variable is discrete or continuous.

Similarly there are primary data and secondary data. Primary data refers to data collected by the researcher by undertaking a field survey. On the other hand, secondary data refers to collection of data from published sources, e.g., census, budget, handbooks, etc. Thus when you undertake a field survey, collect data, analyse the results and present it in some forum, it is primary data. But when I use that data for further analysis it becomes secondary data for me.

Data are collected either as part of a normal routine or specifically for an investigation. Depending on the nature of the problem, they may relate to individuals, families, houses, villages etc. The data collected are known as observations. The individual subjects or objects upon whom the data are collected are known as statistical units. The items or characteristics on which observations are made are known as variables. For example in a study of hypertension among 400 students, we may record student's age, height, weight and blood pressure reading. In this case, we have a data set of 400 students with observations recorded on each of five variables for each. From the point of view of statistical analysis, data can be broadly classified as qualitative data or quantitative data.

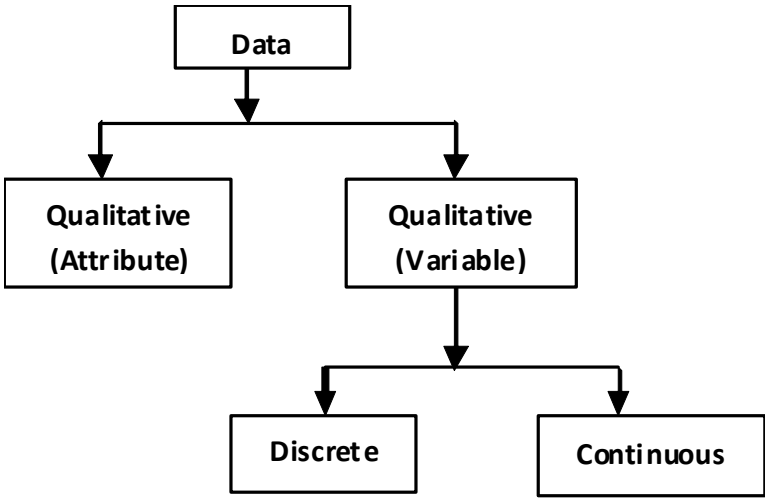


Fig. 1.4

Raw Data – Data when collected in original form is called "raw data". Raw data, or data that have not been summarized in any way, are sometimes referred to as ungrouped data. Table 1.1 contains raw data of the age of 75 persons.

Table 1.1: Data on the age of 75 persons

26	40	42	50	38	22	3	62	51	48
60	37	44	74	19	17	42	57	35	23
29	41	51	36	32	61	50	57	39	68
21	37	52	70	22	42	48	42	53	54
32	17	59	54	50	33	44	39	25	56
35	41	26	33	53	31	16	28	70	52
49	74	60	51	66	41	71	69	45	61
25	60	29	15	40					

1.5.2 Frequency Distribution

One of the major objectives of statistical methods is to describe appropriately the characteristics of a mass of data. In any investigation, after collecting the data, the next step of the researcher should be to organize and simplify the data so that it is possible to get a general overview of the results which will permit conclusions to be drawn directly or by means of further calculations. It is not possible to detect relationships between various factors at issue from the unsorted mass of figures. One method for simplifying and organizing data is to construct a frequency distribution. A frequency distribution is an organized tabulation showing exactly how many individuals are located in each category on the scale of measurement. The preparation of frequency tables thus constitutes an important step when the number of observations is large.

When observations are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. The data of a long series of observations need to be systematically organized and recorded so as to enable analysis and interpretation. In this section, we shall discuss the method of organizing raw data into frequency distributions. We easily can get information from frequency distributions than raw data. Frequency distribution may be of two types.

a) **Ungrouped Frequency distribution**

b) **Grouped Frequency distribution**

We briefly explain each of the above two types below.

a) **Ungrouped Frequency distribution**

Ungrouped frequency distribution can be used when the data is qualitative in nature and also when the variable under consideration is discrete. Let us discuss each situation separately.

i) **Ungrouped Frequency Distribution for Discrete Variable**

Let us consider the following 40 families, in which the number of children per family was recorded.

Table 1.2: Number of children of 40 families

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

This representation of the data does not furnish any useful information and is rather confusing to mind. Now we would like to summarise the data by forming a frequency distribution. A frequency distribution presents an organized picture of the entire set of data. For preparing a frequency table, we have just to count the number of times a particular value is repeated which is called the frequency of that class. We can count more easily if we follow a tallying system.

A frequency distribution table consists of at least three columns.

- In the first column, values of the variable (X) under study are listed from the lowest to highest
- In the second column, tally marks are determined for each value of variable(how often each X value occurs in the data set)
- For the third column, frequency (f) is calculated for each X value. The sum of the frequencies should be equal to N.
- The forth column can be used for the relative frequency (rf) for each value of X. The sum of the relative frequency column should be equal to 1.

In the above table, the number of children is varying from 0 to 6. So we take 7 classes defined by 7 district values 0, 1, 2, 3, 4, 5, 6 in first column. Second column of the table shows the tally marks against each of these values. A bar (|) called tally mark is put against the number when it occurs. Having occurred four times, the fifth occurrence is represented by putting a cross tally (/) on the first four tallies. After counting these tally marks, we write the frequencies in the third column.

Table 1.3: Frequency distribution for number of children of 40 families

No. of Children	Tally Marks	No. of families (frequency)	Relative Frequency
0	𝍸 I	6	0.15
1	𝍸	5	0.125
2	𝍸 III	8	0.2
3	𝍸 II	7	0.175
4	𝍸 I	6	0.15
5	𝍸	5	0.125
6		3	0.075
		40	1

The representation of data as above (Table 1.3) is known as frequency distribution or ungrouped frequency distribution. Number of children are called variables (x) and number of families against the number of children is known as the frequency (f) of the variable. For example, in the above table the frequency of 2 is 8 it means there are 8 families having 2 children.

Alternatively, we can also write these frequencies in terms of proportions. These proportions give the relative frequencies. In the fourth column we have written the relative frequencies. By definition,

$$relative\ frequency\ of\ a\ class = \frac{frequency\ of\ that\ class}{total\ frequency}$$

The frequencies answers the questions of the type "how many families have 2 children?" and the relative frequencies deal with questions like "what is the proportion of families having 2 children?"

Cumulative Frequencies

Now suppose if we want to see how many families have 3 or less children then we have to add up the frequencies of 0, 1, 2 and 3 children to get the number of families having 3 or less children. It will give us the cumulative frequencies of 3 or less. In this way, we can calculate the cumulative frequencies of less than type by adding the frequencies from the top class of the table to downwards. If we add up the frequencies from the bottom class to upwards, we get the cumulative frequencies of the more than type. Table 1.4 below shows both types of cumulative frequencies for the data set given in Table 1.3.

Table 1.4: Cumulative Frequencies of the Number of Children in 40 Families

No. of Children	frequency	cumulative frequencies (less than type)	cumulative frequencies (more than type)
0	6	6	40
1	5	11	34
2	8	19	29
3	7	26	21
4	6	32	14
5	5	37	8
6	3	40	3
	40		

In Table 1.4, the less than type cumulative frequency of 4 is 32 this means that there are 32 families having 4 or less children. Also the more than type cumulative frequency of 5 is 8. Thus there are 8 families who have 5 or more children.

ii) **Ungrouped frequency distribution for qualitative data**

Let us consider the following data of 200 males classified according to their headache status.

Table 1.5: Classification of 200 Males according to their Headache Status

Headache Group	No. of Males (frequency)	Relative Frequency
No headache	58	0.29
Simple headache	63	0.315
Unilateral headache	37	0.185
Migraine	42	0.21
Total	200	1

Source: Biostatistics: A foundation for Analysis in the Health Sciences by W.W. Daniel.

Until now, we have seen how to construct ungrouped frequency tables for discrete variables. For such a table, we count the frequencies of each distinct value taken by the variable, and so there is no loss of information. But this may not always be feasible. For example, suppose we have raw data on the marks of 300 students studying in IGNOU. In this case, if we construct a frequency table taking each distinct value, then the table would be too long. Also, ungrouped frequency tables cannot be constructed for data on continuous variables, because a continuous variable can take infinitely many distinct values. In such cases, we group some variable values together and then construct frequency tables. In the next subsection, we will discuss frequency distribution for continuous data.

b) **Grouped Frequency Distribution**

Sometimes, however, a set of data covers a wide range of values. In these situations,

a list of all the X values would be quite long or too long to be a "simple" presentation of the data. To remedy this situation, a grouped frequency distribution table is used. In a grouped table, the first column lists groups of data, called class intervals, rather than individual values. These intervals all have the same width, usually a simple number such as 2, 5, 10, and so on.

In this type of frequency distribution, we divide the observed range of variable into a suitable number of class-intervals and count the number of observations in each class. We usually make class-intervals by two methods (i) inclusive method (ii) exclusive method. For describing the concept of grouped frequency distribution, we are considering the data of the age of 75 persons shown in Table 1.1:

i) Inclusive Method

The classes in which both the upper and lower limits are included are called inclusive classes. The lowest observation in the above data set is 15 and the highest is 75. If we take our classes as 15-24, 25-34, 35-44, etc. then the total number of classes will be 6. For calculating the frequencies for these classes we again go for the same procedure of tally marks as for table 3

Table 1.6: Grouped frequency distribution of the age of 75 persons

Age in years Class interval	Tally Marks	frequency (f)	Relative frequency
15-24	𝍸 ////	9	0.12
25-34	𝍸 𝍸 //	12	0.16
35-44	𝍸 𝍸 𝍸 𝍸 /	21	0.28
45-54	𝍸 𝍸 𝍸	15	0.2
55-64	𝍸 𝍸 /	11	0.146667
65-74	𝍸 //	7	0.093333
Total		75	1

Such a table showing the distribution of the frequencies in the different classes is called a frequency table and the manner in which the class frequencies are distributed over the class intervals is called the grouped frequency distribution of the variable.

Here, the classes of the type 15-24, 25-34 etc. in which both the upper and lower limits are included are called inclusive classes. For example, the class 15-24 includes all the values from 15 to 24 (both 15 and 24 inclusive).

ii) Exclusive Method

If the inclusive classes, we saw that persons with ages between 24 and 25 years are not taken into consideration. In such a case, we can form the class as 15-25, 25-35, etc. In this type of classes the upper limits of each class are excluded from the respective classes and are included in the immediate next class known as exclusive classes. Here the class intervals are so fixed that the upper limit of one class is the lower limit of the next class.

For the data set given in Table 1.1, the grouped frequency distribution by considering exclusive class-intervals is given in the following table:

Table 1.7: Grouped frequency distribution of the age of 75 persons

Age in years Class interval	Tally Marks	Frequency (f)	Relative Frequency	Mid-v Values
15-25	𝍸 I I I I	9	0.12	20
25-35	𝍸 𝍸 I I	12	0.16	30
35-45	𝍸 𝍸 𝍸 𝍸 I	21	0.28	40
45-55	𝍸 𝍸 𝍸	15	0.2	50
55-65	𝍸 𝍸 I	11	0.146667	60
65-75	𝍸 I I	7	0.093333	70
Total		75	1	

When the variable under study is continuous, then we use exclusive method and the inclusive method should in general be used in case of discrete variable. To ensure continuity and to get correct class interval we should adopt "exclusive method" of classification.

The cumulative frequencies of both types for the above data set are given in the following table

Table 1.8: Cumulative Frequencies

Age in years (Class Interval)	Frequency (f)	Cumulative Frequencies (less than type)	Cumulative Frequencies (more than type)
15-25	9	9	75
25-35	12	21	66
35-45	21	42	54
45-55	15	57	33
55-65	11	68	18
65-75	7	75	7
Total	75		

It is important to note that the upper and lower class limits of the new exclusive type classes are known as 'class boundaries'. The difference between the upper and lower boundaries of a class is known as 'class interval' of that class. For the class 15-25, the class interval is 10 (i.e., 25-15). An important decision while constructing a frequency distribution is about the width of the class interval, i.e. whether it should be 5, 10, 15, 20 etc. the decision would depend upon a number of factors such as the range in the data i.e. the difference between the smallest and largest item; number of classes to be formed

$$class\ interval = \frac{Largest\ value - Smallest\ value}{the\ number\ of\ classes}$$

$$i.e.,\ i = \frac{L - S}{k}$$

For the data given in Table1.1, the age of 75 persons varied between 15 and 74 and suppose we want to form 6 classes, then the class interval would be

$$i = \frac{74-15}{6} = \frac{50}{6} = 9.866 \approx 10$$

The starting class would be 15-25, second class 25-35 and so on. Generally, we fixed the number of classes on the basis of nature of the problem under study or we can also decide it using Sturges' formula. By Sturges' formula, the number of classes will be

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

Where N= Total no of observations

For the above example, the no. of observations are 75, the number of classes will be

$$k = 1 + (3.322 \times \log)_{10} 75)$$

$$= 1 + 3.322 \times 1.8751 = 6.2291 \approx 6$$

The mid value or class mark is the value lying halfway between the lower and upper class limits of a class-interval. i.e.

$$mid\ value\ of\ a\ class = \frac{lower\ boundary\ of\ the\ class + upper\ boundary\ of\ the\ class}{2}$$

For the class 15-25, the mid value will be

$$mid\ value = \frac{15+25}{2} = \frac{40}{2} = 20$$

Preparation of a Table

We have presented frequency distributions for discrete and continuous data in the form of tables. There are certain issues we should take care of while preparing a table.

- It is required to give a *table number* for identification of the particular table.
- There should be a *title of the table* that indicates the type of information contained in the table. Title should be brief and precise. Avoid expressions like 'Table presents…' or 'Table contains….' as part of the title.
- If necessary give a *head note*. It should be given in parentheses and should appear on the right side top just below the title. See, for example, the expression (in Rupees) given in Table 1.9 below.
- *Stub head* describes the nature of stub entry, e.g., 'class interval' in Table 1.9.
- *Stub entries* describe the rows.
- *Caption* describes the nature of data presented in columns followed by column heads and sub-heads. In certain tables it may not be necessary to give sub-heads.
- The *main body of the table* contains numerical information.
- Below the table there may be *footnote*. The purpose of footnote is to caution the readers about the limitations of the table..
- *Source* of the table may be the last component. It is quite important in the case of secondary data. It provides opportunity to the readers to check the data if they desire and get more of it.

Remember that you have to design your own table, keeping your requirements in view. In Table 1.9 we have summarised different parts of a table.

Table 1.9

(----------TITLE----------)

(Head note)

Stub Head	←---------- Caption ----------→			
	Column Head I		Column Head II	
	Sub-head	Sub-head	Sub-head	Sub-head
Stub Entries	MAIN BODY OF THE TABLE			
Totals				

Foot note:
Source:

1.5 GRAPHICAL REPRESENTATION

In addition to tables we often use graphs for presentation of statistical data. They furnish a visual method of examining the data. They make a more lasting impression than detailed numbers and convey an idea forcefully at a single glance. They help us to get a real grasp of the overall picture rather than the details. They facilitate quick and accurate comparison of data relating to different periods of time or different regions.

The researcher needs to decide which type of graph is most appropriate in a given situation. Indeed, a number of software packages allow users to enter the data and decide which method of presentation is most appropriate for their needs. In this section, we discuss about bar graphs, line chart, pie chart and histogram.

1.5.1 Bar Diagram

It is used to describe the frequency of cases belonging to different categories of variable. Bar diagrams are the most commonly used graphical representation. A simple bar diagram is used to represent only one variable. For example, the figures of sales, production or population for various years (points) may be shown by means of simple bar diagram. The gap between one bar and another should be uniform throughout. We take bars of the same width and the length of a bar represents the value of the variable concerned. A bar can be either vertical or horizontal. In practice, vertical bars are more popular. As an example, we consider the data given in Table 1.3.

Table 1.10: Number of Children in 40 Families

No. of Children	No. of families (frequency)
0	6
1	5
2	8
3	7
4	6
5	5
6	3
Total	40

The frequency distribution for the above data given in Table 1.10, is represented by bar diagram. In this bar diagram, we have taken number of families on Y-axis and number of children on X-axis. We plot the bars perpendicular to X- axis and the height of each bar represent the number of families corresponding to the number of children.

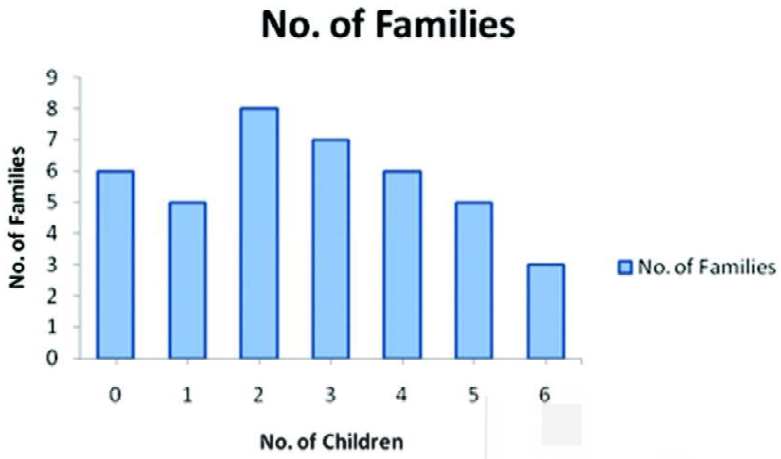


Fig. 1.5

For each graph you should give a caption and indicate the variables measured on x-axis and y-axis. Often we provide the unit of measurement below the caption (for example in cm or Rs crore).

1.5.2 Line Diagram

A line diagram is the way of representing data usually to describe the changes over a period of time. The trends in any time series data is normally represented by a line diagram. We can show one or more variables in a line diagram. Example includes the economical data, share market trends, recovery rate of a patient etc.

On the X-axis (horizontal) we generally take the time and on Y-axis (vertical) the value of the variable and then we plot the points on the graph which are next joined by straight lines.

Table 1.11 given below shows daily temperatures of New Delhi, recorded for 10 days, in degrees Celsius

Table 1.11: Temperature in New Delhi For 10 Days

Day	Temperature (°C)
1	19
2	17
3	21
4	11
5	15
6	17
7	12
8	17
9	15
10	18

The data given in Table 1.11 above is presented in the form of a line diagram. Thus it shows the temperature of New Delhi (in °C) over 10 days:

(in °C)

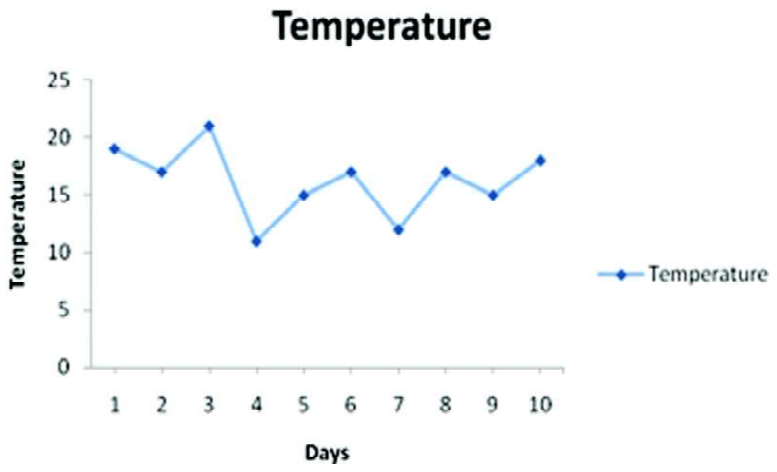


Fig. 1.6

1.5.3 Pie Diagram

Pie diagrams are popularly used in practice to show percentage breakdowns. It is used when the researcher wants to depict the share of various items in total. Household expenditure on various heads, percentage of blood donations of different groups in a blood bank etc. are some examples of pie diagram. For constructing a pie diagram, we first convert the various component values into corresponding degree and draw a circle of appropriate size with a compass. Then we mark the points on the circle representing the size of each component with the help of protractor and divide it into parts according to degrees of angle at the centre.

This diagram makes use of a circle, whose total area is divided into as many components as there are classes by drawing angles at the centre.

To illustrate the use of pie diagram, let us consider the data given in Table 1.5 pertaining to headache status of 200 males.

Table 1.12: Angles to be drawn at the centre of pie diagram

Headache Group	No. of Males (frequency)	Relative Frequency	Angle = 360 × rel. freq.
No headache	58	0.29	104.4
Simple headache	63	0.315	113.4
Unilateral headache	37	0.185	66.6
Migraine	42	0.21	75.6
Total	200	1	360

Relative frequency is obtained by dividing each frequency by the total (for example, 58 ÷ 200 = 0.29). The fourth column indicates the measures (in degrees) of the angle to be drawn for each class. Angle for any given class=360 x relative frequency. The following figure shows the representation of different headache groups by pie diagram. For example, 360 × 0.29 = 104.4).

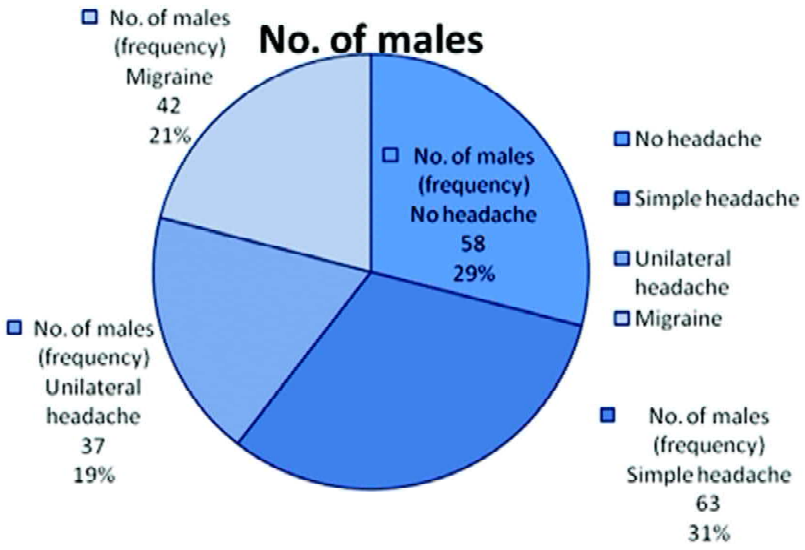


Fig. 1.7

1.5.4 Histogram

Histogram is most widely used method for graphical presentation of grouped frequency distribution by vertical bars against adjacent class intervals. The areas of vertical bars are proportional to the frequencies represented. Remember that in a bar diagram we consider the height of bars whereas in histogram we take into account the area of each bar.

When constructing histogram (when the class intervals are equal) the class intervals (class boundaries) are always taken on the X-axis and their frequencies on Y-axis. The adjacent rectangles are erected on each class interval with heights proportional to the frequency of that class. When the class-intervals are unequal, we take frequency density on Y-axis instead of frequency.

In Figure 1.8, you can see the histogram for the frequency distribution of age of 75 persons drawn on the basis of data given in Table 1.13

Table 1.13: Frequency distribution of 75 persons

Age in years Class interval	**Frequency (f)**
15-25	9
25-35	12
35-45	21
45-55	15
55-65	11
65-75	7
Total	75

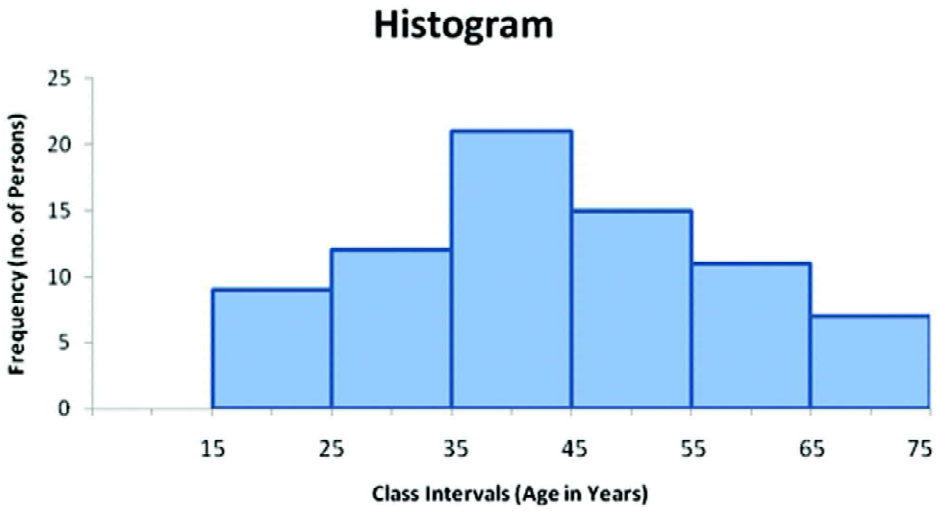


Fig. 1.8

1.6 ACTIVITIES

I) For each of the following variable, identify whether it is quantitative or qualitative.

a) Weight of babies born in a hospital during a year.

b) Gender of students of IGNOU.

c) Occupation of people residing in Delhi.

d) Number of students enrolled in B.Sc. in IGNOU.

II) Identify whether the following are discrete or continuous variables.

a) Area of different oceans in the world.

b) Number of petals in the flowers.

c) Payment made by 25 patients at city hospital for room charges and medicines.

III) The following are the number of accidents that occurred at 50 red lights in New Delhi during 1st week of August, 2011

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

i) Group these data into a frequency distribution.

ii) What number of accidents are less than 4?

iii) Draw a line diagram.

IV) Following are the data of hours worked by 50 workers for a period of a month in a certain factory.

112	170	138	108	140	146	87	103	94	42	130	25
167	128	187	145	195	122	162	79	61	154	148	61
175	179	160	151	112	148	98	161	182	123	188	71
123	156	83	111	133	146	143	157	135	58	95	28
131	120										

i) Construct a frequency distribution taking class interval 20.

ii) Draw a histogram.

V) The blood group of 20 persons are given below. Summarise the data using a frequency distribution.

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

Also draw bar diagram for the above data.

VI) The yields of different oil seeds are given in the following table.

Type of oil seed	**yield (in tonnes)**
Cotton	25
Olive	12
Almond	5
Sunflower	20
Peanut	8

Use the pie diagram to represent this data set.

Answers to Activities

I) (a) Quantitative, (b) Qualitative, (c) Qualitative, (d) Quantitative

II) (a) Continuous, (b) Discrete, (c) Continuous,

III)

i)

No. of accidents (x)	0	1	2	3	4	5	6
No. of Red lights (f)	9	10	10	6	8	3	4

ii) 35

iii) Draw the line diagram according to Figure 1.6

IV)

Class Interval	**Frequency**
40-60	2
60-80	2
80-100	5
100-120	5
120-140	12
140-160	10
160-180	10
180-200	4
Total	50

i) Draw the same as given in Figure 1.8

V)

Blood Group	No. of Persons
A	5
B	5
AB	3
O	7
Total	20

ii) Draw the same as given in Figure 1.5

VI) Draw the same as given in Figure 1.7

APPENDIX A

Random Number Table (generated by computer)

	1	2	3	4	5	6	7	8
1	22943	22529	22948	86451	56522	70994	90466	80248
2	37804	04602	89768	51618	64374	03707	32051	13427
3	09776	92391	43631	64045	79462	62309	86729	61219
4	95918	60512	88090	50835	56912	36936	14242	81144
5	21964	25277	22405	02626	17336	17828	69433	59794
6	23368	47822	91724	34136	22967	14049	19298	67750
7	04518	31856	00118	37920	05193	75380	24018	56273
8	73739	73577	56115	69183	24027	15648	81569	89796
9	31980	90205	50855	86264	33354	11695	72907	53613
10	18224	91613	75442	69469	65903	81742	49752	35555
11	64237	88472	63586	45037	94952	65514	98630	83002
12	28380	18060	39641	40664	80223	82711	02864	11242
13	59788	89771	49791	96817	16489	39716	14644	89874
14	63665	81537	51837	26361	22084	30182	53600	71015
15	06369	77886	31759	50653	94149	84925	49741	86877
16	48965	47830	42997	90187	64582	89666	99771	64067
17	73109	15867	72203	00801	93354	38504	60434	17069
18	49519	78978	01271	64372	67069	58378	09766	37387
19	70542	09994	93873	13297	42591	44672	65323	81459
20	95424	28017	88962	63721	94767	38863	20692	23015

1.7 SUMMARY

In this unit we discussed issues related to collection and presentation of data. In most cases, we resort to a sample survey rather than a census for data collection. There could be various methods of drawing a sample as described in unit. Whatever be the method, the sample should be representative of the population. Sample survey has several advantages: cost effective, lesser time, few personnel, and lower non-sampling errors. Sampling techniques to be followed should be decided keeping objectives of research in view. As far as possible random sampling methods should be pursued as it eliminates personal bias.

Data presentation involves frequency distribution and graphical presentation. Frequency distribution is somewhat similar to grouping of data. We can divide the data range into several class intervals and count the member of observations belonging to a particular class interval. Remember that frequency distribution helps us in better perception of the data distribution.

Data can be presented in the forms of bar diagram, line diagram or pie diagram. In pie diagram we can easily find out the percentage distribution of various components. Histogram is quite different from bar diagram: Bar diagram is one-dimensional (we consider only the height of a bar); Histogram is two-dimensional (we consider the area of a bar).

Suggested Reading

Goon, Gupta and Dasgupta. 1986. Fundamentals of Statistics. Delhi: World Press.

Kothari, C. R. 1985. *Research Methodology: Methods and Techniques*. Delhi: New Age International (P) Limited.

Nagar, A. L., and R K Dass. 1983. Basic Statistics. Delhi: Oxford University Press.

Sample Questions

1) Distinguish between the following terms

   a) Sample survey and census

   b) Parameter and statistic

   c) Sampling error and non-sampling error

   d) Stratified random sampling and systematic random sampling

2) Define the following terms:

   a) Sampling frame

   b) Population

   c) Random number table

   d) Snowball sampling

3) Write a brief note on the advantages and disadvantages of sampling.

UNIT 2 MEASURES OF CENTRAL TENDENCY AND DISPERSION

Contents

Learning Objectives

After going through this unit, you will be in a position to:

- explain the concepts of central tendency and dispersion;
- compute mean, median and mode from raw data as well as frequency distribution;
- compute mean deviation, variance, standard deviation and coefficient of variation from raw data and frequency distribution; and
- compute and interpret coefficient of correlation.

2.1 INTRODUCTION

In unit 1 of this block, we have explained how to present data in the form of tables and graphs. A more complete understanding of data can be attained by summarizing the data using statistical measures. The present unit deals with various measures of central tendency and dispersion in a variable. It also explains how to measure correlation between two variables. As the computation of these measures is different for ungrouped and grouped data, we present some measures for both ungrouped and grouped data.

2.2 MEASURES OF CENTRAL TENDENCY

The most commonly investigated characteristics of a set of data are measures of central tendency. Measures of central tendency provide us with a summary

that describes some central or middle point of the data. There are five important measures of central tendency, viz., i) arithmetic mean, ii) median, iii) mode, iv) geometric mean, and v) harmonic mean. Out of these, the last two measures, viz., geometric mean and harmonic mean, have very specific uses and thus less frequently used. Therefore, we will discuss the first three measures in this unit.

Remember that all these measures may not have the same value for a particular group of observations; because the formula is different for each measure. Which one of these measures, should be used in a particular case depends upon the type of data and the way in which the observations in the group cluster around a point.

Before dealing with these measures let us be familiar with certain notations which we will use. The standard notation is: *X,* which is a variable that takes values, $X_1, X_2, X_3 \ldots X_N$. Suppose we have data on number of children in a family obtained from a household survey of 40 households. The total number of children (*n*), as we know from our survey, is 40. We present these data in the form a frequency distribution such that 6 families have no child; five families have one child each, eight families have two children each and so on (see Table 2.1). Here the number of children is our variable 'X' and it takes seven values, viz, $X_1, X_2, X_3 \cdots X_7$ such that $X_1, = 0, X_2 = 1, X_3 = 2, \cdots\cdots X_7 = 6$. The corresponding frequency for each observation are: 6, 5, 8, 7, 6, 5 and 30. These are denoted as $f_1, f_2, \ldots, f_N$.

Many times we refer to a typical observation; it could be any of the observations under consideration. We call the typical observation as the 'i^{th} observation' and denote it as with corresponding frequency. Here '*i*' is the sub-script. For greater clarity we provide a range for the variable. In the current example '*i*' ranges between 0 and 6.

Table 2.1: Number of Children in families

No. of Children (x_i)	**No. of families (frequency)** f_i
0	6
1	5
2	8
3	7
4	6
5	5
6	3
Total	$\sum_{i=0}^{6} f_i = 40$

In the case of continuous variable we take the mid-values of class intervals as $X_1, X_2, X_3 \ldots X_n$ and the corresponding frequencies as $f_1, f_2, \ldots, f_N$. The sum of the frequencies is given below the table as $\sum_{i=0}^{6} f_i$. The symbol $\sum$ (read as sigma) is usually used to denote the sum of a variable. Adjacent

to $\sum$ are two numbers, i = 0 and 6 which denote the lower and upper range of the variable respectively. When there is no confusion in notations, we omit the subscripts and superscripts and just write $\sum f_i$ or simply $\sum f$.

2.2.1 Arithmetic Mean

The average or the arithmetic mean or simply the mean is the most commonly used measure of central tendency. It is computed by dividing the sum of all observations by the number of observations. It is denoted by $\bar{x}$ (read as 'x bar'). We explain the methods of computing arithmetic mean in the case of ungrouped data and grouped data.

In the case of ungrouped data

If the value of observations in the data is denoted by $x_1, x_2, ..., x_n$ then the arithmetic mean is given by

$$\bar{x} = \frac{x_1 + x_2 \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where n is the number of observations. In this formula, the Greek letter $(x_i \sum_{i=1}^{n})$ denotes summation of all the values , i.e., $\sum_{i=1}^{n} = x_1 + x_2 + \cdots + x_n$.

Example 2.1 Suppose we have the following data on the minimum temperature (°C) of New Delhi for 10 days.

19 17 21 11 15 17 12 17 15 18.

For finding the average temperature, we have total no. of observations = n =10,

Total of all these temperature $\sum x_i = 19 + 17 + 21 + \cdots + 18 = 162$

Therefore, the arithmetic mean $\bar{x} = \frac{\sum x_i}{n} = \frac{162}{10} = 16.2$

In the case of grouped data

In the case of grouped data we are provided with a frequency distribution. Let x_i (i = 1, 2, …, n) be the value of i^{th} observation in the data and it occurs with frequency f^i (i =1, 2, …., n). For the grouped frequency distribution the arithmetic mean is given by

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 \cdots + f_n x_n}{N} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

where $N = \sum_{i=1}^{n} f_i$ = Total no. of observations.

Remember that in the case of grouped frequency distribution, x_i is the mid value of the i^{th} class interval.

Example 2.2 Let us consider the data given in Table 1.3 of Unit 1 and compute the mean.

No. of Children (x_i)	Number of families (frequency) f_i	$f_i x_i$
0	6	0
1	5	5
2	8	16
3	7	21
4	6	24
5	5	25
6	3	18
Total	N=40	$\sum f_i x_i$ =109

Let us compute the arithmetic mean of the data given in the above table.

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{0\times 6 + 1\times 5 \cdots + 6\times 3}{40} = \frac{109}{40} = 2.725$$

Thus, the average number of children per family is 2.725

Example 2.3 Consider the data given in Table 1.7 of Unit 1 and compute the mean.

Table2.2

Age in years Class interval	frequency (f_i)
15-25	9
25-35	12
35-45	21
45-55	15
55-65	11
65-75	7
Total	**75**

For the computation of the mean, we have to construct the following table.

Table2.3

Class Interval	Mid value (x_i)	Frequency (f_i)	$f_i x_i$
15-25	20	9	180
25-35	30	12	360
35-45	40	21	840
45-55	50	15	750
55-65	60	11	660
65-75	70	7	490
Total		**N=75**	$\sum f_i x_i$ = 3280

Thus, the mean of the age of 75 persons is

$$\bar{x} = \frac{\sum f_i x_i}{N} = \frac{3280}{75} = 43.73 \approx 44 \text{ Years}$$

2.2.2 Median

Median is a positional average. Median is the middlemost value of the set of observations which divides the data set into two equal parts, where all the observations are arranged in either ascending or descending order. So there are 50 per cent observations below the median and the remaining 50 per cent are above the median.

Calculation of Median from Raw Data

For calculation of median from raw (unorganised) data you should take the following three steps.

a) Arrange the data either in ascending or in descending order of magnitude (both methods give the same value for median).

b) If there are odd number of observations (n), median is calculated by

$$\text{Median} = \text{value of } \left(\frac{n+1}{2}\right)^{th} \text{ observation}$$

where n = number of observations

c) If there are even numbers of observations, median is calculated by

$$\text{Median} = \frac{value\ of \left(\frac{n}{2}\right)^{th}\ observation + value\ of \left(\frac{n}{2}+1\right)^{th}\ observation}{2}$$

Example 2.4 The following is the data of the hemoglobin level of 11 women in gm/dL:

12.1 13.6 14.2 12.4 14.3 13.2 12.8 14.6 13.9 13.8 12.4

For finding the median of the above data, we arrange the values of hemoglobin in ascending order as follows:

12.1 12.4 12.4 12.8 13.2 13.6 13.8 13.9 14.2 14.3 14.6

Here, the number of observations are odd , since n = 11

Median hemoglobin level is given by

$$\text{Median} = \text{value of } \left(\frac{n+1}{2}\right)^{th} \text{ observation}$$

$$= \text{value of } \left(\frac{11+1}{2}\right)^{th} \text{ observation}$$

= value of (6^{th}) observation = 13.6 gm/dL (Arrange the above data in descending order and calculate the median. You should obtain the value, i.e., 13.6)

Example 2.5 The following is the data of the hemoglobin level of 12 women in gm/dL:

12.1 13.6 14.2 12.4 14.3 13.2 12.8 14.6 13.9 13.8 12.4 14.8

For finding the median of the above data, we arrange the values of hemoglobin in ascending order as follows:

12.1 12.4 12.4 12.8 13.2 13.6 13.8 13.9 14.2 14.3 14.6 14.8

Here, the number of observations are even , i.e., n = 12

Median hemoglobin level is given by

$$\text{Median} = \frac{value\ of\left(\frac{n}{2}\right)^{th}\ observation + value\ of\left(\frac{n}{2}+1\right)^{th}\ observation}{2}$$

$$= \frac{value\ of\left(\frac{12}{2}\right)^{th}\ observation + value\ of\left(\frac{12}{2}+1\right)^{th}\ observation}{2}$$

$$= \frac{value\ of\ (6)^{th}\ observation + value\ of\ (7)^{th}\ observation}{2}$$

$$= \frac{13.6+13.8}{2} = \frac{27.4}{2} = 13.7\ \text{gm / dL}$$

Calculation of Median from ungrouped frequency distribution

a) First of all, arrange the data in ascending or descending order of magnitude.

b) Next find the cumulative frequencies.

c) Apply the following formula:

$$\text{Median} = \text{value of}\ \left(\frac{N+1}{2}\right)^{th}\ observation$$

where N= $\sum f_i$ = Total number of observations

d) Finally, we will find the cumulative frequency which is either equal or just higher to $\frac{N+1}{2}$ and the value of the variable which corresponds to that cumulative frequency will be our required median.

Example 2.6 Consider the data given in Table 2.1 and calculate the median.

Table 2.4

No. of Children (x_i)	**No. of families (frequency) f_i**	**cumulative frequencies (less than type)**
0	6	6
1	5	11
2	8	19
3	7	26
4	6	32
5	5	37
6	3	40
Total	**N=40**	

In the above data set, the values on number of children are already in ascending order. In the third column of the table, we have calculated the cumulative frequencies of less than type then

Median = value of $\left(\frac{N+1}{2}\right)^{th}$ *observation*

= value of $\left(\frac{40+1}{2}\right)^{th}$ *observation*

= value of $(20.5)^{th}$ *observation*

Now the cumulative frequency which is either equal or just higher to 20.5 is 26 then the corresponding value of the variable is 3. Thus the median value for the above data set is 3 children per family.

Calculation of Median from Grouped Frequency Distribution

a) First of all, we find the value of $\frac{N}{2}$, where N = $\sum f_i$ = Total number of observations

b) Next, we calculate the cumulative frequencies and identify the class interval for which the cumulative frequency is either equal to or just higher than $\frac{N}{2}$. This class will contain the median and called the 'median class'.

c) Finally, we use the following formula to compute the median

Median = $L + \frac{\frac{N}{2} - c.f}{f} \times h$

where L = lower limit of median class

c.f. = cumulative frequency of the class preceding the median class

f = frequency of the median class

h = class interval of the median class

Example 2.7 for the data set given in **Table 2.2**, calculate the median age.

Table 2.5

Age in years Class interval	**frequency (f)**	**cumulative frequencies (less than type)**
15-25	9	9
25-35	12	21 = cf
35-45	21 = f	42
45-55	15	57
55-65	11	68
65-75	7	75
Total	**N=75**	

For the frequency distribution of age of 75 persons, $\frac{N}{2} = \frac{75}{2} = 37.5$

Therefore, the above table indicates that median would lie in the class interval 35-45. Thus

L = 35, C.f. = 21, f= 21, h = 45-35 = 10

Hence, the median is given by

$$\text{Median} = L + \frac{\frac{N}{2} - c.f}{f} \times h$$

$$\text{Median} = 35 + \frac{37.5 - 21}{21} \times 10$$

$$\text{Median} = 35 + \frac{16.5}{21} \times 10 = 42.86$$

The median age is 42.86 years.

2.2.3 Mode

Mode is the value of given data set which occurs maximum number of times, i.e., the value which has the highest frequency. Mode is the most commonly used measure of central tendency when we have to decide which is the most fashionable (most demanded or most preferable) item at this time. For example, to decide the most preferable size of shoes, clothes, etc., we find their mode.

Calculation of Mode from Raw Data

Example 2.8 Let us consider the temperature of 10 days in New Delhi, i.e.,

19, 17, 21, 11, 15, 17, 12, 17, 15, 18

In this data, the observation 17 is occurring maximum number of time (i.e., 3). Hence the mode is 17 (Note that Mode is 17, not 3).

Calculation of Mode from Ungrouped Frequency Distribution

Example 2.9 Consider the data given in Table 2.1, and find out the mode.

Table 2.6

No. of Children (x_i)	**No. of families (frequency) f_i**
0	6
1	5
2	8
3	7
4	6
5	5
6	3
Total	**N=40**

In this data set, the value 2 has the maximum frequency (, i.e., 8). Thus 2 is the most commonly occurring value. We can say that the modal value for the above data is 2 children per family.

Calculation of Mode for Grouped Frequency Distribution

i) First of all, we will find the class having maximum frequency which is called modal class.

ii) Then, we will calculate the mode by the following formula.

$$\text{Mode} = L + \frac{f_i - f_0}{2f_1 - f_0 - f_2} \times h$$

where L = lower limit of the modal class

$f_1 = frequency\ of\ the\ modal\ class$

$f_o = frequency\ of\ the\ class\ preceding\ the\ modal\ class$

$f_2 = frequency\ of\ the\ class\ succeeding\ the\ modal\ class$

$h = Class\ interval$

Example 2.10 Consider the data given in **Table 2.2** and find out the mode.

Table 2.7

Age in years Class interval	**frequency (f)**
15-25	9
25-35	12
35-45	21
45-55	15
55-65	11
65-75	7
Total	**N=75**

In the above data set, modal class is 35-45 since it has maximum frequency (i.e., 21).

We find that L = 35 $f_1 = 21, f_0 = 12, f_2 = 15, h = 45 - 35 = 10$, h = 45-35 = 10

Therefore, mode can be calculated as

$$Mode = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

$$= 35 + \frac{21 - 12}{2 \times 21 - 12 - 15} \times 10$$

$$= 35 + \frac{9}{15} \times 10 = 41$$

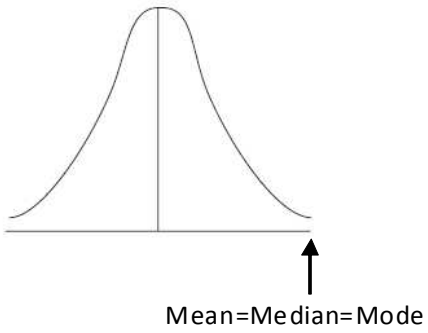
Thus the modal value for the above data is 41 years.

2.2.4 Relationship between Mean, Median and Mode

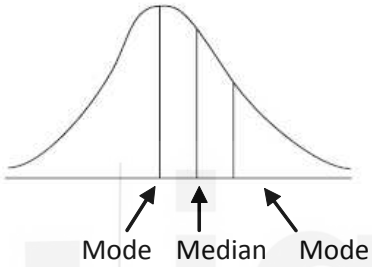
If a distribution is symmetric, values of mean median and mode coincide (as in the case of a normal distribution). If a distribution is moderately asymmetrical or skewed (positively or negatively) mean, median and mode have the following relationship given by Karl Pearson:

(Mean - mode) = 3(mean-median)

(a)



(b)



(c)

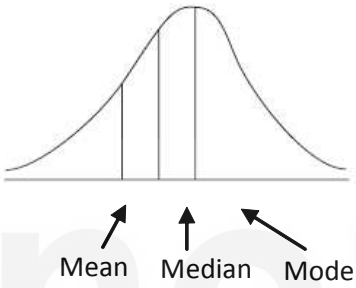


Figure 2.1: Relationship between mean, median, and mode in case of (a) Symmetric, (b) Positively skewed and (c) Negatively skewed distributions.

2.3 MEASURES OF DISPERSION

The various measures of central tendency discussed in previous section gives us an idea about the concentration of data around its central part. But these measures of central tendency cannot be used alone to describe the data. The following data on the marks of 3 students tells us that a single measure of central tendency cannot sufficient to describe the data and it is needed to use another measure called dispersion to get the complete idea about the entire data.

Table 2.8: Marks of the students A, B and C in 5 subjects

Subject	Student A	Student B	Student C
1	30	50	10
2	40	50	20
3	50	50	30
4	60	50	90
5	70	50	100
Total	250	250	250
Mean	50	50	50

You can observe from the above table, that the average marks of the students A, B and C are the same. After a thorough examination of the marks of all the students, we can find that the distribution of the marks of three students differ widely from one another. Marks in all the subjects vary widely from one another. So it is necessary to use other measures called measures of dispersion or variability (along with measures of central tendency) to get a complete idea about the distribution of the data.

In measures of dispersion, we measure the extent of scatterness or deviation to which all observations of the data varies from its central value. Measures of dispersion give an idea about homogeneity & heterogeneity of the distribution. In this section, we will discuss the commonly used measures of dispersion, viz., mean deviation, variance and standard deviation. The following are the commonly used measures of dispersion:

2.3.1 Mean Deviation

Mean deviation is the average of the absolute deviation of all the observations from its mean value. Here, first of all we compute the deviations of the data values from the mean. Secondly we, obtain the 'absolute values' for these deviations (it means you take you take only the numerical part of a number and ignore the minus sign). Finally, we calculate the average for these deviations. The mean deviation is also called the average deviation.

For Raw Data

If there are n observations say $x_1, x_2, ..., x_n$ of a variable under study and $\bar{x}$ is the mean of these n observations, then the mean deviation about mean is given by

$$M.D = \frac{1}{n}\sum_{i=1}^{n}\left|x_1 - \bar{x}\right|$$

Here $\left|x_1 - \bar{x}\right|$ (read as mod $\left(x_1 - \bar{x}\right)$) is the absolute value of the difference between x_i and $\bar{x}$. For finding the absolute value of a number, we ignore the minus signs. Thus (5) = 5.

Also (-5) = 5.

Example 2.11 Calculate the mean deviation from the following data of the hemoglobin level of 10 women in gm/dL:

12.1 13.6 14.2 12.4 14.3 13.2 12.8 14.6 13.9 13.9

For computing mean deviation, we will prepare the following table:

Table 2.9

x_i	$\left(x_i - 13.5\right)$	$\left\|x_i - 13.5\right\|$
12.1	-1.4	1.4
13.6	0.1	0.1
14.2	0.7	0.7
12.4	-1.1	1.1
14.3	0.8	0.8
13.2	-0.3	0.3
12.8	-0.7	0.7
14.6	1.1	1.1
13.9	0.4	0.4
13.9	0.4	0.4
$\sum x_i = 135$		$\sum\left\|x_i - \bar{x}\right\| = 7$

From the above table, n=10, $\sum x_i = 135$,

Then, $\bar{x} = \frac{\sum x_i}{n} = \frac{135}{10} = 13.5$

The mean deviation about mean is given by

$$M.D = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

$$= \frac{7}{10} = 0.7$$

Thus the mean deviation of the above data on hemoglobin level is 0.7 gm/dL.

For frequency distribution

Let x_i (i=1, 2, …, n) be the value of i^{th} observation in the data and it occurs with frequency fi,(i =1, 2, …., n). For the ungrouped frequency distribution the mean deviation about mean is given by

$$M.D = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Where $N = \sum f_i$

And $\left|x_i - \bar{x}\right|$ is the deviation from mean after ignoring the minus signs.

In case of grouped frequency distribution, we consider x_i as the mid value of i^{th} class interval.

Example 2.12 Calculate the mean deviation about mean from the data set given in Table 2.1

For the computation of the mean deviation, we have to construct the following table

Table 2.10

No. of Children (x_i)	**No. of families** f_i	$f_i x_i$	$x_i - 2.725$	$x_i - 2.725$	$f_i\left\|x_i - 2.725\right\|$
0	6	0	-2.725	2.725	16.350
1	5	5	-1.725	1.725	8.625
2	8	16	-0.725	0.725	5.800
3	7	21	0.275	0.275	1.925
4	6	24	1.275	1.275	7.650
5	5	25	2.275	2.275	11.375
6	3	18	3.275	3.275	9.825
Total	N=40	109.000			61.550

From the above table, $\bar{x} = \frac{\sum f_i x_i}{N} = \frac{109}{40} = 2.725$

In the fourth column we compute $\left(x_i - \bar{x}\right)$ and in the fifth column we compute $\left|x_i - \bar{x}\right|$.

Mean deviation about mean is given by

$$M.D = \frac{1}{N}\sum_{i=1}^{n} f_i \left|x_1 - \bar{x}\right|$$

$$= \frac{61.55}{40} = 1.539 \approx 1.54$$

Example 2.13 Calculate the mean deviation from the data given in table 2.3

Thus the mean deviation is 1.54 children per family

We will construct the following table for computation of the mean deviation:

Table 2.11

Mid values (x_i)	**frequency** (f_i)	$f_i x_i$	$(x_i - 43.733)$	$\lvert x_i - 43.733 \rvert$	$f_i \lvert x_i - 43.733 \rvert$
20	9	180	-23.733	23.733	213.600
30	12	360	-13.733	13.733	164.800
40	21	840	-3.733	3.733	78.400
50	15	750	6.267	6.267	94.000
60	11	660	16.267	16.267	178.933
70	7	490	26.267	26.267	183.867
Total	N=75	3280		90.000	913.600

From the above table, $\bar{x} = \frac{\sum f_i x_i}{N} = \frac{3280}{75} = 43.733$

Mean deviation about mean is given by

$$M.D = \frac{1}{N}\sum_{i=1}^{n} f_i \left|x_i - \bar{x}\right|$$

$$= \frac{913.6}{75} = 12.181\ years$$

2.3.2 Variance and Standard Deviation

The variance and the standard deviation are the most commonly and popularly used measures of dispersion. The average of the squared deviation from the mean is known as variance and it is denoted by σ^2 (read as sigma square). That

is, first of all we compute the deviations of the data values from the mean, then find the square of the values for these deviations and finally we find the average of these squared values.

The positive square root of the variance is called standard deviation. It is also known as root mean square deviation because it is the square root of the mean of the squared deviation from the arithmetic mean. It is denoted by (read as sigma). We can calculate variance and standard deviation as follows:

For Raw Data

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

We can rewrite it for computational convenience

$$\sigma^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

Or, $$\sigma^2 = \frac{1}{n}\sum x_i^2 - \left(\frac{\sum x_i}{n}\right)^2$$

And the standard deviation is given by

$$S.D = \sigma = +\sqrt{variance}$$

Remember that standard deviation is always positive.

For frequency distribution

When we have frequency distribution, we can calculate variance and standard deviation by following formulae:

Variance $\sigma^2 = \frac{1}{N}\sum f_i\left(x_i - \bar{x}\right)^2$

Or, $$\sigma^2 = \frac{1}{N}\sum f_i x_i^2 - \bar{x}^2$$

Or, $\sigma^2 = \frac{1}{N}\sum f_i x_i^2 - \left(\frac{\sum f_i x_i}{N}\right)^2$, where $N = \sum f_i$

$$S.D = \sigma = +\sqrt{variance}$$

The three formulae given above will provide the same result. Thus you can use any one of the above. For computation of variance we usually prepare a table from the given data as per our requirements. As mentioned earlier standard deviation is the positive square root of variance. Thus, in the case of grouped frequency distribution, we consider as mid value of the ith class interval.

Let us now consider the following examples to understand the computational method of variance and standard deviation:-

Example 2.14: Calculate the variance and standard deviation from the data set given in Example 2.11.

For computation of variance and standard deviation, we prepare the following table:

Table 2.12

x_i	x_i^2
12.1	146.41
13.6	184.96
14.2	201.64
12.4	153.76
14.3	204.49
13.2	174.24
12.8	163.84
14.6	213.16
13.9	193.21
13.9	193.21
$\sum x_i = 135$	$\sum x_i^2 = 1828.92$

From the above table, n=10, $\sum x_i^2 = 135$,

Then, $\bar{x} = \frac{\sum x_i}{n} = \frac{135}{10} = 13.5$

Variance is given by

$$\sigma^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

$$= \frac{1828.92}{10} - (13.5)^2$$

$$= 182.892 - 182.25 = 0.642$$

Standard deviation is given by

$$S.D = \sigma = +\sqrt{variance}$$

$$= \sqrt{0.642}$$

$$= 0.801$$

Example 2.15 Calculate the variance and standard deviation from the data set given in table 2.1

For the computation of variance and standard deviation, we have to construct the following table.

Table 2.13

No. of Children (x_i)	**No. of families** f_i	$f_i x_i$	x_i^2	$f_i x_i^2$
0	6	0	0	0
1	5	5	1	5
2	8	16	4	32
3	7	21	9	63
4	6	24	16	96
5	5	25	25	125
6	3	18	36	108
Total	N=40	$\sum f_i x_i = 109$	91	$\sum f_i x_i^2 = 429$

From the above table, $\bar{x} = \frac{\sum f_i x_i}{N} = \frac{109}{40} = 2.725$

Variance is given by

$$\sigma^2 = \frac{1}{N}\sum f_i x_i^2 - \bar{x}^2$$

$$= \frac{429}{40} - (2.725)^2$$

$$= 10.725 - 7.426$$

$$= 3.299$$

Standard deviation is given by

$$S.D = \sigma = +\sqrt{variance}$$

$$= \sqrt{3.299}$$

$$= 1.816$$

Example 2.16 Calculate the variance and standard deviation from the data given in table 2.3

We construct the following table for the computation of variance and standard deviation:

Table 2.14

Class Interval	**Mid values** (x_i)	**frequency** (f_i)	$f_i x_i$	x_i^2	$f_i x_i^2$
15-25	20	9	180	400	3600
25-35	30	12	360	900	10800
35-45	40	21	840	1600	33600
45-55	50	15	750	2500	37500
55-65	60	11	660	3600	39600
65-75	70	7	490	4900	34300
Total		N=75	$\sum f_i x_{i=}$ 3280	13900	$\sum f_i x_i^2 = 159400$

From the above table, $\bar{x} = \frac{\sum f_i x_i}{N} = \frac{3280}{75} = 43.733$

Variance is given by

$$\sigma^2 = \frac{1}{N}\sum f_i x_i^2 - \bar{x}^2$$

$$= \frac{159400}{75} - (43.7333)^2$$

$$= 2125.333\text{-}1912.604$$

$$=212.729$$

Standard deviation is given by

$$S.D = \sigma = +\sqrt{variance}$$

$$= \sqrt{212.729}$$

$$=14.585$$

Note: To get an unbiased estimate of population variance from sample, we divide the quantity $\sum (x_i - \bar{x})^2$ by (n-1) instead of by n. and this is denoted by s^2. Thus the formula to compute the sample variance is

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

Or $s^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right)$

2.3.3 Coefficient of Variation (C.V.)

As, the measures of central tendency and measures of dispersion specify the characteristics of a data set. A limitation with these two measures is that they are not free from the unit of measurement. For example if I take height instead of inches in centimeters then I get different values of mean and standard deviation. In order to avoid this problem we often use the coefficient of variation.

When we want to compare two or more data sets in respect to variability then we will use coefficient of variation. The coefficient of variation is also useful even in comparison of data sets having different measurement units because it is a unit free measure. It is given by

Coefficient of variation $= \frac{S.D}{mean} \times 100$

Or, $C.V. = \frac{\sigma}{\bar{x}} \times 100$

The data set for which coefficient of variation is less is said to be more consistent or more uniform or more homogeneous. For the above examples we can calculate the coefficient of variation as:

For example 2.14, $C.V. = \frac{\sigma}{\bar{x}} \times 100 = \frac{0.801}{13.5} \times 100$

$= 5.93\ \%$

For example 2.15, $C.V. = \frac{\sigma}{\bar{x}} \times 100 = \frac{1.816}{2.725} \times 100$

$= 66.64\ \%$

For example 2.16, $C.V. = \frac{\sigma}{\bar{x}} \times 100 = \frac{14.585}{43.733} \times 100$

$= 33.35\ \%$

Example 2.18: The following data gives the means and standard deviations of the marks of two students in MA (Anthropology) examination.

	Student A	Student B
Mean ($\bar{x}$)	60	70
Standard Deviation (σ)	11	10

Which student is the better performer in the examination?

To find a better performer, we will calculate their coefficient of variations. The student having less coefficient of variation will have better performance in the examination.

Coefficient of variation of student A is given by

$$CV_A = \frac{\sigma_A}{\bar{x}_A} \times 100$$

$$= \frac{11}{60} \times 100$$

$$= 18.33\ \%$$

Coefficient of variation of student B is given by

$$CV_B = \frac{\sigma_B}{\bar{x}_B} \times 100$$

$$= \frac{10}{70} \times 100 = 14.29\ \%$$

Since the coefficient of variation of student B is less than that of student A so student B is the better performer in the examination.

2.4 CORRELATION

So far we have dealt with a single characteristic of data. But, there may be cases when we would be interested in analyzing more than one characteristic at a time. For example, you may like to study the relationship between the age and the number of books a person reads. Such data, having two characteristics under study are called bivariate data. One of the measures to find out the extent or degree of relationship between two variables is correlation coefficient.

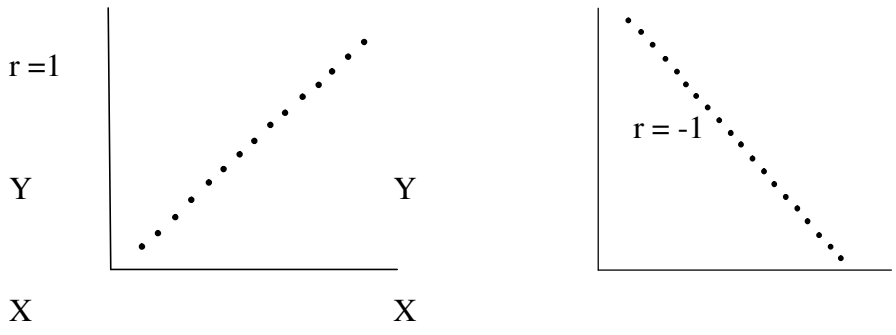
An analysis of the covariation of two or more variables is usually called correlation. If two characteristics vary in such a way that movement in one is accompanied by movement in the other, these characteristics are correlated. For example, there are relationships between age and blood pressure of individuals, the price and demand of a product, the height and weight of a person, the number of hours devoted in study and performance in the examination etc. are some examples of correlated variables. Correlations coefficient measures the strength and direction of the relationship between two variables. The value of correlation coefficient (r) remains between -1 and +1. A positive value of r indicates a positive relationship and negative value indicates a negative relationship.

In order to have a rough idea about the nature of relationship between two variables we plot the data on graph paper, called the scatter plot or scatter diagram. In the case of quantitative variables we can however have a unique value of the relation in the form of **Karl Pearson's Coefficient of Correlation**. In the case of ordinal data where ranks only are available we use **Spearman's rank correlation** method to obtain the degree of relationship.

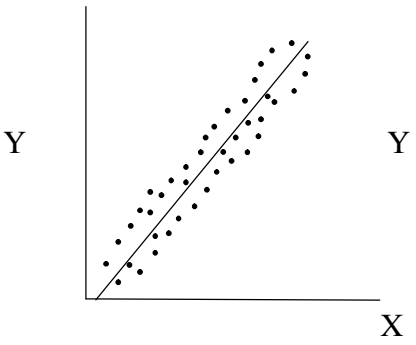
a) **Scatter Diagram**

If we are interested in finding out the relationship between two variables, the simplest way to visualize it is to prepare a dot chart called scatter diagram. Using this method, the given data are plotted on a graph paper in the form of dots. For example, for each pair of X and Y values, we put a dot and thus obtain as many point as the number of observations. Now, by looking into the scatter of various dots, we can ascertain whether the variables are related or not. The greater the scatter of the plotted points on the chart, the lesser is the relationship between the two variables. The more closely the points come to a straight line, the higher the degree of relationship.

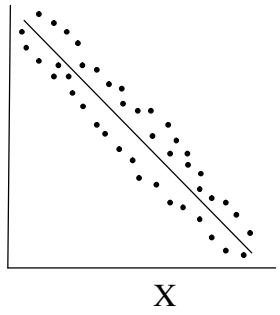
The following figures show the different types of Correlation.



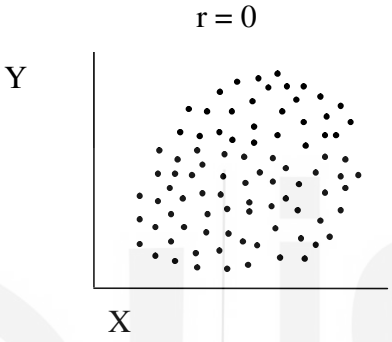
Perfect Positive Correlation Perfect Negative Correlation



High Degree Positive Correlation



High Degree Negative Correlation



No Correlation

b) **Karl Pearson's Coefficient of Correlation**

Let X and Y be the two variables representing two characteristics which are known to have some meaning full relationship.

The Karl Pearson's coefficient of correlation is given by

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

Or, $$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

The method of computation of correlation coefficient will be more clear by the following example.

Example 2.19: Following are the heights and weights of 10 students.

Table 2.15

Heights (in inches)	70	61	73	67	58	65	71	65	63	60
Weight (in kgs)	64	50	64	66	50	54	60	61	54	55

i) Make a Scatter diagram

ii) Find the correlation coefficient between height and weight

In the following figure, we considered height on X-axis and weight on Y axis then plotted the corresponding points.

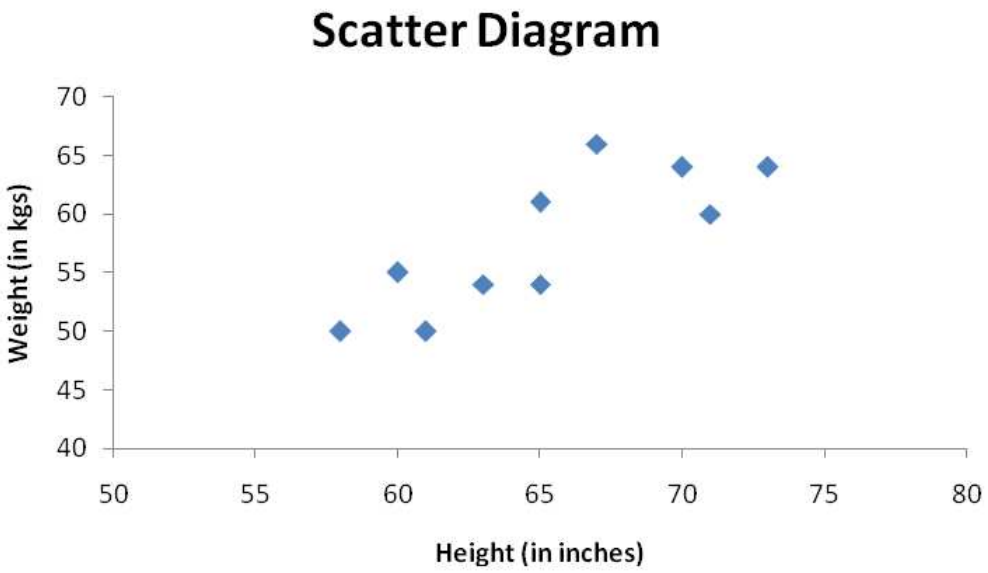


Fig. 2.1

By looking at the scatter diagram, we can say that the height and weight are correlated. It is clear from the above diagram that correlation is positive because the points are in upward rising from the lower left hand corner to the upper right hand corner and all the points are close to a line, so there is a high degree positive correlation.

For calculating Karl Pearson's Correlation Coefficient, we will construct the following table:

Table 2.16

Height (x_i)	**Weight** (y_i)	x_i^2	y_i^2	$x_i y_i$
70	64	4900	4096	4480
61	50	3721	2500	3050
73	64	5329	4096	4672
67	66	4489	4356	4422
58	50	3364	2500	2900
64	54	4225	2916	3510
71	60	5041	3600	4260
65	61	4225	3721	3965
63	54	3669	2916	3402
60	55	3600	3025	3300
$\sum x_i = 653$	$\sum y_i = 578$	$\sum x_i^2 = 42863$	$\sum y_i^2 = 33726$	$\sum x_i y_i = 37961$

Here, n=10

$$\bar{x} = \frac{\sum x_i}{n} = \frac{653}{10} = 65.3$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{578}{10} = 57.8$$

The Karl Pearson's coefficient of correlation is given by

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

$$r = \frac{(37961) - 10(65.3)(57.8)}{\sqrt{(42863) - 10(65.3)^2}\sqrt{(33726) - 10(57.8)^2}}$$

$$= \frac{(37961) - (37743.4)}{\sqrt{222.1}\sqrt{(317.6)}}$$

$$= \frac{(217.6)}{(14.903) \times (17.821)} = 0.819$$

c) **Spearman's Rank Correlation**

This is denoted by ρ (read as 'rho') instead of 'r'. Here the raw data are converted to their ranks. For example, suppose two examiners rank individual students in a class according to their performance in a viva voce test. It may so happen that both examiners will assign different ranks to a particular student. If there is too much difference in ranks assigned by both the examiners, then the evaluation of students may not be appropriate. Thus we need to study the relationship between the ranks assigned by the examiners and the degree of relationship will judge how appropriate the evaluation process has been. There could be several similar situations where rank correlation can be applied.

In rank correlation method we take into account the difference in ranks assigned to an observation. By considering such difference in ranks for all observations we arrive at the rank correlation coefficient. The formula for rank correlation is given by

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference in ranks assigned to an observation.

The Spearman's rank correlation also ranges from +1 to -1. Thus, positive values indicate direct relationship between the variables, while negative values indicate inverse relationship. The value $\rho = 0$ indicates absence of association between the variables.

Example 2.20: Given below are the ranks assigned by two examiners, A and B, to a group of 10 students. Find out the degree of relationship between ranks assigned by the examiners.

We prepare a table as given below and find out the difference in ranks assigned by the examiners.

Student	rank by A	rank by B	d_i	d_i^2
1	1	1	0	0
2	2	6	-4	16
3	3	8	-5	25
4	4	7	-3	9
5	5	10	-5	25
6	6	9	-3	9
7	7	3	4	16
8	8	5	3	9
9	9	2	7	49
10	10	4	6	36
total				194

Next, we apply the formula $\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$

We find the value ρ = -0.175757575

Thus we can say that the Spearman's rank correlation in the above case is -0.18 (approx).

2.6 CONCEPT OF REGRESSION

In regression analysis we have two types of variables: i) dependent (or explained) variable, and ii) independent (or explanatory) variable. As the name (explained and explanatory) suggests the dependent variable is explained by the independent variable. Note that correlation coefficient does not reflect cause and effect relationship whereas regression analysis assumes that one variable (or more than one) is the cause and other is the effect.

In the simplest case of regression analysis there is one dependent variable and one independent variable. Let us assume that consumption expenditure of a household is related to the household income. For example, it can be postulated that as household income increases, expenditure also increases. Here consumption expenditure is the dependent variable and household income is the independent variable.

Usually we denote the dependent variable as Y and the independent variable as X. Suppose we took up a household survey and collected n pairs of observations in X and Y. The next step is to find out the nature of relationship between X and Y.

The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equation. The

simplest of these equations is the linear equation. This means that the relationship between X and Y is in the form of a straight line and is termed linear regression. When the equation represents curves (not a straight line) it is called non-linear regression.

Now the question arises, 'How do we identify the equation form?' There is no hard and fast rule as such. The form of the equation depends upon the reasoning and assumptions made by us. However, we may plot the X and Y variables on a graph paper to prepare a scatter diagram. From the scatter diagram, the location of the points on the graph paper helps in identifying the type of equation to be fitted. If the points are more or less in a straight line, then linear equation is assumed. On the other hand, if the points are not in a straight line and are in the form of a curve, a suitable non-linear equation (which resembles the scatter) is assumed.

Regression analysis can be extended to cases where one dependent variable is explained by a number of independent variables. Such a case is termed 'multiple regression'.

You may by now be wondering why the term 'regression', which means 'reduce'. This name is associated with a phenomenon that was observed in a study on the relationship between the stature of father (x) and son (y). It was observed that the average stature of sons of the tallest fathers has a tendency to be less than the average stature of these fathers. On the other hand, the average stature of sons of the shortest fathers has a tendency to be more than the average stature of these fathers. This phenomenon was called regression towards the mean. Although this appeared somewhat strange at that time, it was found later that this is due to natural variation within subgroups of a group and the same phenomenon occurred in most problems and data sets. The explanation is that many tall men come from families with average stature due to vagaries of natural variation and they produce sons who are shorter than them on the whole. A similar phenomenon takes place at the lower end of the scale.

The simplest relationship between X and Y could perhaps be a linear deterministic function given by

$$Y_i = a + bX_i \qquad \ldots(2.1)$$

In the above equation X is the independent variable or explanatory variable and Y is the dependent variable or explained variable. You may recall that the subscript i represents the observation number, i ranges from 1 to n. Thus Y_i is the first observation of the dependent variable, X_i is the fifth observation of the independent variable, and so on.

Equation (2.1) implies that Y is completely determined by X and the parameters a and b. Suppose we have parameter values a = 3 and b = 0.75, then our linear equation is Y = 3 + 0.75 X. From this equation we can find out the value of Y for given values of X. For example, when X = 8, we find that Y = 9. Thus if we have different values of X then we obtain corresponding Y values on the basis of (2.1).

Linear Regression

Let us consider the following data on the amount of rainfall and the agricultural production for ten years.

Rainfall (in mm.)	Agricultural production (in tonne)	Rainfall (in mm.)	Agricultural production (in tonne)
60	33	75	45
62	37	81	49
65	38	85	52
71	42	88	55
73	42	90	57

We assume that rainfall is the cause (X) and agricultural production is the effect (Y). We plot the data on a graph paper. The scatter diagram looks something like Fig. 2.2. We observe from Fig. 2.2 that the points do not lie strictly on a straight line. But they show an upward rising tendency where a straight line can be fitted.

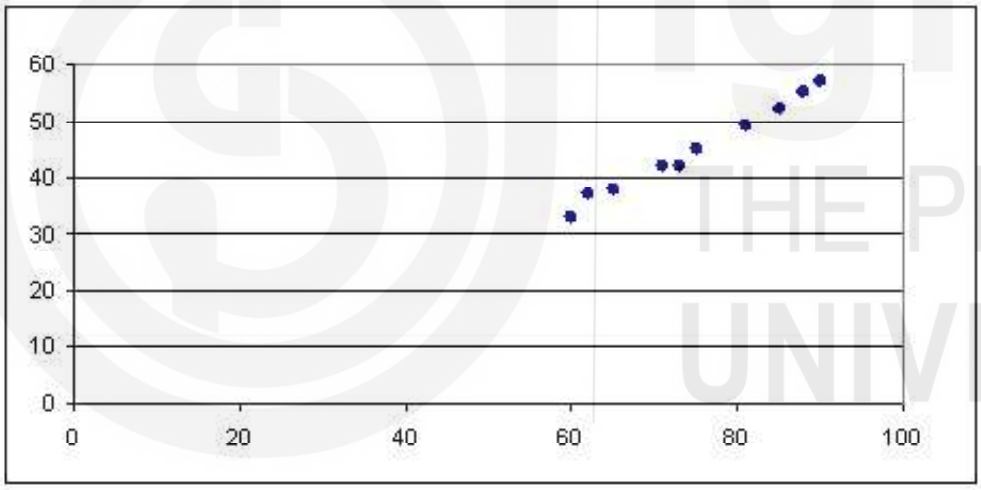


Fig. 2.2: Scatter Diagram

When we fit a straight line to the data there is some sort of error we are committing – the observations are not on a straight line but we are forcing a straight line. The vertical difference between the regression line and the observations is the 'error'. Our objective is to minimize the error values. This is usually done by the method of 'least squares'. We will not go into the details of the method here. Instead, two equations derived on the basis of least squares method and known as normal equations are given below.

These are:

$$\sum Y = na + b\,X \qquad \ldots(1)$$

$$\sum XY = a \sum X + b \sum X^2 \qquad \ldots (2)$$

From our sample survey we have data on X and Y variables; we also know the number of observations (n). The unknowns in the above two equations are 'a' and 'b'; we estimate these two values.

Example 2.21: Estimate the regression equation from rainfall data given above.

We apply the normal equations to the rainfall data. For that purpose we prepare a table as given below.

Table 9.2: Computation of Regression Line

X_i	Y_i	X_i^2	X_iY_i	$\overline{Y}_i$
60	33	3600	1980	33.85
62	37	3844	2294	35.34
65	38	4225	2470	37.57
71	42	5041	2982	42.03
73	42	5329	3066	43.51
75	45	5625	3375	45.00
81	49	6561	3969	49.46
85	52	7225	4420	52.43
88	55	7744	4840	54.66
90	57	8100	5130	56.15
Total $\sum_i X_i = 750$	$\sum_i Y_i = 450$	$\sum_i X_i^2 = 57294$	$\sum_i X_iY_i = 34526$	$\sum_i \overline{Y}_i = 450$

We obtain the normal equations as

450 = 10a + 750b

34526 = 750a + 57294b

By solving these two equations we obtain the values a = -10.73 and b = 0.743

Thus the estimated regression equation is

$$\overline{Y}_i = -10.73 + 0.743X_i$$

Multiple Regression

In many cases you have more than one independent variables which together explain the dependent variable. This sort of models are termed 'multiple regression'. A typical example of a multiple regression is $Y = a + bX_1 + cX_2$.

In the above equation Y is the dependent variable while X_1 and X_2 are independent variables.

These days many statistical software are available which can compute the estimates for you. Once you have the estimates, you can formulate the regression equation.

2.6 SUMMARY

In this unit we discussed the methods of presentation of data, particularly measures of central tendency and dispersion. These measures are summary statistics of the dataset. In addition to the above, we dealt with correlation and regression also. These two indicate the relationship between two variables. Correlation coefficient is a summary statistic of the strength of relationship between two variables. A major limitation of correlation coefficient is that it does not show cause and effect relationship; it just says that both the variable move together – either in the same direction (positive correlation) or in opposite directions (negative correlation). Regression analysis shows cause and effect relationship – changes in the independent variable causes changes in the dependent variable. These measures help us in interpretation of data.

Suggested Reading

Kothari, C. R. 1985. *Research Methodology: Methods and Techniques.* Delhi: New Age International (P) Limited.

Nagar, A.L. and R.K. Dass, 1983, Basic Statistics, Oxford University Press, Delhi.

Sundar Rao, P.S.S. and Richard, J. 1996. An Introduction to Biostatistics. New Delhi: Prentice-Hall of India.

Sample Questions

1) Consider the following data set.

   91 83 60 58 73 48 79 85 92 80.

   On the basis of the above data

   i) Calculate mean, median and mode.

   ii) Calculate mean deviation, standard deviation and variance.

   iii) Compute coefficient of variation.

2) The following are the number of injured persons in 50 accidents that took place in New Delhi during 1st week of August.

No. of injured persons (x)	0	1	2	3	4	5	6
No. of accidents (f)	9	10	10	6	8	3	4

   i) On an average how many persons were injured in an accident?

   ii) Calculate the median and mode number of accidents.

   ii) Calculate mean deviation, standard deviation and variance of number of injured persons.

   iii) Find the coefficient of variation of the above data.

3) Following are the data of hours worked by 50 workers for a period of a month in a certain factory.

Hours worked (class interval)	Number of workers (Frequency)
40-60	2
60-80	2
80-100	5
100-120	5
120-140	12
140-160	10
160-180	10
180-200	4
Total	50

i) Calculate mean, median and mode hours worked.

ii) Calculate mean deviation, standard deviation and variance of the hours worked.

iii) Find coefficient of variation of hours worked.

4) Following are the data of the hours worked by two workers for seven days in a factory.

Worker A	8	5	7	4	6	9	5
Worker B	2	8	4	3	7	6	5

i) Find the average hours of work done by both workers.

ii) Which worker is more consistent (hint: the worker with less variance)?

5) Ten persons were advised by their physicians to lose weight for health reasons. They enrolled in a special weight loss program. The following table gives the time spent in the program (in days) and weight lost after completion of the program (in kg.).

Time Spent(x)	25	39	12	30	52	41	67	92	10	11
Weight Loss (y)	12	18	5	14	20	17	25	47	7	6

i) Present a scatter plot for the data.

i) Compute correlation coefficient between number of days enrolled and weight lost.

Answers and Hints

1) i) 74.9, 79.5, mode does not exist

ii) 12.12, 14.13, 199.69

iii) 18.87%

2) i) 2.38

ii) median = 2. The data is bi-modal (1 and 2)

iii) 1.56, 1.83, 3.35

iv) 76.89%

3) i) 135.2, 138.33, 135.56

ii) 28.608, 35.51, 1260.96

iii) 26.27%

4) i) $\bar{x}_A = 6.29, \bar{x}_B = 5$

ii) CV_A=26.50%, CV_B= 40% and worker A is more consistent

5) i) Scatter diagram can be plotted in the same manner as in Example 2.19

ii) 0.9704

UNIT 3 STATISTICAL DISTRIBUTIONS AND INFERENCE

Structure

Learning Objectives

After going through this unit, you should be in a position to:

- explain the normal and standard normal distributions and their applications;
- explain the difference between concepts of type I and type II errors;
- explain the difference between one-tailed and two-tailed tests;
- judge which test can be used in a given problem;
- draw inferences about the problems related to t & x^2 tests.
- apply ANOVA for both one-way and two-way classified data; and
- drawinference by using appropriate test.

3.1 INTRODUCTION

As mentioned in Unit 1 we undertake a sample survey instead of complete census of the population concerned because of certain constraints. These constraints could be availability of money, manpower and time. After collection of data through questionnaire, interview or participatory observation method we follow certain steps such as tabulation, presentation and analysis of data.As you know, we can present data in the form of tables and graphs. Also data can be put to various statistical analyses. Thus we can find out i) measures of central tendency such as mean, median and mode, ii) measures of dispersion such as variance and standard deviation, and iii) correlation and regression coefficients. We have discussed these issues in Unit 2.

Many a time we conduct random experiments and the outcome can be considered as a random variable. When we select a random sample, there is an element of probability or chance attached to each unit. Thus the characteristics of a random sample, such as arithmetic mean, can be considered as a random variable. The probability with which the various outcomes of a random variable take place can be modelled in line with certain theoretical distribution available

in literature. In the present unit we will discuss some of the important theoretical distributions such as normal distribution, t-distribution and chi-square (read as 'ki-square') distribution.

Recall that the objective of our study is to analyse the behaviour of the population or the universe, not the sample. In order to make things feasible we are studying the sample and hence whatever results we have got are based on sample information. Naturally a question arises: Are the sample results valid for the population? In other words, can we draw inferences on the basis of sample results? We will deal with hypothesis testing and apply some of the statistical tests in this unit.

3.2 NORMAL DISTRIBUTION

The normal distribution perhaps is the most frequently used probability distribution. Its usefulness comes from the fact that many variables in real life are closer to normal distribution. A normal random variable X is completely defined by its two parameters μ (mean) (pronounced 'mu') and σ^2 (variance) (pronounced 'sigma-squared'). It is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ where } -\infty < x < \infty$$

If Xis normally distributed with mean μ and variance σ^2 then we write it as X $X \sim N(\mu,\sigma^2)$ and read as X is normally distributed with mean μ and variance σ^2. Normal distribution is a continuous distribution in the sense that the random variable describing normal distribution is continuous. Thus the probability of the random variable taking values within a specified range can be found out from the normal distribution. Thus, if we know the mean and standard deviation of a variable which follows normal distribution, we can find out the probability of X taking values between any two points. The probability of the normal distribution is described by the area under the curve.

3.2.1 Major Characteristics of Normal Distribution and Normal Curve

In Fig. 3.1 below we measure the variable 'X' on the x-axis and the probability with which a particular value of 'x' takes place is on the y-axis.

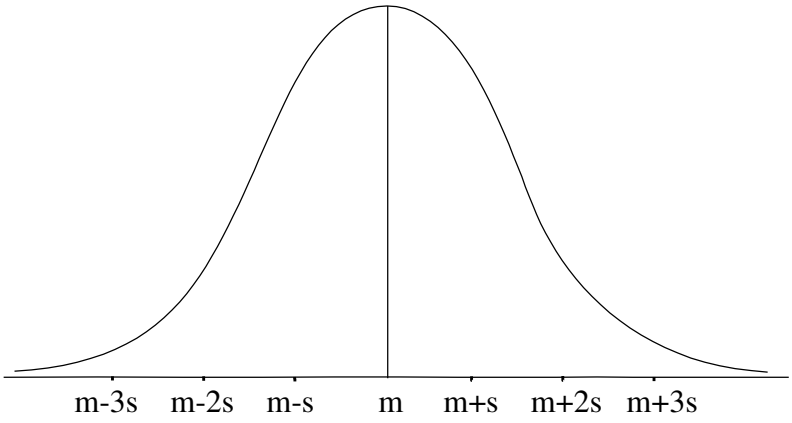


Fig. 3.1

From Figure 3.1 we observe that

1) The normal curve is bell shaped curve and is symmetrical about mean
2) Mean, Median & Mode of normal distribution are equal for normal distribution, i.e., Mean = Median = Mode
3) It is described by mean μ and standard deviation σ.
4) The random variable ranges between $-\infty$ *and* ∞.

Importance of Normal Distribution

Importance of normal distribution may be pointed out in the following heads.

1) Most of the practical characteristics/variables follow normal distribution, e.g., height, weight, biological characteristics, births in a country over a number of years.
2) Most of discrete distributions follow normal distribution as size of the population tends toinfinity.
3) It is user friendly.

3.2.2 Standard Normal Distribution

There could be different combinations of mean and standard deviation for different normal variables. For each case, therefore, we need to draw a different curve and find out the area under the curve. In order to simplify things, we subtract mean and divide it by its standard deviation. In this method we obtain a random variable which has zero mean and unit (means – one) standard deviation. Such a random variable is called standard normal variable (snv). Its probability density function (p. d. f.) is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad \text{where} -\infty < z < \infty$$

The standard normal variable has some interesting area properties:

1) The total area under an snv is 1.00
2) For a snv, 68.2 per cent area remains within a range of one sigma. In symbols we can write it as P[μ-$\sigma \prec X \prec \mu+\sigma$]=P[-1<*Z*<1]=68.27% =0.6827(see Figure 3.2)
3) For a snv, 95.44 per cent area remains within a range of two sigma. In symbols, we can write it as P[$\mu - 2\sigma < X < \mu + 2\sigma$]=P[-2<*Z*<2]=95.44% =0.9544 (see Figure 3.3)
4) 4 For a snv, 99.73 per cent area remains within a range of three sigma. In symbols we can write it as P[$\mu - 3\sigma < X < \mu + 3\sigma$]=P[-3<*Z*<3]=99.73% =s0.9973 (see Figure 3.4)

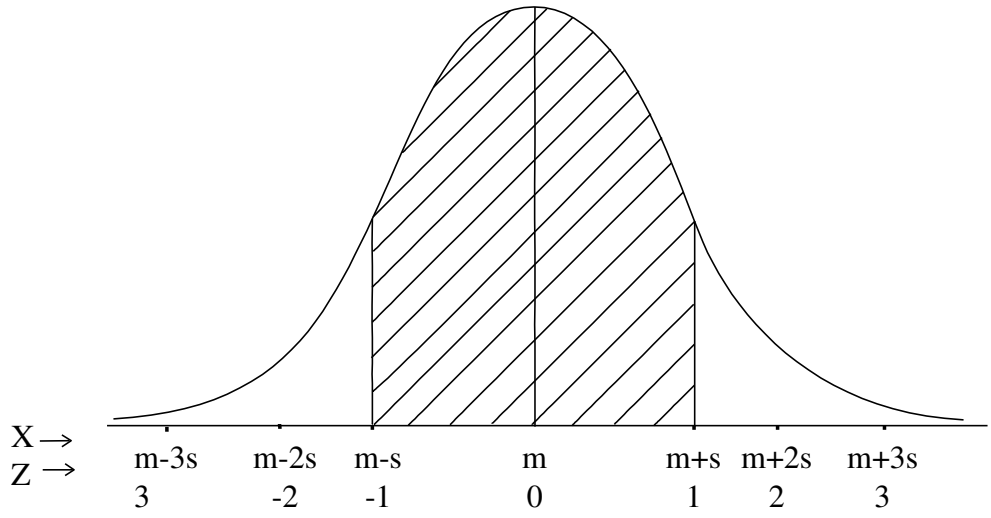


Fig. 3.2

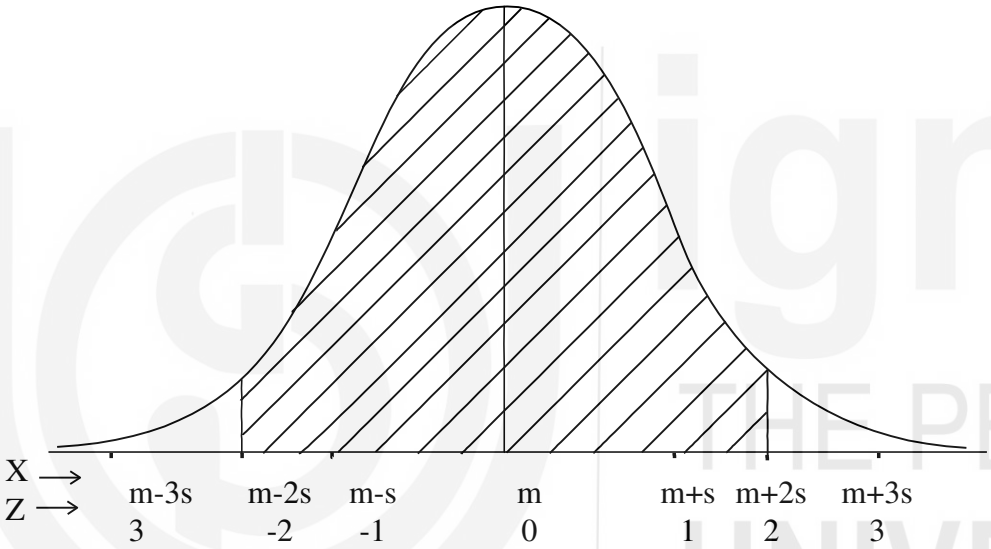


Fig. 3.3

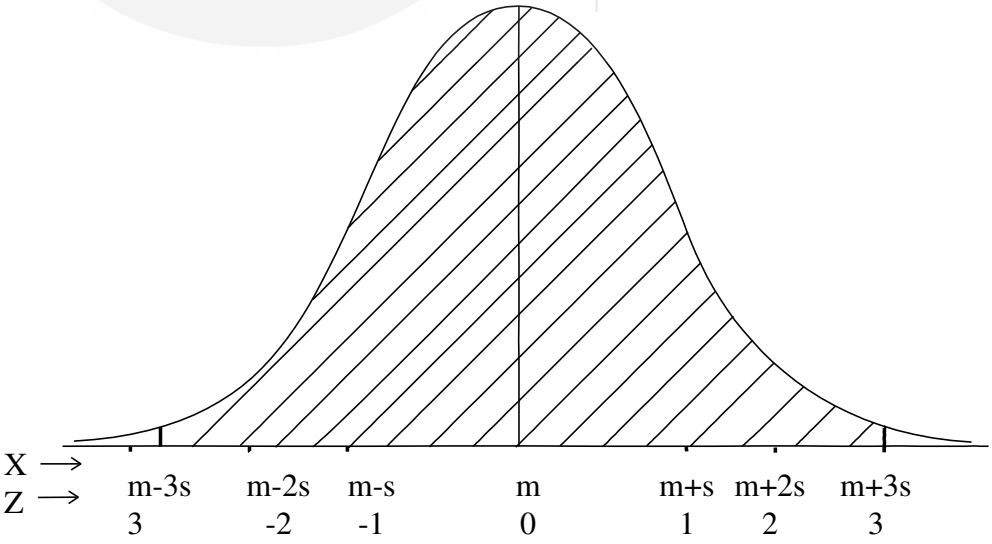


Fig. 3.4

Thus even though and snv can vary between- $-\infty < z < \infty$,we see that 99.73% observations are within 3σ plus and minus the mean.The area under snv is calculated by statisticians and is available to us in various forms – appendix table to books on statistics, Internet, etc. We can refer to these tables and find out the area for any specified range.

Let us discuss some simple applications of standard normal distribution.

Example 3.1: Assume that the weight of students in a university is normally distributed. The mean weight of 1000 students is 60 kg & standard deviation is 16 kg. Find the number of male students having their weights.

a) less than 55 kg, b) more than 70 kg, and c) between 45 kg& 65 kg.

Solution: a) In such a problem we have to apply the snv and find out the area under the curve. Let the random variable X denote the weights of the male students of the university.

We are given that

N = 1000, μ = 60 kg, σ = 16 kg

i.e., X ~N (60,256)

If Z= $\frac{x-\mu}{\sigma}$ then we know that Z ~ N (0, 1). Let us solve each part of the above problem one by one.

For X = 55, $Z = \frac{55-60}{16} = -0.3125$

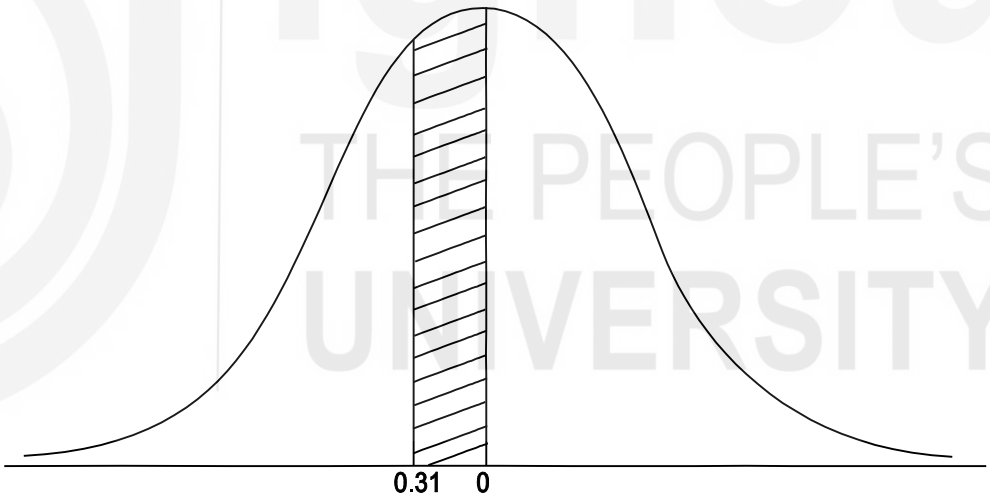


Fig. 3.5

P(X<55) = P(Z<-0.31) = P(Z>0.31)

$$= 0.5\text{-P } (0<Z<0.31) \left[\begin{matrix} since, Area\ of\ on\ the\ both\ side\ of\ the \\ z = 0\ is\ 0.5 \end{matrix} \right]$$

$$= 0.5\text{-}0.1217 \left[from\ the\ table\ of\ areas\ under\ Normal\ Curve \right]$$

$$= 0.3783$$

$\therefore$ Number of male students having weight less than 55 kg = N × P(X<55)

$$= 1000 \times 0.3783 = 378$$

Thus, out of the 1000 students in the university, 378 students have weight less than 55kg.

b) For X = 70, $Z = \frac{70-60}{16} = 0.625 \approx 0.63$

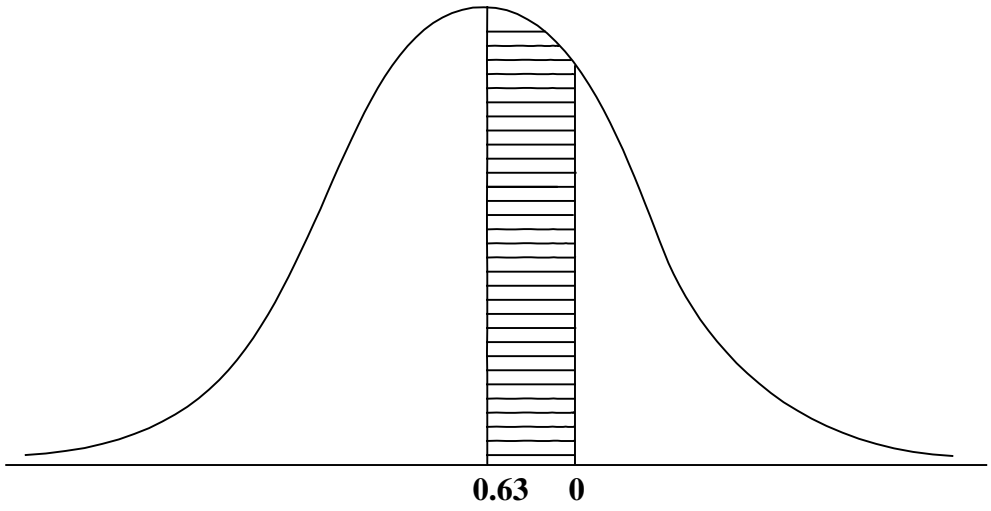


Fig. 3.6

P(X>70) = P (Z>0.63)

= 0.5-P (0<Z<0.63)

= 0.5-0.2357

= 0.2643

$\therefore$ Number of male students having weight more than 70 kg = N P(X>70)

=1000×0.2643 = 264

Thus, out of the 1000 male students in the university, 264 students have weight of more than 70 kg.

c) For X = 45, $Z = \frac{45-60}{16} = -0.9375 \approx 0.94$

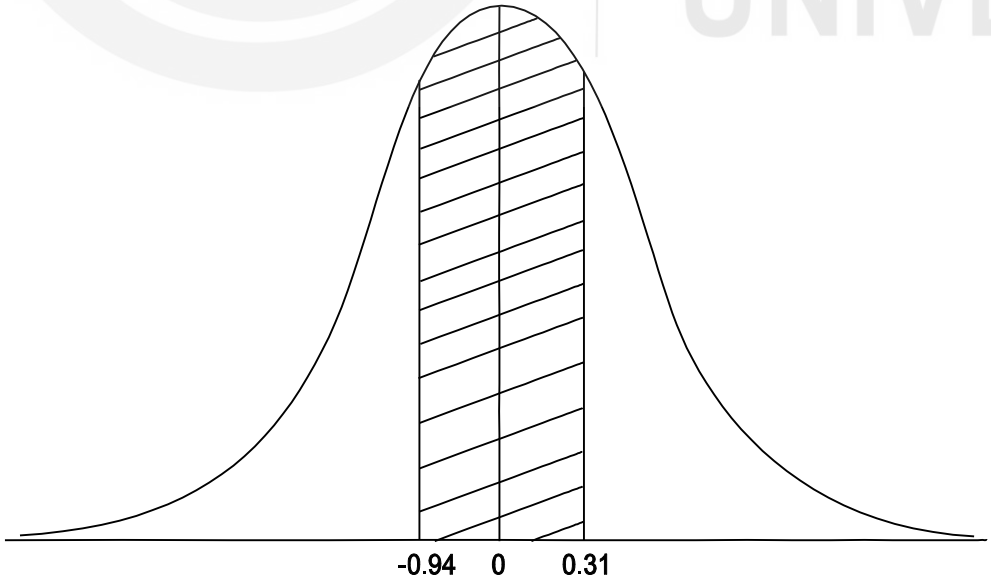


Fig.3.7

For X = 65, Z = (65-60)/16 = $0.3125 \approx 0.31$

P (45<X<65) = P(-0.94<Z<0.31)

= P(-0.94 < Z < 0) + P(0 < Z < 0.31)

= P (0<Z<0.94)+P(0<Z<0.31)

= 0.3264+0.1217 = 0.4481

Thus, the number of male students having weight between 45 kg & 65 kg

= P(45 < X <65) =1000×0.448 = 0.4481

3.3 STATISTICAL HYPOTHESIS

In our day to day life we come across many statements such as:

i) Taller parents have taller children

ii) Product X is better than product Y

iii) Average life of electric bulbs of a particular company is more compared to another

iv) Smoking and cancer are associated.

The above statements are assertions amenable to testing through statistical tools. Such statements are known as statistical hypothesis. In other words statistical hypothesis is a claim or assertion about certain characteristic or property of a population. The process of rejecting or accepting these statements is known as testing of hypothesis.

Null Hypothesis: A statistical hypothesis which nullify the difference between two or more than two statistic/treatments/attributes. It is denoted by H_0

e.g., in statement (iii) given above null hypothesis will be

$$H_o : \mu_x = \mu_y,$$

where μ_x, μ_y denote the average life of electric bulb of two brands

Alternative Hypothesis: A hypothesis complementary to the null hypothesis is called an alternative hypothesis and is denoted by H_1 or H_A

For example, in statement (iii) given above the alternative hypothesis is

$$H_1 : \mu_x \neq \mu_y,$$

The alternative hypothesis specified above can take two forms:

$$H_1 : \mu_x > \mu_y \text{ or } H_1 : \mu_y$$

For example, our alternative hypothesis will be $H_o : \mu_x > \mu_y$, if we want to test the hypothesis that the second brand hashigher average life than the second brand.

One-tailed and two-tailed Tests

If our alternative hypothesis is similar to equation (3.4) then we use two-tailed test while we use one-tailed test in case the alternative hypothesis is similar to equation (3.5). Thus the alternative hypothesis

Type I and Type II Errors

In testing of statistical hypothesis we make decision about a characteristic of the population based on the information provided by a sample (part of the corresponding population). So there are chances of wrong decisions. It means our hypothesis is true but our test procedure rejects the null hypothesis (type I error). Similarly our hypothesis does not hold for the population, but our test procedure does not reject the null hypothesis (type II error). These wrong decisions in statistics are named as type I and type II errors. Following table explains the difference between the two types of errors.

Table 3.1: Type I and Type II Errors

	H_0 is true :	H_1 is true :
Reject H_0	Wrong decision Type I error	Right decision
Accept H_0	Right decision	Wrong decision Type II error

If we reject H_0 when it is actually true we commit type I error.The probability of type I error is generally denoted by α (*alpha*).

If we accept H_0 when it is actually false we commit type II error.The probability of type II error is generally denoted by β (*beta*).

Level of Significance or Size of the 'Critical Region'

The probability of type I error (which is denoted by α) is known as the level of significance or the size of the critical region or critical area.

Tests of Hypotheses

Tests of Hypotheses are scientific techniques which put us in a position to reject or accept a given hypothesis on the basis of the information provided by the random sample drawn from the population under study.

3.4 CHI-SQUARE (χ^2)VARIATE

Degrees of Freedom

Number of independent observations which make up the statistic is known as degrees of freedom.

Degrees of freedom = total number of independent observations *minus* the number of constraints or conditions improved on the observations.

For example, suppose we want to select 4 numbers having sum 100, then we can select only 3 numbers freely according to our choice.The fourth number is automatically selected.

If we select 3 numbers as 10, 15, 40 then definitely 4^{th} number will be 35

$\therefore$ Degree of freedom = No. of observations - No. of constrains = (4 - 1) = 3

Chi-square (χ^2)Variate

It is the square of the standard normal variate (S.N.V). Remember that, if X ~ N $\left(\mu,\sigma^2\right)$, then

$$Z = \frac{x - \mu}{\sigma} \text{ is a SNV}$$

and $\chi^2 = \left(\frac{x-\mu}{\sigma}\right)^2$. This is a chi-square ($\chi^2$) variate (pronounced 'ki-square') with one 'degree of freedom'. If we have n number of SNVs and we sum them up then it follows chi-square distribution with 'n' degrees of freedom.

For example, suppose

$$x_1 \sim N\left(\mu_1, {\sigma_2}^2\right)$$

$$X_2 \sim N\left(\mu_2, {\sigma_2}^2\right)$$

…

…

$$X_n \sim N\left(\mu_n, {\sigma_n}^2\right)$$

Then

$$\chi^2 = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 + \ldots\ldots\ldots + \left(\frac{x_n - u_n}{\sigma_n}\right)^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$

The above is a χ^2 variable with '*n*' degrees of freedom.

Applications of χ^2 Distribution

The important applications of χ^2 variate are

a) It tests if the hypothetical value of the population variation is $\sigma^2 = \sigma_o^2$ (say)

b) It tests the goodness of fit

c) It tests the independence of attributes

a) Chi- Square Test for Single Variance

Assumptions:

i) The parent population is normal

ii) Sample drawn is random

Hypothesis Formulation and Test Statistic

Suppose we have a normal population with specified variance σ_o^2 (say) and we have drawn a random sample $x_1, x_2, \ldots, x_n$ (n <30) from this population. The x^2 test helps us in answering questions such as whether this random sample has actually come from the normal population having specified variance σ_o^2 or not.

For testing this sort of a hypothesis we take $H_0 : \sigma^2 = \sigma_0^2$

Alternative Hypothesis may be $H_1 : \sigma^2 \neq \sigma_0^2$

or, $H_1 : \sigma^2 > \sigma_0^2$

or, $H_1 : \sigma^2 < \sigma_0^2$

depending upon the statement of the question.

Test statistic is $x^2 = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$

$$= \frac{n}{\sigma_0^2} \left[\frac{n}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right]$$

$$= \frac{ns^2}{\sigma_0^2}, \quad \text{where } s^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

i.e., $x^2 = \frac{ns^2}{\sigma_0^2}$, where $s^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$ … (3.6)

It follows x^2distribution with (n-1) degrees of freedom

i.e., $\frac{ns^2}{\sigma_0^2} \sim x^2_{n-1}$

Note: We see that if $x^2 = \frac{(n-1)s^2}{\sigma_0^2}$, where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$

Then, still $\frac{(n-1)s^2}{\sigma_0^2} \sim x^2_{n-1}$

But we will use $x^2 = \frac{ns^2}{\sigma_0^2}$ for numerical problems.

Example 3.3: A manufacturer of a particular product claims that the variance of his products is 2 cm^2. A random sample of size 25 is taken and found that itsvariance is 2.3 cm^2 Comment on the claim of the manufacturer at 5% level of significance.

Solution: Here we are given n = 25, s^2= 2.3 cm^2 , $\sigma_0^2 = 2 cm^2$

In this case our null hypothesis is H_0: $\sigma^2 = 2 \left(= \sigma_0^2 \right)$

& alternative Hypothesis is H_1: $\sigma^2 \neq 2$

Test statistic is $x^2 = \frac{ns^2}{\sigma_0^2}$

$$= \frac{24 \times 2.3}{2} = 27.6$$

Calculated $\chi^2 = 27.6$

Tabulated $\chi^2_{0.05,24} = 36.41$ (We obtain this value from the chi-square table)

i.e., value of χ^2 at 5% level of significance for 24 degrees of freedom (df) is 36.41

Conclusion: Since calculated x^2 is less thanTabulated x^2 at 5% level of significance for 24 degrees of freedom (df), we do not reject our null hypothesis and conclude that manufacturer's claim is justified.

Chi-Square Test of Goodness of Fit

Assumptions

i) $\sum_{i=1}^{n} o_i = \sum_{u=1}^{n} = E_i$, where O_i= Observed frequencies

E_i= Expected frequencies

ii) $E_i \geq 5, 1 \leq i \leq n$

If some E_i<5, then for the application of x^2 test it has to be pooled with the preceding or succeeding cell so that expected frequency of the pooled cell is ≥ 5

Hypothesis Setup and Test Statistic

This test tells us whether the difference between the observed and expected frequencies is significant or not. Typically our null Hypothesis is H_0: the experimental results support the theory

The Alternative Hypothesis is H_1: the experimental results do not support the theory

Test statistic is $x^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i}$, O_i = Observed frequencies

E_i = Expected frequencies

It follows x^2 distribution with (n-1) degrees of freedom.

Example 3.4: In a telephone directory of a particular region it is observed that various digits (0 to 9) are used as per frequencies given in the table below.

Digits	0	1	2	3	4	5	6	7	8	9
Frequencies	900	1150	1200	1050	1040	950	860	910	980	960

Comment whether the digits 0 to 9 are equally frequently in the directory or not

Solution: Here our null Hypothesis is H_0: the digits 0 to 9 are equally frequently occurring in the directory andAlternative Hypothesis is H_1: the digits 0 to 9 are not equally frequently used in thedirectory.

Test statistic is $x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$, O_i= Observed frequency

E_i= Expected frequency

n = 10

UnderH_0, $E_i = \frac{10,000}{10} = 1000, 1 \le i \le 10$

Table 3.2: Computation of Chi-Square Value

Digits	**Observed frequencies O_i**	**Expected frequencies E_i**	**$(O_i$-$E_i)^2$**	$\frac{(O_i - E_i)^2}{E_i}$
0	900	1,000	10,000	10
1	1150	1,000	22,000	22.5
2	1200	1,000	40,000	40
3	1050	1,000	2,500	2.5
4	1040	1,000	1,600	1.6
5	950	1,000	2,500	2.5
6	860	1,000	19,600	19.6
7	910	1,000	8,100	8.1
8	980	1,000	400	0.4
9	960	1,000	1,600	1.6
Total	10,000	10,000		x^2=108.8

Calculated x^2 = 108.8

Tabulated value x^2at 5% level of significance & 10-1 = 9 degrees of freedom (df) is 16.919

Conclusion: Since calculated x^2>tabulated x^2 at 5% level of significance and 9 d.f .

So, we conclude that H_0 is rejected, i.e., the digits 0 to 9 are not equally frequently occurring in the telephone directory.

Chi-Square Test for Independence of Attributes

An important use of the chi-square test is to test for the independence of attributes.

Assumptions:

i) $\sum_{i=1}^{n} O_i = \sum_{u=1}^{n} E_i$ where O_i= Observed frequencies

E_i = Expected frequencies

ii) $E_i \ge 5, 1 \le i \le n$

If some E_i<5, then for the application of x^2 test it has to be pooled with the preceding or succeeding cell so that expected frequency of the pooled cell is ≥ 5

Hypothesis Set up and Test Statistic: Let A and B be two attributes and let them further bedivided in m & n classes $A_1, A_2, \ldots, An; B_1, B_2, \ldots, B_n$ respectively as shown in the table below.

B / A	$B_1\ B_2 \cdots B_j \cdots B_n$	Total
A_1	$O_{11}\ O_{12} \cdots O_{1j} \cdots O_{1n}$	(A_1)
A_2	$O_{21}\ O_{22} \cdots O_{2j} \cdots O_{2n}$	(A_2)
.	.	.
.	.	.
.	.	.
A_i	$O_{i1}\ O_{i2} \cdots O_{ij} \cdots O_{in}$	(A_i)
.	.	.
.	.	.
.	.	.
A_m	$O_{m1}\ O_{m2} \cdots O_{mj} \cdots O_{mn}$	(A_m)
Total	$(B_1)(B_2) \cdots (B_j) \cdots (B_n)$	N

Here our null hypothesis is H_0: Attributes A& B are independent

Alternative hypothesis is H_1: Attributes A&B are associated

Corresponding to eachO_{ij}, we have to calculate the expected frequency E_{ij} (say) under H_0

i.e.,$E_{ij}=\frac{(A_i)(B_i)}{N}$, where (A_i) = sum of i^{th} row

$$= O_{i1} + O_{i2} + O_{ij} \cdots + O_{in}$$

(B_j) = Sum of j^{th} column

$$= O_{1j} + O_{2j} + \text{---} + O_{mj}$$

$$\text{N = Grand Total} = \sum_{i=1}^{m}(A_i) = \sum_{i=1}^{n}(B_i)$$

Test statistics is $x^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

It follows x^2 distribution with (m-1) (n-1) degrees of freedom.

1) In x^2 test of goodness of fit the expected frequencies are calculated using a theoretical relationship.

2) In x^2 test of independence of attributes the expected frequencies are calculated using only observed frequencies.

3) x^2 tests for goodness of fit & independence of attributes both depend only on the observed frequencies, expected frequencies and degrees of freedom, i.e., these two tests do not involve any parameter of the parent population from which the samples are drawn. That's why these two tests are known as non- parametric tests.

4) Test Statistic in both the tests is

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$, where O_i = Observed frequencies

E_i = Expected frequencies

Note that the x^2 statistic is independent of the size of the sample

That's why these tests can be used for any sample size.

Degree of Freedom: Total no of observation in m×n contingency table = mn

Sum of each row and each column is given = (m+n) constraints

But sum of rows = sum of columns

$\therefore$ independent constraints remains = (m+n-1)

$\therefore$ Degree of freedom = (Total no of observation - Total no. of independentConstraints)

= mn-(m+n -1)

= (mn - m - n+1)

= m (n-1) -1(n-1)

= (m-1) (n-1)

Example 3.5: Calculate value of x^2 for 2×2 contingency table given below for independence of attributes.

a	b
c	d

Solution:

	B_1	B_2	Total
A_1	a	b	a + b
A_2	c	d	c + d
Total	a + c	b + d	N= a +b + c + d

Under the hypothesis of independence of attributes

$$E(a) = \frac{(a+b)(a+c)}{N}, E(b) = \frac{(b+d)(a+b)}{N}, E(c) = \frac{(a+c)(c+d)}{N},$$

$$E(d) = \frac{(b+d)(c+d)}{N} \qquad \ldots(1)$$

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \qquad O_i = \text{Observed frequencies}$$

E_i = Expected frequencies

n = 4

$$\frac{[a-E(a)]^2}{E(a)} + \frac{[b-E(b)]^2}{E(b)} + \frac{[c-E(c)]^2}{E(c)} + \frac{[d-E(d)]^2}{E(d)} \qquad \ldots(2)$$

Now a-E (a)$= a - \frac{(a+c)(a+b)}{N} = \frac{a(a+b+c+d)-(a^2+ab+ac+bc)}{N} = \frac{ad-bc}{N}$

b-E (b) $= b - \frac{(b+d)(a+b)}{N} = \frac{b(a+b+c+d)-(ab+b^2+ad+bd)}{N} = \frac{bc-ad}{N} = -\frac{ad-bc}{N}$

c-E (c) $= c - \frac{(a+c)(c+d)}{N} = \frac{c(a+b+c+d)-(ac+ad+c^2+cd)}{N} = \frac{bc-ad}{N} = -\frac{ad-bc}{N}$

d-E (d)$= d - \frac{(b+d)(c+d)}{N} = \frac{d(a+b+c+d)-(bc+bd+cd+d^2)}{N} = \frac{ad-bc}{N}$

Using these values in (2), we have

$$x^2 = \frac{(ad-bc)^2}{N}\left[\frac{1}{E(a)} + \frac{1}{E(b)} + \frac{1}{E(c)} + \frac{1}{E(d)}\right]$$

$$= \frac{(ad-bc)^2}{N}\left[\frac{1}{(a+c)(a+b)} + \frac{1}{(b+d)(a+b)} + \frac{1}{(a+c)(c+d)} + \frac{1}{(b+d)(c+d)}\right]$$

$$= \frac{(ad-bc)^2}{N}\left[\frac{b+d+a+c}{(a+c)(a+b)(b+d)} + \frac{b+d+a+c}{(a+c)(c+d)(b+d)}\right]$$

$$= \frac{(ad-bc)^2}{N}\left[\frac{N}{(a+c)(a+b)(b+d)} + \frac{N}{(a+c)(c+d)(b+d)}\right]$$

$$= (ad-bc)^2\left[\frac{c+d+a+b}{(a+c)(a+b)(b+d)(c+d)}\right]$$

$$x^2 = \frac{N(ad-bc)^2}{(a+c)(a+b)(b+d)(c+d)}$$

Example 3.6: On the basis of the information given below regarding the stature of the fathers and their sons at the age of 26 years.

Stature of Fathers

Stature of Sons	Tall	Short	Total
Tall	40	12	52
Short	32	28	60
Total	72	40	112

Can we assume that the stature of sons and the fathers are associated?

Solution: Null Hypothesis H_0: The stature of son and father is not associated.

Alternative Hypothesis H_1: Stature of son and the father is associated

$x^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}$, O_i = Observed frequencies

E_i = Expected frequencies

E (40) = $\frac{72\times 52}{112}$ = 33.43, E (12) = $\frac{52\times 40}{112}$ = 18.57

E (32) = 72-E(40) = 38.57, E(28) = 40-E (12) = 21.43

O_i	E_i	$(O_i-E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
40	33.43	43.1649	1.2912
32	38.57	43.1649	1.1191
12	18.57	43.1649	2.3244
28	21.43	43.1649	3.0142
Total = 112	112		x^2 = 6.7489

Calculated $x^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E}$ =6.7489

Tabulated x^2 at 5% level of significance for (2-1) (2-1)=1 d.f is 3.841

Conclusion: Since Calculated x^2 > Tabulated value of x^2 we reject our null hypothesis and conclude that stature of son & father is associated.

3.5 STUDENT'S T-TEST

T-test is a small sample test and known as student's t-test. There is a very interesting reason to call it as a student's t-test. It is named after its founder William Sealy Gosset (1876-1937). He worked in Guinness Brewery in Dublin, Ireland as a chemist. He published the test in Biometrika in 1908 with the pen name 'student' as his employer did not allow employees to publish scientific papers.

The t-test is basically of two types

a) One sample test ORt- test for single mean

b) Two sample test ORt-test for difference of two means

Two sample tests is further divided into two parts

i) **Independent samples t-test:** It is used to compare the means from independent groups

i) **Paired samples t- test:** It is used to compare the means that are repeated measures for the same participants-scores across time

3.5.1 Student's t test for single mean

There are certain assumptions we make while applying t-test. If these assumptions do not hold, then the validity of inference drawn on the basis of t-test is questionable. The assumptions are given below.

a) The parent population is normal, i.e.,the population from which sample is drawn should be normal.

b) Sample should be random, i.e., all the observation in the sample should be independent.

c) Standard deviation σ of the parent population should be unknown.

d) If sample size is n then $n < 30$

e) Population mean is fixed known value μ_0 (say).

Null Hypothesis and Test Statistic

Suppose a random sample $x_1, x_2, \cdots, x_n$ (n<30) is drawn from a normal population with mean μ_0 and known S.D. If we want to knowwhether this sample has actually come from the population with mean μ_0. Then student's t- test for single mean is used and we set up the null hypothesis

$$H_0 : \mu = \mu_0$$

Against the alternative Hypothesis $H_1 : \mu \neq \mu_0$ (Two tailed)

Or, $H_1 : \mu > \mu_0$ (right tailed)

Or, $H_1 : \mu : < \mu_0$ (left tailed)

(Which alternative hypothesis is used depends upon the statement of research problem)

Here our test statistic will be

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}},$$

where $\bar{x}$ = mean of sample = $\frac{1}{n}\sum_{i=1}^{n} x_i$

s = S.D. of sample = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

n = sample size

It follows student's t distribution with (n-1) degrees of freedom.

Conclusion: If calculated t < tabulated t at some given level of significance α (say) generally 5% or 1% then we may accept or fail to reject H_0 and accept H_1 otherwise we reject H_0

You should note that the tabulated t for one tailed test, given α level of significance and n-1 degrees of freedom (is obtained from the two tailed table at 2α level of significance and (n-1) degrees of freedom. Secondly, If $\bar{x}$ comes out in fraction then to avoid heavy calculation, we can use the following formula for s which is given by $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}$

Example 3.7: A domestic gas filling station claims that mean weight of its cylinder is 30 kg. A sample of 21 cylinders taken and the mean weight was noted as 28.4 kg with standard deviation of 0.50 kg. Comment on the claim of the filling station.

Solution: In the notation, we are given

Sample size = n = 21, sample mean = $\bar{x}$ = 28.4 kg

Sample standard deviation s = 0.05 kg

Population mean μ_0 = 30 kg

Here we set up null hypothes is $H_0: \mu = \mu_0 (=30)$

against the alternative hypothesis $H_0: \mu \neq \mu_0 (=30)$

Test statistic is

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}} = \frac{28.4 - 30}{\frac{0.05}{\sqrt{20}}} = \frac{-1.6}{0.118} = -14.31$$

$$\therefore |t| = 14.31$$

Tabulated value of t at 5% level of significance for 20 degrees of freedom (df) is 2.09.

Conclusion: Since calculated $|t|$ > tabulated t, so we reject H_0 and conclude that claim of filling station is false.

Example 2.8: A sample of 11 randomly selected boys had the IQs: 70, 118, 107, 99, 94, 87, 95, 97, 105, 101, 102. Can we assume that this sample is taken from a population having mean IQ of 100?

Solution: Null Hypothesis is $H_0: \mu = 100 (= \mu_0)$

Alternative Hypothesis is $H_1: \mu \neq 100$

Test statistic is $t = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n-1}}}$, $\bar{x}$ = sample mean

s = S.D. of sample

$n = 11$

x_i	A=95 $d_i = x_i$-A	d_i^2
70	-25	625
118	23	529
107	12	144
99	4	16
94	-1	1
87	-8	64
95	0	0
97	2	4
105	10	100
101	6	36
102	7	49
Total = 1075	30	1568

$$\bar{x} = \frac{\sum_{i=1}^{11} x_i}{n} = \frac{1075}{11} = 97.7273$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} d_i\right)^2} = \sqrt{\frac{1568}{11} - \left(\frac{30}{11}\right)^2}$$

$$= \sqrt{142.5455 - 7.4380} = \sqrt{135.1075} = 11.6236$$

Since, $t = \dfrac{\dfrac{97.7273}{11.6236}}{\sqrt{10}} = \dfrac{2.2727}{3.6757} = -0.6183$

Calculated t at 5% level of significance for 11-1=10 degrees of freedom (df) is 2.23

Conclusion: Since calculated $|t|$ < Tabulated t, so we may accept H_0 and conclude that it may be assumed that this sample is taken from a population having mean I.Q. of 100.

3.5.2 Independent Samples t-test

As the name suggests, this test is used to test for the difference in means between two independent samples.

Assumptions

a) Parent population should be normal i.e. parent population from which the samples have been drawn should be normal.

b) Two population have equal & unknown variances i.e. if σ_1^2 & σ_2^2 denote variances of two populations then $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say), where σ^2 is unknown

c) Two samples drawn one from each population should be random and independent.

d) Sample size should be small, i.e, n<30

Hypothesis set up & test statistic

Suppose we have two normal populations having equal and unknown variances i.e. σ_1^2, σ_2^2 denotes the variances of two populations then $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say) where σ^2 is unknown.

Suppose we want to know whether these two populations are same(i.e. $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$) where μ_1, μ_2 denote the mean of two populations.

For this purpose, we have to draw two random and independent samples one from each population. Let $x_1, x_2, \cdots, x_{n1}$; $y_1, y_2, \cdots, y_{n2}$ be two random and independent samples one from each population. Let $\bar{x}$, $\bar{y}$ respectively be their means and s_1, s_2 be their S.D.

Here we set up the null hypothesis

$$H_0: \mu_1 = \mu_2$$

Against the alternative hypothesis $H_0: \mu_1 \neq \mu_2$(two tailed test)

Or $H_1: \mu_1 > \mu_2$ (right tailed)

Or $H_1: \mu_1 < \mu_2$ (left tailed)

(nature of H_1 depends upon the statement of the problem)

Under H_0 our test statistic will be

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where } S^2 = \frac{\sum_{i=1}^{n_1}\left(x_i - \bar{x}\right)^2 + \sum_{j=1}^{n_2}\left(y_i - \bar{y}\right)^2}{n_1 + n_2 - 2}$$

$$\text{or } S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

It follows student-test with $(n_1 + n_2 - 2)$ degrees of freedom.

Example 2.9: Diet A is given to a group of 11 boys and diet B is given to another group of 8 boys. Increase in their weight is given below.

Diet A	4	5	7	8	6	2	2	2	7	8	10
Diet B	2	2	4	6	3	6	3	8			

Is diet A superior to diet B?

Solution:Let μ_1, μ_2 denote the mean increase in weights of populations which take diet A & diet B respectively

Null Hypothesis is $H_0: \mu_1 = \mu_2$

Against the alternative Hypothesis $H_1: \mu_1 > \mu_2$

Test statistic is $t = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, $S^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$$S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n_2} x_i\right)^2$$

$$S_2^2 = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j^2 - \left(\frac{1}{n_2}\sum_{j=1}^{n_2} j_i\right)^2$$

We construct a table as given below.

x	y	x^2	y^2
4	2	16	4
5	2	25	4
7	4	49	16
8	6	56	36
6	3	36	9
2	6	4	36
2	3	4	9
7	8	49	64
8		64	
10		100	
59	34	403	178

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = \frac{59}{10} = 5.9$$

$$\bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_i = \frac{34}{8} = 4.25$$

$$s_1 = \sqrt{\frac{403}{10} - \left(\frac{59}{10}\right)^2} = \sqrt{40.3 - 34.81} = \sqrt{5.49} = 2.3431$$

$$s_2 = \sqrt{\frac{178}{8} - \left(\frac{34}{8}\right)^2} = \sqrt{22.25 - 18.0625} = \sqrt{4.1875} = 2.0463$$

$$s = \sqrt{\frac{10 \times 5.49 + 8 \times 4.1875}{10 + 8 - 2}} = \sqrt{\frac{88.4}{16}}\sqrt{5.525} = 2.3505$$

Now $t = \dfrac{5.9 - 4.25}{2.3505\sqrt{\dfrac{1}{10} + \dfrac{1}{8}}} = \dfrac{1.65}{2.35.5 \times 0.4743} = \dfrac{1.65}{1.1149} = 1.48$

Calculated value of $t = 1.48$

Tabulated value of t at 5% level of significance for (10+8-2) = 16 degrees of freedom (df) is 1.75 (for one tailed test).

Conclusion: Since calculated $t <$ tabulated t, so we may accept H_0 and conclude that two diets are the same as far as increase in weight is concerned.

3.5.3 Paired Samples t-test

Paired t-test is applied when we are taking repeated samples. For example the same set of patients being before the drug and after the drug.

Assumptions:

i) Parent populations should be normal, i.e., parent population from which the samples have been drawn should be normal.

ii) Two population have equal & unknown variances i.e. if σ^2, & $\sigma_2{}^2$ denote variance of two populations than $\sigma^2, = \sigma_2{}^2 = \sigma^2$ (say), where σ^2 is unknown

iii) The sample sizes are equal and small i.e. < 30

iv) The samples are not independent but samples observations are paired together i.e. the pair of observations corresponding to the same sample unit

Hypothesis Set Up and Test Statistic

In this test we have two samples one before the treatment and one after the treatment (keep in mind that a sample unit remains same in both samples)

For example, (i) Measures of size of 10 cancer patient's tum or before the treatment constitute first sample and measures of size of the same 10 cancer patient's tumor after the treatment constitute second sample.

(ii) If we consider the case of blood pressure patient's then measures of blood pressure before and after the treatments constitute two paired samples.

In other words, we can say that paired t-test is often used in before-after situation across time. Let $x_1, x_2, \ldots\ldots, x_n$ & $y_1, y_2, \ldots\ldots, y_n$ be two samples before & after the treatment from the respective populations.

If μ_1, μ_2 denote the means of the populations before and after treatment

Then null Hypothesis will be

H_0: $\mu_1 = \mu_2$

Against the alternative Hypothesis H_0: $\mu_1 \neq \mu_2$ (two tailed test)

or H_1: $\mu_1 > \mu_2$ (right tailed)

or H_1: $\mu_1 < \mu_2$ (left tailed)

Test Statistic is

$$t = \frac{\bar{d}}{\frac{s}{\sqrt{n-1}}}, \text{ where } d_i = x_i - y_i$$

$$\bar{d} = \frac{1}{n} = \sum_{i=1}^{n} d_i$$

$$s_2 = \frac{1}{n}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2$$

It follows student's t test with n-1 degrees of freedom.

Example 3.10: Food A is given to 10 pigs for 1 month and their increase in weight is noted. After a gap of 2 months food B is given to the same 10 pigs for 1 month and their increase in weight is again noted. From the following information can we assume that food B is better than food A.

Increase in weight due to										
	Food A	38	44	40	42	37	41	43	42	47	48
	Food B	42	47	41	45	39	45	42	46	52	53

Solution: Null Hypothesis is H_0: $\mu_1 = \mu_2$, where H_0: μ_1, μ_2 are means of the increase in weights of populations having food A & food B respectively

Against the alternative Hypothesis H_1: $\mu_1 < \mu_2$

Test statistic is $t = \frac{\bar{d}}{\frac{s}{\sqrt{n-1}}}$, where $d_i = y_i - x_i$

$$\bar{d} = \frac{1}{n} = \sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} = d_i$$

$$n = 10$$

We construct a table as given below.

x_i	y_i	$d_i = y_i - x_i$	$d_i - \bar{d}$	$\left(d_i - \bar{d}\right)^2$
38	42	+4	1	1
44	47	+3	0	0
40	41	1	-2	4
42	45	3	0	0
37	39	2	-1	1
41	45	4	1	1
43	42	-1	-4	16
42	46	4	1	1
47	52	5	2	4
48	53	5	2	4
		30		32

$$\bar{d} = \frac{30}{10} = 3$$

$$s^2 = \frac{1}{10}(32) = 3.2 \Rightarrow s = \sqrt{3.2} = 1.7889$$

$$t = \frac{\bar{d}}{\frac{s}{\sqrt{n-1}}} = \frac{3}{\frac{1.7889}{3}} = \frac{9}{1.7889} = 5.031$$

Calculated $t = 5.031$

Tabulated t at 5% level of significance for 9 d.f is 1.83 (This is at 10% level of significance for 9 degrees of freedom (df) from two-tailed test table)

Conclusion: Since calculated t is greater than tabulated t, we reject the H_0 and conclude that food B is better than food A as far as increasing weight is concerned.

3.6 ANALYSIS OF VARIANCE (ANOVA)

Independent two samples test of significance for difference of two means have been discussed under t-test. But if we have more than two samples and we want to test the significance of difference of their means then ANOVA fulfils our purpose.

ANOVA is discussed by expressing total variation as a sum of its non-negative components associated with the nature of classification of data.

E.g. Suppose the mean yields of wheat per acre due to five different fertilizers (treatments) are 20, 18, 21, 24, 22 quintals respectively. If we want to test whether there is any significant difference in these means then we have to apply ANOVA.

3.6.1 ANOVA of One-way Classified Data

Let there be k treatments $T_1, T_2, \cdots\cdots, T_k$ replicated $n_1, n_2, \cdots\cdots, n_k$ times respectively. Then total number of observations will be

$n_1, n_2 + \cdots + , n_k = \sum_{i=1}^{k} n_i = n$ (say).

Let y_{ij} denotes the j^{th} observation in the i^{th} treatment

where $i = 1, 2, 3, \ldots, k$

$j = 1, 2, 3, \ldots, n_i$

Treatments

T_1	T_2	$T_3 \cdots$	$T_i \cdots$	T_k
y_{11}	y_{21}	$y_{31} \cdots$	$y_{i1} \cdots$	y_{k1}
y_{12}	y_{22}	$y_{32} \cdots$	$y_{i2} \cdots$	y_{k2}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
y_{1j}	y_{2j}	y_{3j}	y_{ij}	y_{kj}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
y_{1n_1}	y_{2n2}	$y_{3n_3} \cdots$	$y_{ini} \cdots$	y_{kn_k}
Mean $\bar{y}_{10}$	$\bar{y}_{20}$	$\bar{y}_{30} \cdots$	$\bar{y}_{i0} \cdots$	$\bar{y}_{k0}$
Total T_{10}	T_{20}	$T_{30} T_{i0} \cdots$		T_{k0}

The total variation in the observation y_{ij} can be split into two components as follows.

i) **The variation between treatments:** It is due to different treatments involved and can be noted and controlled

ii) **The variation within Treatments:** It is due to chance factor and beyond the human control.

Our mathematical model is

$$y_{ij} = \mu_i + e_{ij} \quad \text{..... (3.8)}$$

Where μ_i = the fixed effect due to i^{th} treatment or the mean of the i^{th}treatment of the population and e_{ij} are the errors which are supposed to be normally distributed with mean 0 and variance σ_e^2 i.e. $e_{ij}^2 \sim N(0, \sigma_e^2)$

Let $\mu = \frac{1}{n} \sum_{i=1}^{k} n_i \mu_i$, $\beta_i = \mu_i - \mu$

Where μ =general mean effect

β_i = Additional effect due to i^{th} treatment over the general mean effect

$\therefore$ equation (3.8) becomes

$$y_{ij} = \mu + \beta_i + e_{ij}$$

By the method of least square, the estimates of μ, β_i iaregiven by

$$\mu = y_{00}, \beta_i = y_{i0} - y_{00}$$
$$\therefore y_{ij} = y_{00} + (y_{i0} - y_{00}) + (y_{ij} - y_{i0})$$

Squaring both sides and then summing over all the values of *i* and *j*, we have

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^{\,2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{00}^{\,2} + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i0} - y_{00})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - y_{i0})^2$$

+ Product terms which vanish

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^{\,2} - n y_{00}^{\,2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i0} - y_{00})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - y_{i0})^2$$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}^{\,2} - y_{00}^{\,2}\right) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i0} - y_{00})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - y_{i0})^2$$

i.e., $$\begin{pmatrix}\text{Total sum of}\\ \text{squares}\\ \text{(TSS)}\end{pmatrix} = \begin{pmatrix}\text{Sum of squares due to}\\ \text{treatment effect}\\ \text{(SSA)}\end{pmatrix} + \begin{pmatrix}\text{Sum of squares due to}\\ \text{errors}\\ \text{(SSE)}\end{pmatrix}$$

Now, the TSS is computed from the quantities like $(y_{ij} - y_{00})$ and thus possesses (n – 1) independent values since $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - y_{00}) = 0$.

$\therefore$ Degree of freedom of TSS is (n-1)

SSA is calculated from k quantities like $(y_{i0} - y_{00})$ and hence possesses (k –1) degrees of freedom (df)

Similarly, SEE possesses n-k degrees of freedom (df) since it is obtained from quantities like $(y_{ij} - y_{i0})$ and $\sum_{j=1}^{n_i}(y_{ij} - y_{i0}) = 0$ for i = 1, 2, 3,…, *k*

Now TSS = SSA + SSE

And degrees of freedom (df) (n –1) = degrees of freedom (df)(*k* – 1) + degrees of freedom (df) of (n – *k*)

Now on dividing sum of squares by corresponding degrees of freedom, we get respective mean squares

$$\text{i.e.,} MSA = \frac{SSA}{K-1}\,[\text{mean squares due to treatment effect}]$$

and $MSE = \frac{SSE}{n-k}[mean\ squares\ due\ to\ error]$

Here, we note that SSA and SSE will add up to TSS

Also degree of freedom (k–1), degrees of freedom (df) of (n–k) will add up to (n–1),

But MSA and MSE will not add upto MST (Mean squares of total)

We are to test the null Hypothesis

$$H_0: \mu_1 = \mu_2 = \cdots \mu_K = \mu$$

$$\text{Or, } H_0: \beta_1 = \beta_2 = \cdots \beta_K = 0\ [Since, \beta_1 = \mu_i - \mu]$$

Now, for appropriate test, we find expected values of mean squares given by

$$E\ (MSA) = E\left(\frac{SSA}{k-1}\right) = \frac{1}{k-1} E(SSA) = \frac{1}{K-1}\left[\sum_{i=1}^{k} n_i {\beta_i}^2 + (k-1){\sigma_e}^2\right]$$

$$= \frac{1}{k-1}\sum_{i=1}^{k} n_i {\beta_i}^2 + {\sigma_e}^2$$

$$E\ (MSE) = E\left(\frac{SSE}{n-k}\right) = \frac{1}{n-k} E(SSE) = \frac{(n-k){\sigma^2}_e}{n-k} = {\sigma_e}^2$$

Under H_0 we have

$$E\ (MSA) = E\ (MSE) = {\sigma_e}^2 \quad [Since, under H_0, \beta_1 = \beta_2 = \cdots = \beta_k = 0]$$

$\therefore$ we can apply F-test

Which is $F = \frac{MSA}{MSE}$

If calculated |F|> Tabulated |F| then we reject the null Hypothesis and conclude that there is a significant difference between treatments effect and if calculated |F|< Tabulated |F| then null hypothesis may be accepted and conclude that there is no significant difference between treatments effect.

ANOVA TABLE

Source of variation	S.S	S.S	M.S	F
Between Treatments	SSA $= \sum_{i=}^{k} n_i (y_{i0} - y_{00})^2$	k–1	MSA $= \frac{SSA}{k-1}$	
With Treatments (Due to error)	SSE $= \sum_i \sum_j (y_{ij} - y_{00})^2$	n–k	MSE $= \frac{SSE}{n-k}$	F $= \frac{MSA}{MSE}$
Total	TTS $= \sum_i \sum_j (y_{ij} - y_{00})^2$	n–1		

Formulae Used for Numerical Problems

1) Grand Total = G= $\sum_i \sum_j y_{ij}$

2) Correction factor $(C.F) = \frac{G^2}{n}$, n=$n_1 + n_2 + \cdots + n_k$

3) Raw sum of squares (RSS) $= \sum_i \sum_j {y_{ij}}^2$

4) Total sum of squares (TSS) = Raw sum of squares – C.F

5) Sum of squares due to treatments (SSA)

$$= \frac{{T_{10}}^2}{n_1} + \frac{{T_{20}}^2}{n_2} + \cdots + \frac{{T_{k0}}^2}{n_k} - C.F$$

6) Sum of squares due to errors (SSE) = TSS – SSA

7) MSA $= \frac{SSA}{k-1}$, $MSE = \frac{SSE}{n-k}$

8) Finally, $F = \frac{MSA}{MSE}$

Example 3.11: An investigator wants to know the level of knowledge of students about the history of India of 4 different schools in a city A test is given to 5, 6, 7, 6 students of 8^{th} class of 4 schools. Their score out of 10 is given below.

School I:	8	6	7	5	9		
School II:	6	4	6	5	6	7	
School III:	6	5	5	6	7	8	5
School IV:	5	6	6	7	6	7	

Solution: If $\mu_1, \mu_2, \mu_3, \mu_4$ denote the average score of students of 8^{th} class of schools I, II, III, IV respectively. Then

Null Hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

Alternative hypothesis H_1: difference among $\mu_1, \mu_2, \mu_3, \mu_4$ are significant.

s_1	s_2	s_3	s_4	${s_1}^2$	${s_2}^2$	${s_3}^2$	${s_4}^2$
8	6	6	5	64	36	36	25
6	4	5	6	36	16	25	36
7	6	5	6	49	36	25	36
5	5	6	7	25	25	36	49
9	6	7	6	81	36	49	36
	7	8	7		49	64	49
		5				25	
35	34	42	37	255	198	260	231

Grand Total G = 35+34+42+37=148

Correction factor $(C.F.) = \frac{G^2}{n} = \frac{(148)^2}{24} = 912.6667 \left[since\ n = n_1 + n_2 + n_3 + n_4 = 5 + 6 + 7 + 6 = 24\right]$

Raw sum of squared RSS $= \sum_i \sum_j {y_{ij}}^2 = 255 + 198 + 260 + 231 = 944$

Total sum of square (TSS) = RSS – C.F = 944 – 912.6667 = 31.3333

Sum of squares due to treatments (SSA) $= \frac{{T_{10}}^2}{n_1} + \frac{{T_{20}}^2}{n_2} + \frac{{T_{30}}^2}{n_3} + \frac{{T_{40}}^2}{n_4} - C.F$

$= \frac{35^2}{5} + \frac{34^2}{6} + \frac{42^2}{7} + \frac{37^2}{6} - 912.6667$

$= 245+192.6667+252+228.1667-912.6667$

$= 5.1667$

Sum of squares due to errors = SSE = TSS – SSA = 31.3333 – 5.1667 = 26.1666

$$MSA = \frac{SSA}{k-1} = \frac{5.1667}{3} = 1.7222$$

$$MSE = \frac{SSE}{n-k} = \frac{26.1667}{20} = 1.3083$$

ANOVA TABLE

Source of variation	S.S	d.f	M.S	F
Between schools	5.1667	3	1.7222	$F=\frac{1.7222}{2.30833} = 1.3164$
Within schools	26.1666	20	1.3083	

Calculated F = 0.7461

Tabulated F at 5% level of significance with (3,20) degree of freedom is 3.10

Conclusion: Since calculated F < tabulated F,we may accept H_0 and conclude that level of knowledge of school I, II, III, IV do not differ significantly.

3.6.2 ANOVA of Two-way Classified Data

In one way classified data, we have only one independent factor of variation. But there are situations where more than one factor of variation is present, e.g.in a field experiment, if whole of the experimental area is not homogeneous (as far as fertility is consider) and fertility gradient is only in one direction, then ANOVA of two way classified data is used for test of significance of difference of means. In such a case experimental area is divided into homogenous sub-groups called blocks. This stratification is done in such a way that the plots within a block are relatively similar in comparison to plots belonging to different blocks.

Let y_{ij} be the yield from the plot receiving the i^{th} treatment in the j^{th} block. Since the experimental material is relatively homogenous in each block, the yield may be assumed to depend only on particular block and treatment

$\therefore$ our mathematical model is

$$y_{ij} = \mu + \tau_i + \beta_j + e_{ij} \quad \ldots \quad (1), i=1, 2, 3, \ldots, t \quad J= 1, 2, 3, \ldots, r$$

where $\mu =$ the general effect

$\tau_i = i^{th}$ Treatment effect

$\beta_j = j^{th}$ Block effect

And e_{ij} are independent normally distributed random variables with mean 0 & variance $\sigma_e{}^2$

Here null hypothesis are

$$H_{01}: \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

$$H_{02}: \beta_1 = \beta_2 = \cdots = \beta_r = \cdots = \beta_r = 0$$

Proceeding as in one way ANOVA, we have

$$E(SST) = r \sum_i \tau_i{}^2 + (t-1)\sigma_e{}^2$$

$$E(SSB) = t \sum_i \beta_j{}^2 + (r-1)\sigma_e{}^2$$

$$E(SSE) = (t-1)(r-1)\sigma_e{}^2$$

$$E(MST) = E\left(\frac{SST}{t-1}\right) = \frac{1}{t-1} E(SST) = \frac{r}{t-1}\sum_i \tau_i{}^2 + \sigma_e{}^2$$

$$E(MSB) = E\left(\frac{SSB}{r-1}\right) = \frac{1}{r-1}\sum_i \beta_j{}^2 + \sigma_e{}^2$$

$$E(MSE) = E\frac{SSE}{(t-1)(r-1)} = \frac{1}{(t-1)(r-1)} E(sse) = \sigma_e{}^2$$

Under null hypothesis

$$H_{01}: \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

We have,

$$E(MST) = E(MSE) = \sigma_e{}^2$$

And thus, the test statistic is

$$F_1 = \frac{MST}{MSE}$$

Under null hypothesis

$$H_{02}: \beta_1 = \beta_2 = \cdots = \beta_r = 0$$

We have,

E (MSB) = E (MSE) = $\sigma_e{}^2$

And thus, the test statistic is

$$F_2 = \frac{MSB}{MSE}$$

If calculated |F|> Tabulated |F|, then we reject the null hypothesis

If calculated |F|< Tabulated |F|, then we may accept the null hypothesis

ANOVA TABLE

Source of variation	**S.S**	**d.f**	**M.S**	**F**
Between Treatments	SST	t-1	$MST = \frac{SST}{t-1}$	$F_1 = \frac{MST}{MSE}$ $F_2 = \frac{MSB}{MSE}$
Between Blocks	SSB	r-1	$MSB = \frac{SSB}{r-1}$	
Due to error	SSE	(t-1)(r-1)	$MSE = \frac{SSE}{(t-1)(r-1)}$	
Total	TSS	rt-1		

Formula used for numerical problems

1) Grand Total = $G = \sum_i \sum_j y_{ij}{}^2$

2) Correction factor $(C.F) = \frac{G^2}{n}$

3) Raw sum of squares $(RSS) = \sum_i \sum_j y_{ij}{}^2$

4) Total sum of squares (TSS) = RSS – CF

5) Sum of squares due to treatments $(SST) = \frac{T_{10}{}^2}{t} + \frac{T_{20}{}^2}{t} + \cdots + \frac{T_{to}{}^2}{t} - C.F$

6) Sum of squares due to Blocks $(SSB) = \frac{T_{01}{}^2}{b} + \frac{T_{02}{}^2}{b} + \cdots + \frac{T_{0r}{}^2}{b} - C.F$

7) Sum of squares due to errors (SSE) = TSS – SST – SSB

8) Mean sum of squares due to treatments $(\text{MST}) = \frac{SST}{t-1}$

9) Mean sum of squares due to blocks $(\text{MSB}) = \frac{SSB}{r-1}$

10) Mean sum of squares due to errors $(\text{MSE}) = \frac{SSE}{(t-1)(r-1)}$

11) Finally $F_1 = \frac{MST}{MSE}, F_2 = \frac{MSB}{MSE}$

Example 3.12: A researcher wants to test four diets A, B, C, D on growth rate in mice. These animals are divided into 3 groups according to their weights. Heaviest 4, next 4 and lightest 4 are put in Block I, Block II and Block III respectively. Within each block, one of the diets aregiven at random to the animals. After 15 days increase in weight is noted and given in the following table.

Blocks	**Treatments/Diets**			
	A	B	C	D
I	12	8	6	5
II	15	12	9	6
III	14	10	8	5

Perform an ANOVA and find whether data indicate any significant difference between the four diets due to different blocks.

Solution: Null Hypotheses are H_{01}: There is no significant difference between mean effect of diets.

H_{02}: There is no significant difference between mean effects of different blocks.

Against the alternative hypothesis

H_{11}: There is significant difference between mean effects of diets

H_{12}: There is significant difference between mean effects of different blocks.

We construct a table as given below.

Blocks	**Treatments/Diets**				
	A	**B**	**C**	**D**	**Totals**
I	12	8	6	5	31 = T_{01}
II	15	12	9	6	42 = T_{02}
III	14	10	8	5	37 = T_{03}
Totals	41	30	23	16	110
	T_{10}	T_{20}	T_{30}	T_{40}	Grand total

Squares of observations

Blocks	Treatments/Diets				Totals
	A	B	C	D	
I	144	64	36	25	269
II	225	144	81	36	486
III	196	100	64	25	385
Totals	565	308	181	86	1140

Grand Total $= G = \sum_i \sum_j y_{ij} = 110$

Correction factor $(C.F) = \frac{G^2}{n} = \frac{(110)^2}{12} = 1008.3333$

Raw sum of squares $(RSS) = \sum_i \sum_j {y_{ij}}^2 = 1140$

Total sum squares (TSS) = RSS – C.F = 1140 – 1008.3333 = 131.6667

Sum of squares due to treatments/ Diets $(SST) = \frac{{T_{10}}^2}{3} + \frac{{T_{20}}^2}{3} + \frac{{T_{30}}^2}{3} + \frac{{T_{40}}^2}{3} - C.F$

$$1008.3333 = \frac{1}{3}[(41)^2 + (30)^2 + (23)^2 + (16)^2] -$$

$$= \frac{1}{3}[1681 + 900 + 529 + 256] - 1008.3333$$

$$= 1122 - 1008.3333$$

$$= 113.6667$$

Sum of squares due to block (SSB)

$$= \frac{1}{4}\left({T_{01}}^2 + {T_{02}}^2 + {T_{03}}^3 + {T_{04}}^2\right) - C.F$$

$$= \frac{1}{4}[(31)^2 + (42)^2 + (37)^2] - 1008.3333$$

$$= 1023.5 - 1008.3333 = 15.1667$$

Sum of squares due to errors

$$(SSE) = TSS - SST - SSB$$

$$= 131.6667 - 113.6667 - 15.1667$$

$$= 2.8333$$

Mean sum of squares due to treatments

$$(\text{MST}) = \frac{SST}{t-1} = \frac{113.6667}{3} = 37.8889$$

Mean sum of squares due to blocks

$$(\text{MSB}) = \frac{SSB}{r-1} = \frac{15.1667}{2} = 7.58335$$

Mean sum of squares due to errors

$$(\text{MSE}) = \frac{SSE}{(r-1)(t-1)} = \frac{113.6667}{3\times 2} = 0.4722$$

ANOVA TABLE

Source of variation	S.S	d.f	M.S	F
Between Treatments/ Diets	113.6667	3	$\frac{113.6667}{3} = 37.8889$	$F_1 = \frac{37.8889}{0.4722} = 80.2391$
Between blocks	15.1667	2	$\frac{15.1667}{2} = 7.58335$	$F_2 = \frac{7.58335}{0.4722} = 16.0596$
Due to errors	2.8333	6	$\frac{2.8333}{6} = 0.4722$	
Total	131.6667	11		

Tabulated F at 5% level of significance with (3, 6) degree of freedom is 4.76

&Tabulated F at 5% level of significance with (2, 6) degree of freedom is 5.14

Conclusion: Since calculated F_1> Tabulated F, so we reject H_{01}, and conclude that there is significant difference between mean effect of diets.

Also calculated F_2> Tabulated F, so we reject H_{02} and conclude that there is significant difference between mean effect of different blocks.

3.7 SUMMARY

- Let us summarise the topics what we have gone through in this unit.
- Definitions, characteristic, importance and simple applications of normal & standard normal distributions.
- Statistical hypotheses, null hypothesis, alternative hypothesis, one tailed & two tailed hypotheses, one tailed & two tailed tests type I& type II errors, critical region.
- Assumptions, hypotheses set up, test statistic and simple applications of chi-square t-test.
- Assumptions, hypotheses set up, test statistic and simple applications of student's test.

- Assumptions, hypotheses set up, test statistic and simple applications of ANOVA.

Sample Questions

1) Let the random variable X denote the chest measurement (in cm.) of 2000 boys, where X ~ N (70, 36). Then find no of boys having chest measurement,

   a) Less than or equal to 68 cm

   b) Between 71 cm & 75 cm

   c) More than 65 cm

2) In a particular branch of a bank, it is noted that the duration/waiting time of the customers for being served by the teller is normally distributed with mean 5.5 minutes & standard deviation 0.6 min. Find the probability that a customer has to wait a) between 2.2 & 4.5, (b) For less than 5.2 minutes, and (c) more than 6.8 minutes.

3) Suppose that temperature of a particular city in the month of March is normally distributed with mean 24°c and standard deviation 6°c. Find the probability that temperature of the city is a) less than 20°c, (b) more than 26°c, and (c) between 23°c and 27°c.

4) Variance of a random sample of size 20 is found to be 0.25. Test whether this sample is drawn from a normal population with variance 0.12 (hint: apply chi-square test).

5) 2000 students of a university were classified according to their intelligence and economic conditions as shown below.

		Intelligence				
Economic Conditions		Excellent	Good	Medicos	Dull	Total
	Good	60	350	300	90	800
	Poor	110	300	400	390	1200
	Total	170	650	700	800	2000

6) A surveys of 1000 people gave the following information

Smoking/Literacy	**Smoker**	**Non Smoker**	**Total**
Literates	200	250	450
Illiterates	320	230	550
Total	520	480	1000

Test whether the literacy is related to the habit of cigarette smoking.

7) Scores of two similar tests of 10 candidates one before the training and one after the training of one week are given below.

Scores Before training	75	50	45	48	65	62	85	87	90	55
Scores After training	85	60	65	50	68	65	86	90	92	72

Can we assume that training is effective?

8) In 25 plots four varieties v_1, v_2, v_3, v_4 of wheat are randomly put and their yield in kg are shown below.

(v_1) 2000	(v_3) 2270	(v_2) 2230	(v_4) 2270	(v_4) 2180
(v_2) 2160	(v_1) 2100	(v_2) 2050	(v_3) 2300	(v_2) 2280
(v_1) 2200	(v_1) 2300	(v_4) 2040	(v_3) 2420	(v_1) 2240
(v_4) 2370	(v_1) 2250	(v_2) 2040	(v_2) 2360	(v_1) 2460
(v_3) 2210	(v_1) 2340	(v_2) 2190	(v_1) 2150	(v_3) 2020

Perform the ANOVA and test the effect of varieties.

UNIT 4 USING SPSS FOR DATA ANALYSIS CONTENTS

Contents

Learning Objectives

It is expected that after going through Unit 4, you will be able to

- understand the use of SPSS in your data analysis;
- start and Exit SPSS program;
- enter the data into a SPSS data editor; and
- import a data file from excel program.

4.1 INTRODUCTION

SPSS (Statistical Package for Social Sciences) computer software program provides access to a wide range of data management and statistical analysis procedures. This program can perform a variety of data analysis including tables, statistical analysis and graphical presentation of data. Also, SPSS is particularly well suited to sample survey research.

It is assumed here that you have a basic understanding of the basic concepts and techniques of statistical analysis. In Unit 4, you will learn how to use the SPSS to perform data analysis.

You can run SPSS program on a Personal Computer (PC) within the Windows (95, 98, 2000, XP, or NT) operating system. Since it is a windows based program, you can use the program without any difficulty and more interactively like Word, Excel, or PowerPoint programs. The command instructions given and examples shown in this Unit are Windows based SPSS version 11.5.

Please note that Unit 4 does not carry any information in boxes as it has plenty of graphics to understand the details without the aid of boxes. For Reflection and Action exercises, there are some straight questions for you to answer as the reflection part of the exercise is going to take place during your reading of the text along with its graphics. It is a good idea for you to repeat the viewing of these graphics as many times as possible and practice using them for carrying out first simpler and later more complex tasks.

4.2 STARTING AND EXITING SPSS

Normally, SPSS program will be located in the **Programs** folder of your PC. To start the SPSS,

1) Click the left mouse button on the **Start** button located at the lower left of the screen. A number of items will be listed on the screen.

2) Select **Programs**. The program menu will open.

3) Select **SPSS for Windows** from the programs menu and then select **SPSS 11.5 for Windows** from the SPSS menu. Click and release the mouse button. Symbolically, these actions are shown as: select **Start** → **Programs** → **SPSS for Windows** → **SPSS 11.5 for Windows** command from the Start button of your PC. (Throughout this Unit, we will be showing the symbol → to indicate the director (steps)you have to move your cursor with mouse.).

Fig. 4.1: SPSS for Windows

4) After a few moments, you will see the **Data Editor** window dialog® box along with a **SPSS for Windows** menu dialog box asking you "What would you like to do?".

Fig. 4.2: SPSS for Data Editor

Exit SPSS Data Editor: Whenever you have finished using SPSS and want to quit it, then select **File** → **Exit** command on the menu bar.

4.3 CREATING A DATA FILE

Normally, the first thing you would like to do is to create a data file. For this, check mark on the Type in data box and then click OK button on the SPSS for Windows menu dialog box. The menu dialog box disappears from the screen leaving the **Data Editor** on the screen.

Data Editor: The Data Editor helps you:

1) To enter a series of data you have in a specified format required for data analysis.
2) Open an existing file.
3) Edit the data.
4) Converting other data files into SPSS data files.
5) Will be active throughout your session of using SPSS data entry and data analysis.

The Data Editor looks like a worksheet made up of a series of rows and columns. The intersection of a row and column is called a cell. The cells may contain numbers or text. Each column will contain information/data for each variable. Similarly, each row contains information/data for each case.

The first row of the cells located at the top of each column is shaded and contains a faint **Var**. These cells contain the names of variables. Similarly, the first shaded column contains faint numbers (1,2,3,...). These are called case numbers.

The Data Editor dialog box contains a Menu bar at the top of the window. The menu bar identifies broad categories of SPSS's features called commands. This menu bar helps you in defining and selecting commands.

Fig. 4.3: Menu bar and Tool Bar of SPSS Data Editor

The Toolbar is below the menu bar and allows you to quickly access basic SPSS commands. By clicking on the respective buttons you can access some commands which will interest you quite often.

Observe the cell at the intersection of row 1 (Case 1) and column 1 (Var 1) with a heavy border. The heavy border cell indicates that the cell is an active cell. You can enter or edit data in the cell. You can activate any cell in the worksheet by simply pointing the mouse cursor® at it and clicking once.

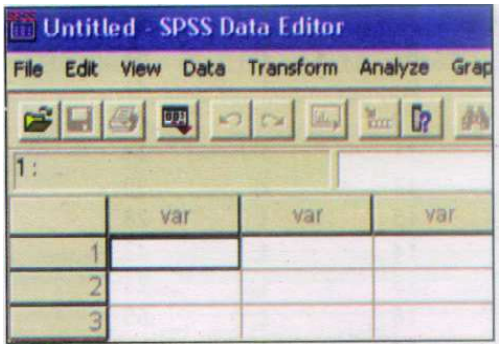


Figure 4.4 Untitled SPSS Data View

There are two views available in the Data Editor: Data View and Variable View, In the Data View, you can see the data the way you have typed. In the Variable View, you can see the properties of each variable defined. To access these views click on the respective buttons located at the left bottom of the Data Editor screen.

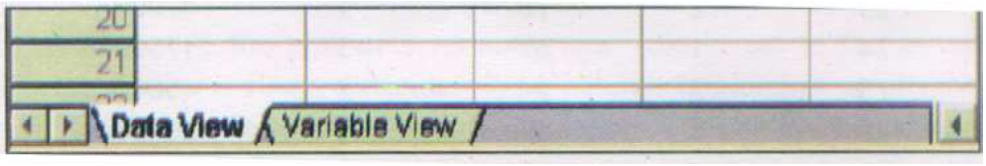


Fig. 4.5: Untitled SPSS Variable View

Entering data into Data Editor

As you have learned earlier, your data may contain a number of variables. Examples of variables are sex (male or female), marital status (married, unmarried, widowed, etc.), income, attitude (likes, dislikes, etc.), a score, etc. Again, for each variable you may have a number of observations called cases. You have also learned in an earlier unit how to assign the codes to the qualitative data. For example, you have a data set containing information of 30 people on sex distribution, age, marital status, and income. You also assigned the following codes for the data set.

Sex distribution: male = 1, female = 2

Age in years

Marital status: married = 1, unmarried = 2, widowed = 3

Income in Rs.

The data you might have collected may look like as follows;

Table 4.1 Data Set by Sex, Age, Marital Status and Income for 30 Persons

Case Number	Sex	Age	Marital Status	Income (Rs.)
1	2	24	1	150000
2	1	52	1	345000
3	2	65	3	45000
4	1	35	3	245000
5	1	42	1	23000
6	1	25	1	670000
7	2	23	2	345000 I
8	2	63	1	156000
9	1	41	2	65300
10	2	48	1	150000
11	1	34	2	354000
12	2	55	3	23000
13	1	28	1 .	452000
14	1	43	2	120000
15	1	23	2	456000

16	2	65	1	765000
17	2	67	3	235000
18	2	32	2	54000
19	1	30	2	200000
20	2	25	2	180000
21	1	47	1	210000
22	2	36	3	350000
23	2	70	1	42000
24	1	67	3	175000
25	2	24	2	45000
26	1	32	1	234000
27	1	20	2	125000
28	2	25	1	36000
30	2	40	3	560000
30	2	45	1	234000

The process of data entry into the Data Editor involves four basic steps

1) Define variables
2) Define labels
3) Define missing values
4) Enter the data into the cells

We will explain these steps with the help of the data given in our earlier example. For this move your cursor to the left bottom of the Data Editor and click on Variable View button, if the Data Editor is not in the Variable View.

Step 1: Define Variables: You need to define a variable to name it, specify the data type (qualitative, quantitative, number of decimal places, etc.), assign labels to the variable and data values, define missing values, and specify levels of measurement (nominal, ordinal, interval/ratio scale). In addition, you can also define the column format. For this,

1) Activate a cell in the first column by clicking on it.
2) Click on the Variable View button. The grid will change to a new format as shown below. For each variable you create, you need to specify all or most of the attributes described by column headings.
3) Activate the first cell in row 1 under Name column heading to change the variable name. Type the name of the variable say Sex and then press enter key. Observe that the variable name Sex replaces with the default variable name Var and in the other cells of the first row the default properties of the variable will appear.

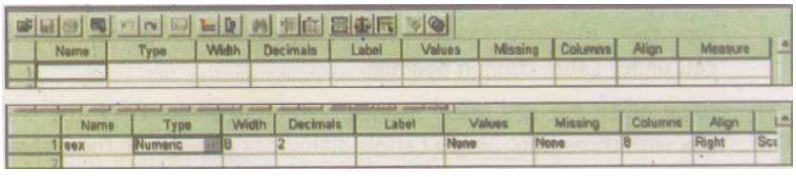


Fig. 4.6: Changes in Data Grid and Data Variable

Remember that a variable name can have a maximum of 8 characters containing only letters or both letters and numbers. The first character of a variable name should start with a letter when it contains numbers. There should not be any special characters (like &, ?, !, ',*) in the variable name. Also, in the same data file no two variables should have the common names.

4) To change the type of a variable, move the cursor to the second cell of the first row under **Type** column heading. A small grey button marked with three dots will appear. Click on it. The **Variable Type** dialog box appears on the screen. Notice that Numeric is the default Variable Type. If you have only numeric values for that variable (say Sex variable) check mark the **Numeric** box. You can enter the width of the number (the default width is 8 characters) in the **Width** text box. Sometimes you may need to enter numbers with decimal places. Enter the number of decimal places in the **Decimal Places** text box. The default setting is 2 decimal places. If your data Figure 4. 7 Variable contained only integer values, type 0 in the **Decimal Places** text box. If you have a variable string characters (like names of people, places, etc.), check mark the String button in Variable type dialog box and enter the maximum number of characters that particular string variable can hold. Similarly, other data variable types such as date, currency, etc., can also be defined.

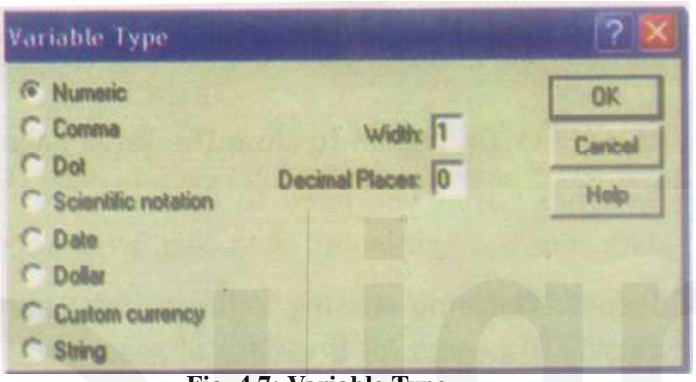


Fig. 4.7: Variable Type

5) Specify the level of measurement for the variable (for example, Sex is a nominal variable). By clicking the cell under **Measure** column heading.

Step 2: Define Labels: Now you can assign the text labels to the coded values of the variables. A variable label is a longer description of the variable that can be included in the variable name you have defined earlier: This may be necessary since the variable name is restricted to only 8 characters and at later stages to understand the characteristics of that variable. To define labels,

1) Type the name of the label (say Sex distribution of persons) in the cell under Label column heading.

2) Move on to the cell under **Values** column heading. Click the grey box with three dots. **Value Labels** dialog box appears on the screen. Type the numerical value assigned for the label under Value text box and type label name for that value under **Value** Label text box. For example, you may type 1 in the Value text box and Male in the **Value Label** text box. To add the label, click **Add** button. Again type 2 in the Value text box and Female in the Value Label text box and then click Add button.

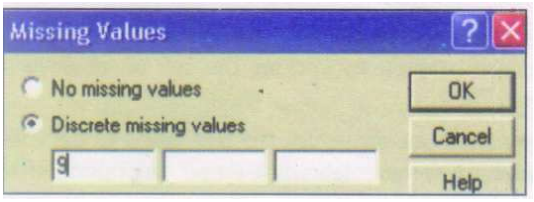


Fig. 4.8: Variable Labels

3) Continue this process until you add all the values and the respective labels.

4) Click **OK** button to close the Value Labels dialog box.

Remember that you need to define Value Labels only for categorical data. For the continuous data this is not required.

Step 3: Defining Missing Values: Sometimes, your data may contain missing responses for a variety of reasons. Assign a missing value to the variable if necessary. For example, if the Sex category of a person is not available, you may assign the value 9 to indicate the missing value. The missing value indicates to SPSS that the response is not available and should not be included in the data analysis. To define the missing values to the variable,

1) Click on the cell under **Missing** column heading. The **Missing Values** dialog box appears on the screen.

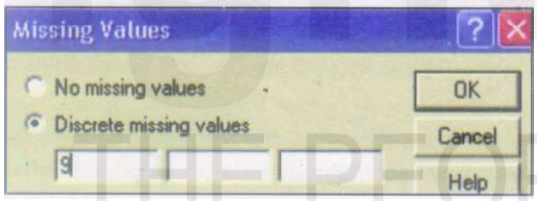


Fig. 4.9: Missing Values

2) If there are no missing values in the variable, check mark **No missing values** button. Otherwise check mark **Discrete missing** values button. You can type a number assigned by you as Missing value. For example, you may assign 9 if Sex category of a person is not available. If you have check marked for missing values, type the assigned missing value (say 9) in the text box.

3) Click **OK** button to close the dialog box.

Remember that the missing values you have assigned for a variable are present in their respective positions in a data file.

You can also specify the width of each column in the cell under **Column** column heading and the alignment (Right, Left, and Center) in the cell under **Align** column heading. Specify the type of measure in the cell under Measure heading.

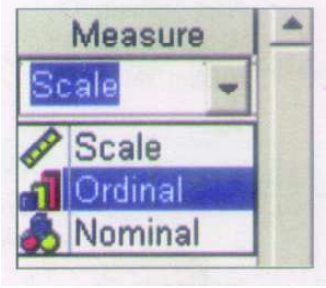


Fig. 4.10: Specifying under Measure

Now you have defined all the information for the variable. Move on to the second row (to define 'age' variable), the third row (to define 'marital status' variable), etc. to define the rest of the variables in that order.

	Name	Type	Width	Decimals	Label	Values	Missing	Column	Align	Measure
1	sex	Numeric	1	0	Sex distributio	{1, Male}...	9	8	Right	Nominal
2	age	Numeric	1	0	Age in years	None	99	8	Right	Scale
3	marital	Numeric	1	0	Marital status	{1, Married}...	9	8	Right	Nominal
4	income	Numeric	6	0	Income in Rs.	None	99	8	Right	Scale
5										

Fig. 4.11: (a) Defining all the variables

you have defined all the variables, you may like to see the generated variable definitions. For this, select **Utilities → File Info** command on the menu bar. This will generate file information in the output window that can be printed if you need for future reference.

File Information

List of variables on the working file

Name Position

SEX **Sex distribution of persons**

1) Measurement Level: Nominal

Column Width: 8 Alignment: Right Print

Format: F1

Write Format: F1

Missing Values: 9

Value	Label
1	Male
2	Female

AGE **Age in years**

2) Measurement Level: Scale

Column Width: 8 Alignment: Right

Print Format: F1

Write Format: F1

Missing Values: *

MARITAL **Marital status**

3) Measurement Level: Nominal

Column Width: 8 Alignment: Right

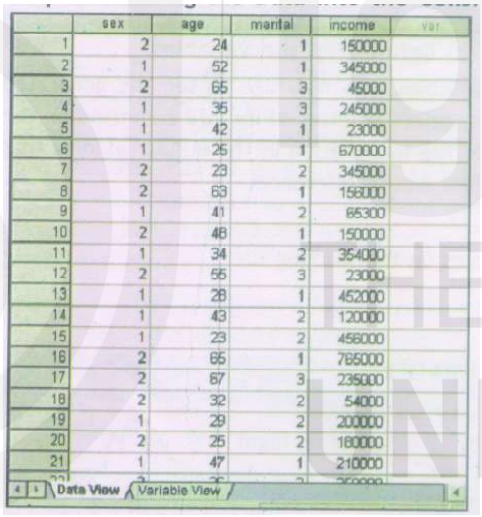
Print Format: F1

Write Format: F1

Missing Values: 9

	Value	Label
	1	Married
	2	Unmarried
	3	Widowed
INCOME	**Income in Rs.**	
4)	Measurement Level: Scale	
	Column Width: 8	
	Alignment: Right	
	Print Format: F6	
	Write Format: F6	
	Missing Values: 99	

Step 4: Entering the Data into the Cells:



	sex	age	marital	income	var
1	2	24	1	150000	
2	1	52	1	345000	
3	2	65	3	45000	
4	1	35	3	245000	
5	1	42	1	23000	
6	1	25	1	670000	
7	2	28	2	345000	
8	2	63	1	156000	
9	1	41	2	65300	
10	2	48	1	150000	
11	1	34	2	354000	
12	2	56	3	23000	
13	1	28	1	452000	
14	1	43	2	120000	
15	1	29	2	456000	
16	2	65	1	785000	
17	2	67	3	235000	
18	2	32	2	54000	
19	1	29	2	200000	
20	2	25	2	180000	
21	1	47	1	210000	

Data View | Variable View

Fig. 4.11: (b) Entering Values for Each Variable

Once all the variables for your data file are defined, the data can be directly entered into the cells. For this, first change the view to Data View by clicking the **Data View** button.

1) Click on cell 1 of the (Sex) variable to activate the cell.

2) Type the value of the variable (say '2') and then press **Enter** key.

Observe that now number '2' appears in cell 1 and cell 2 (the cell below cell-1) becomes active.

3) Type 1 and press Enter key. This indicates that the value for case of the sex variable is also entered. Continue this procedure until all the values of the 30 cases are entered for sex variable.

4) Activate the case 1 cell below Age variable (column) and start entering values for that variable. Continue this procedure until you enter the data for all cases and all **variables.ng a data file**

Once you have entered the data, it is a good practice to save the data in a file. This will avoid not only repeating the data entry but also for all future uses of your data. SPSS distinguishes between two types of files: data files (with extension .sav) and output files (with extension .spo). The data files contain the data you have entered. The output files contain the output of the data analysis you have performed. You need to save these files in case you may need them for future use.

To save a data file,

1) Select **File → Save As** command from the menu bar. The: **Save Data** As dialog box appears on the screen.

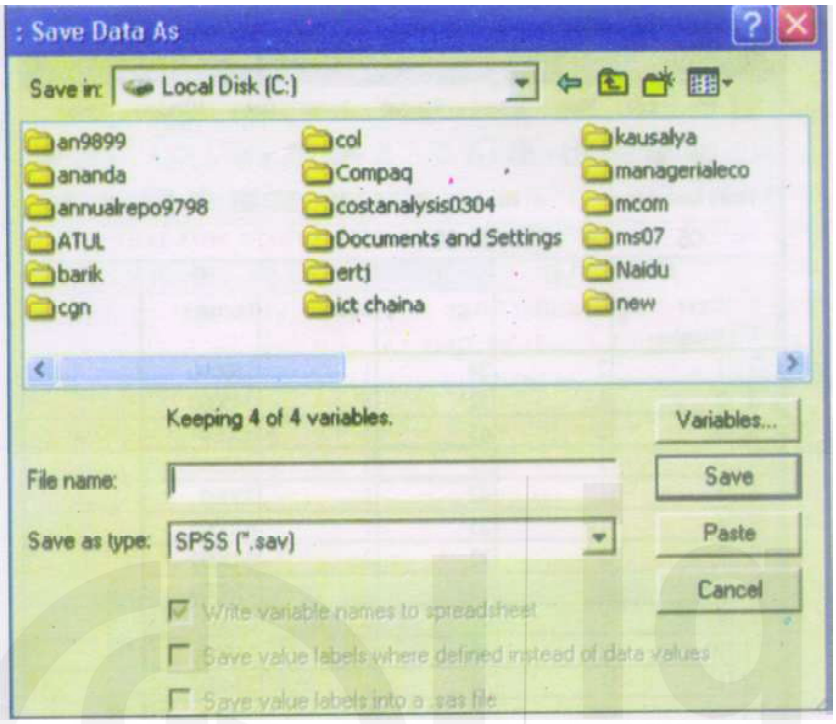


Fig. 4.12: Saving a Data File

2) Select the drive and folder where you are interested that your file should be located.

3) Type the file name in the text box under **File name** box. Click Save button.

Importing a data file from Excel worksheet

Often you might have entered data in an Excel worksheet and want to use the same data set for analysing the data using SPSS. SPSS can easily

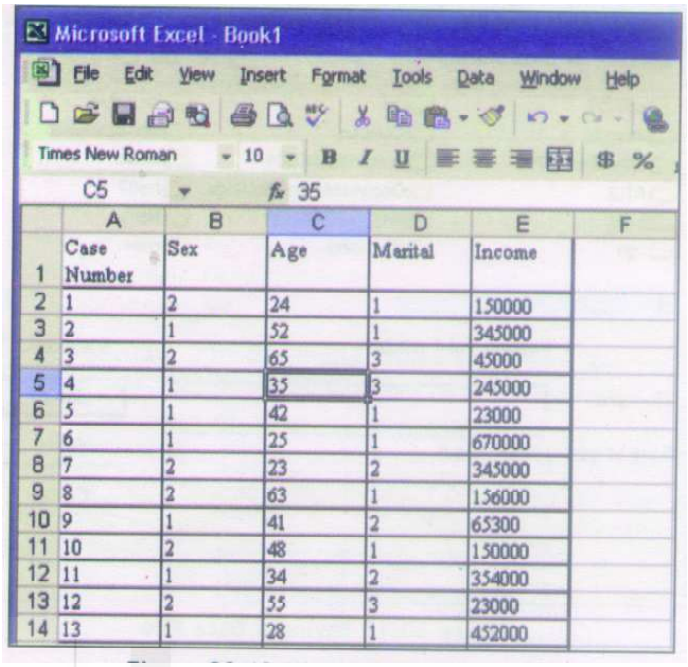
open an Excel data file and some other types of data files. To open an Excel data file,

1) Select Starts Programs Microsoft Excel command from the Start button on your PC screen. In a few moments the Excel worksheet dialog box appears on the screen.

2) Select **File → Open** command from the menu bar. The **Open File** dialog box appears on the screen. Select the drive and folder where the data file is stored. Select (or type) the file name and click OK button. The Excel data file will open as shown here.

Observe that the variable names are at the top row. Let us assume that the worksheet has been saved as Excel (with extension .xls) file called Profile.

3) Select **File → Open → Data** command in the SPSS Data Editor dialog box. The **Open File** dialog box appears. Choose the appropriate directory and folder in **Look** in File of type box. Select the file name Profile.xls in File name box.



	A	B	C	D	E	F
1	Case Number	Sex	Age	Marital	Income	
2	1	2	24	1	150000	
3	2	1	52	1	345000	
4	3	2	65	3	45000	
5	4	1	35	3	245000	
6	5	1	42	1	23000	
7	6	1	25	1	670000	
8	7	2	23	2	345000	
9	8	2	63	1	156000	
10	9	1	41	2	65300	
11	10	2	48	1	150000	
12	11	1	34	2	354000	
13	12	2	55	3	23000	
14	13	1	28	1	452000	

Fig. 4.13: Microsoft Excel - Book 1

Observe that the variable names are at the top row. Let us assume that the worksheet has been saved as Excel (with extension .xls) file called Profile.

4) Select **File → Open → Data** command in the SPSS Data Editor dialog box. The **Open File** dialog box appears. Choose the appropriate directory and folder in **Look** in box. Choose the Excel (+.xls) in the **File of type** box. Select the file name Profile.xls in **Filename** box.

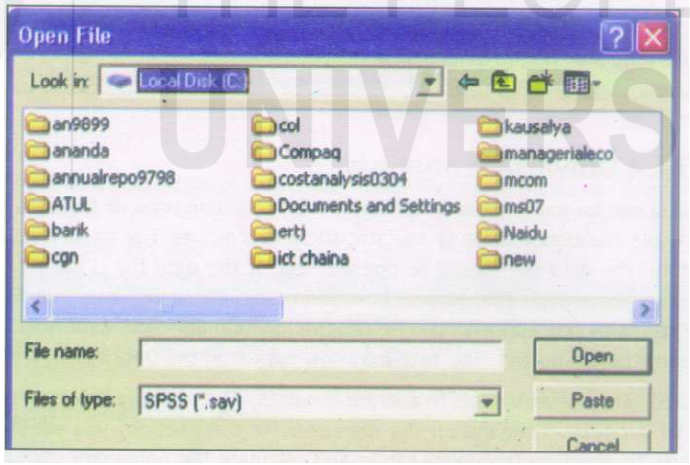


Fig. 4.14: Open Files

5) Click **Open** tab in the Open File dialog box. The **Opening Excel Data Source** dialog box appears on the screen. If your Excel file contains variable names, check mark **Read variable names from the first row of the data** box. If you leave the Range box blank, SPSS will read all the available data in the Excel worksheet. If you wish to read only some rows and columns then type a range. For example, you may type **A1:D30** to select first 4 columns (A,B,C, and D) and 30 rows (1 to 30). Click OK button to close the Opening Excel Data Source dialog box and return to the Data Editor dialog box. Observe that the data is in SPSS Data Editor. Save the SPSS data file.

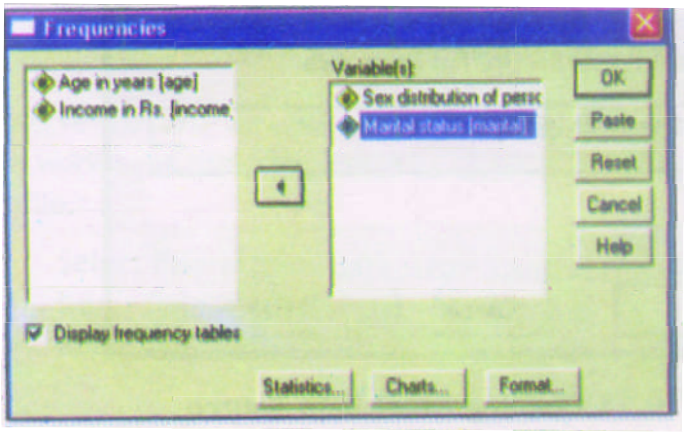


Fig. 4.15: Opening Excel Data Source

Reflection and Action 4.1

Answer the following two questions after going through the text above and viewing the graphics. The idea about answering these questions is to make sure that you have understood the procedures.

1) Explain briefly various steps in creating a data file.

2) Differentiate between SPSS Data Editor and Output SPSS Viewer.

4.4 UNIVARIATE ANALYSIS

Data can be analysed in a number of ways. You can present the data in a simple frequency table or use statistical techniques. For the analysis of data, the data file should be opened first. If the data file is not already open, select the **File → Open** command from the menu bar. The **Open** File dialog box appears. Choose the file location and SPSS data file name and click OK button. The data may now appear in the Data Editor screen.

SPSS offers several tools to analyse the data. The tools are selected from the **Analyze** command on the menu bar. In this section, you will learn how to generate frequency tables and calculate the univariate statistics.

Remember that you can compute the same statistics in SPSS using alternative commands.

Frequency tables

To generate the frequency tables,

1) Select **Analyse → Descriptive Statistics → Frequencies** command from the menu bar. The Frequencies dialog box appears on the screen with two large boxes for variables selection/deselection.

   Observe that the left box contains a list of variables for which data has been entered. On the right side, there is an empty box for the variable(s) that you want to include in the analysis.

2) In the left side box, select the variable to generate a frequency table by clicking the mouse on the name of the variable. Click the arrow between the two boxes. The selected variable appears on the right side box. Follow these steps till all the variables you want to include for the data analysis appear on the right side box.

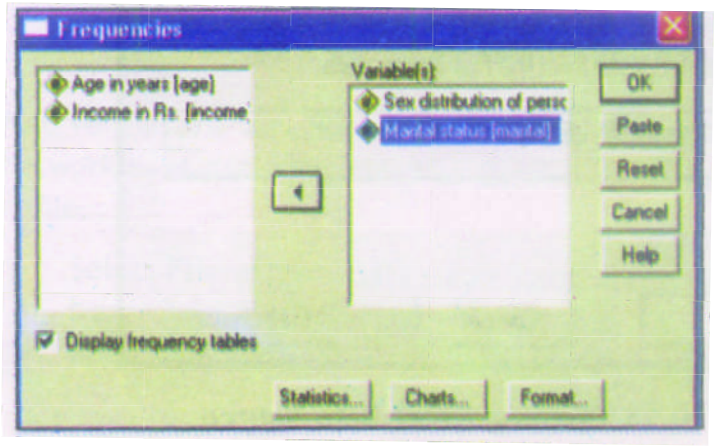


Fig. 4.16: Frequencies

3) When the first variable is moved to the right side box an arrow in the opposite direction appears between the two boxes. If you commit a mistake by selecting a wrong variable, click this arrow to return the variable to the original list.

Remember that this Frequencies tool is appropriate only for the categorical data (like Sex and Marital status in our example). Therefore, do not select any continuous variables (like in Age and Income in our example).

4) When you have selected all the variables, you want to include for data analysis, click the **OK** button. In the **Output SPSS Viewer window**, you should now see the output shown below.

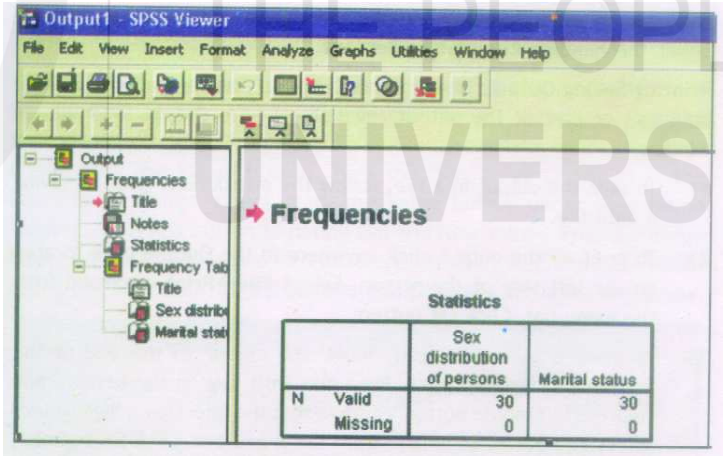


Fig. 4.17: Output SPSS Viewer

Frequencies

Find below two tables (Table 4.4 and 4.5 on Sex\Distribution of Persons and Marital Status, respectively).

Table 4.2 Sex\Distribution of Persons Statistics

		Sex Distribution of Persons	Marital Status
N	Valid	30	30
	Missing	0	0

Table 4.3 Marital Status

Frequencies table

Sex distribution of persons

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	**Male**	14	46.7	46.7	46.7
	Female	16	53.3	53.3	100.0
	Total	30	100.0	100.0	

Table 4.4 Marital Status Marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	**Married**	13	43.3	43.3	43.3
	Unmarried	10	33.3	33.3	76.7
	Widowed	7	23.3	23.3	100.0
	Total	30	100.0	100.0	

Observe that the tables contain all the information you want. Also, you have information on the missing cases for each variable. You also have three types of percentages: one for all the cases including missing cases under the heading **Percent,** the second one for only valid cases under the heading **Valid Percent,** and the third gives cumulative percentage under the heading **Cumulative Percent.**

Printing/Saving Output: Now you may want to print on paper or save in a file all or part of the output available in the Output SPSS Viewer window.

1) To save the output in a file, follow the instructions given at saving a data file.

2) To print all the output click anywhere in the Outline pane located on the left side of the screen. Select **File → Print** command from the menu bar. Click **OK** button.

3) To print a part of output, move the cursor to the end of the portion you want to print. Press the Shift key on the keyboard and click the left mouse button. Observe that the selection is highlighted. Select **File → Print** command from the menu bar. Click **OK** button.

Recode Data: you may want to recode your data for a variety of reasons. For example, the data values for the variable Age are continuous. Now you may want to group them like,

Old Value	New Value
Less than 19	1
20-30	2
30-39	3
40-49	4
50-59	5
60 and above	6

Table 4.5 Data values

To recode a variable, select **Transform → Recode → lnto Different Variables** command from the menu bar. The **Recode into Different Variables** dialog box appears on the screen.

Select the variable you want to recode into different variables in the left side box. Transfer this variable to the right side box using arrow tab that lies between the two boxes. Type the name of the variable in the **Name** text box and label name in the **Label** text box.

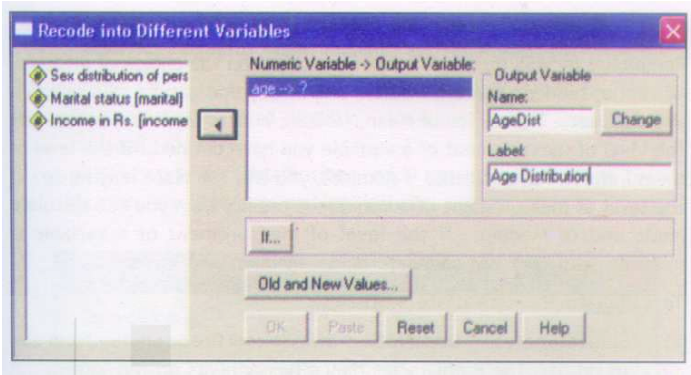


Fig. 4.18: Recode into Different Variables: Old and New Values

Press the **Old and New Values** button in the Recode into Different Variables dialog box. The **Recode into Different Variables: Old and New Values** dialog box appears.

Check mark the **Range** button. Type the first range of values in the boxes under Range. Check mark the **Value** button under **New Name** heading. Press Add button to define Old and New values. Press **Continue** button to close the Recode into Different Variables: Old and New Values dialog box and return to the Recode into Different Variables dialog box.

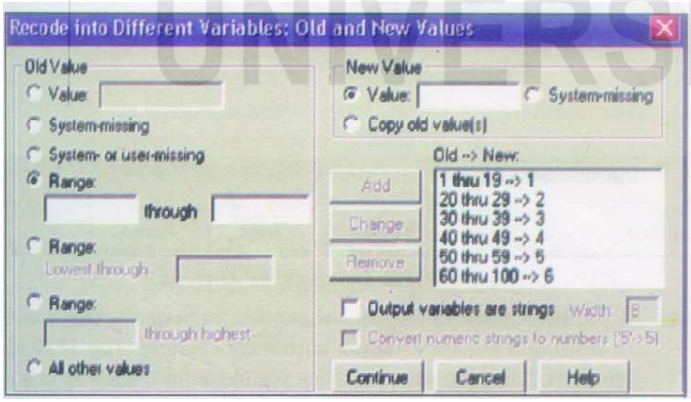


Fig. 4.19: Recode into Different Variables

If you have more variable(s) to redefine, continue these steps for each variable. Otherwise press **OK** button to close the Recode into Different Variables dialog box.

Univariate statistics

For each variable of your data set, you can calculate:

a) Measures of central tendency: Mean, Median, and Mode.

b) Dispersion: Standard deviation, Variance, Range, Standard Error of Mean, etc.

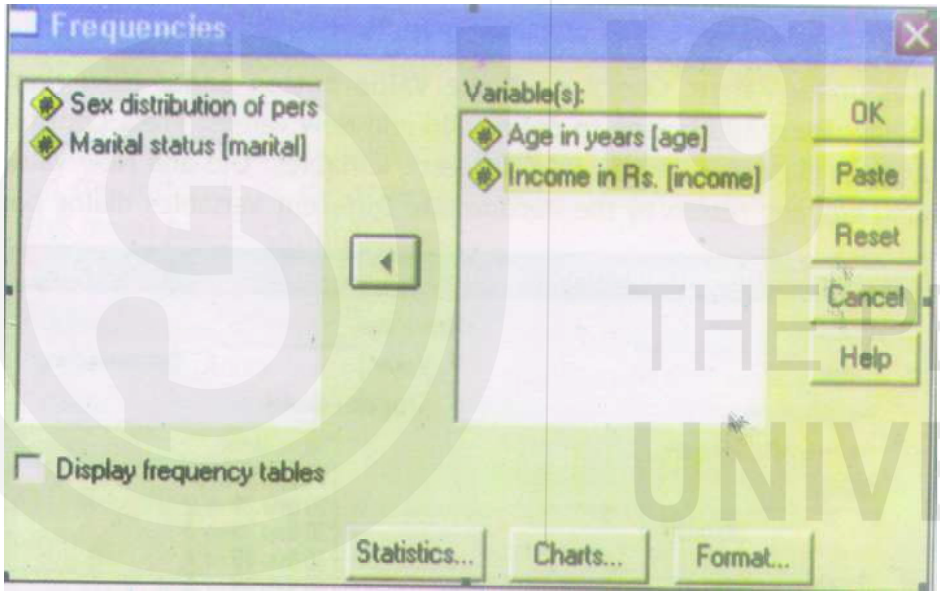
c) **Distribution: Kurtosis and Skewness**

Remember in SPSS there are some restrictions on the choice of measures of central tendency (Mean, Median, and Mode) that can be calculated on any data set. The choice of Mean, Median, and/or Mode is restricted by the level of measurement of a variable you have defined. If the level of measurement for a variable is nominal, you can calculate only mode. If the level of measurement of a variable is ordinal then you can calculate Mode and/or Median. If the level of measurement of a variable is interval/ratio, you can calculate Mode, Median, and/or Mode.

To calculate the univariate statistics:

1) Select **Analyse → Descriptive Statistics → Frequencies** from the menu bar. The Frequencies dialog box appears on the screen.

2) Transfer the variables on which you want to perform the data analysis from left side box to right side box (as you have done for frequencies analysis earlier).



Fi. 4.20: Frequencies Analysis

3) If you don't want to display frequencies, remove the check mark in **Display frequency tables** button by clicking. The **SPSS for Windows** dialog box appears asking you to confirm. Click **OK** button to close that window.

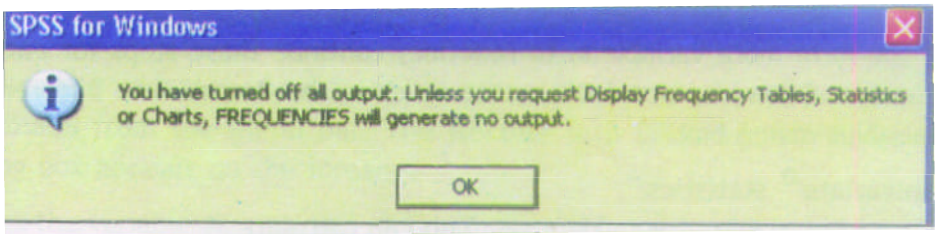


Fig. 4.21: Closing SPSS for Windows

4) Click on the **Statistics** button in the Frequencies dialog box. The **Frequencies:Statistics** dialog box appears on the screen.

5) In the area under **Central Tendency**, check mark the appropriate button to calculate Mean, Median, and/or Mode.

6) In the area under **Dispersion**, check mark the appropriate buttons to calculate Standard deviation, Variance, Range, Standard error or Mean, etc.

7) In the area under **Distribution**, check mark the appropriate buttons to calculate Skewness and/or Kurtosis.

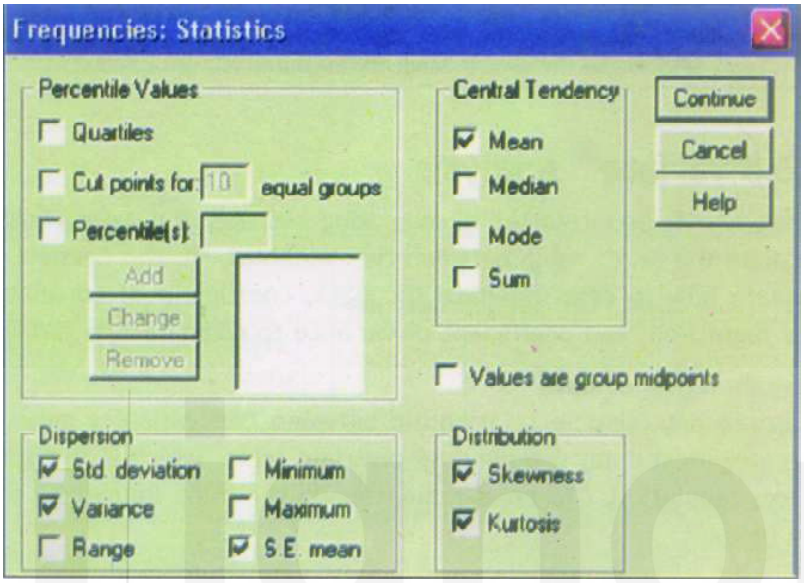


Fig. 4.22: Frequencies Statistics

8) Click on **Continue** button to close the **Frequencies: Statistics** dialog box.

9) Click on OK button in the Frequencies dialog box to close it. Observe that now the output is Descriptive Statistics.

Table 4.6 Statistics

		Age in Years	Income in Rs.
N	Valid	30	30
	Missing	0	0
Mean		40.83	234810.00
Std. Error of Mean			Z.891
35206.370			
Std. Deviation		15.836	192833.232
Variance		250.764	37184655414
Skewness		.531	1.182
Std. Error of Skewness		.427	.427
Kurtosis		-1.012	1.059
Std. Error of Kurtosis		.833	.833

Remember that you should opt for only appropriate statistics. For example, there is no meaning in opting for Mean of a sex Variable since there is nothing like mean of a sex distribution.

Reflection and Action 4.2

You have just finished reading about univariate analysis in which you worked on frequency tables and univariate statistics. In the light of this information answer the following questions.

- What is the command on the SPSS menu bar to perform frequencies data analysis?
- Once the data has been coded and entered into the SPSS Data Editor, is it possible to recode the data? If yes, what is the command to recode?
- You have defined the level of measurement of a variable as ordinal. Is it possible to calculate all the measures of central tendency for this variable using SPSS. Name the central tendency measures you can calculate.

4.5 BIVARIATE ANALYSIS

Often, you may be interested in comparing two sets of data or Variables to explore the relationship between two Variables. In this Section, you will learn how to cross-tabulate the data, coefficient of correlation, linear regression, and coefficient of variance to compare two variables.

Cross-tabulation of data

To capture any possible relationship between two variables measured with categorical data, you may use bivariate table, which is also known as cross-tabulation. To cross-tabulate two variables follow the steps given below.

1) Select **Analyse → Descriptive Statistics → Crosstabs** command from the menu bar. The **cross tabs** dialog box appears on the screen.
2) Click on the Variable in the source list (from left side box) that will form the rows of the table. Shift this Variable to the box under **Row(s)** using arrow key.
3) Similarly, shift the variable that will form the columns of the table from source list to the box under column(s).

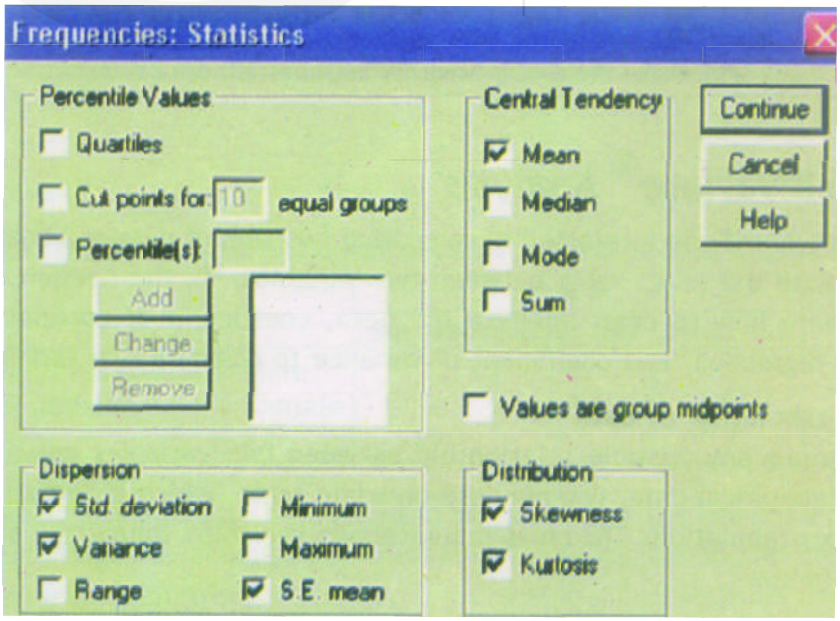


Fig. 4.23: Crosstabs

4) Click OK button. You will find the table in the Output viewers window.

Crosstabs

Table 4.7 Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Sex distribution of persons© Marital status	30	100.00%	0	.0%	30	100.0%

Count

Table 4.8 Sex distribution of persons' Marital status Crosstabulation

		Marital Status			
		Married	Unmarried	Widowed	Total
Sex distribution of persons	Male	6	6	2	14
	Female	7	4	5	16
Total		13	10	7	30

Sometimes, you may be interested in calculating the row/column/total percentages in a table. For this, click on **Cells** tab in the cross tabs dialog box before step-4 above. The **Crosstab: Cell Displays** dialog box appears on the screen.

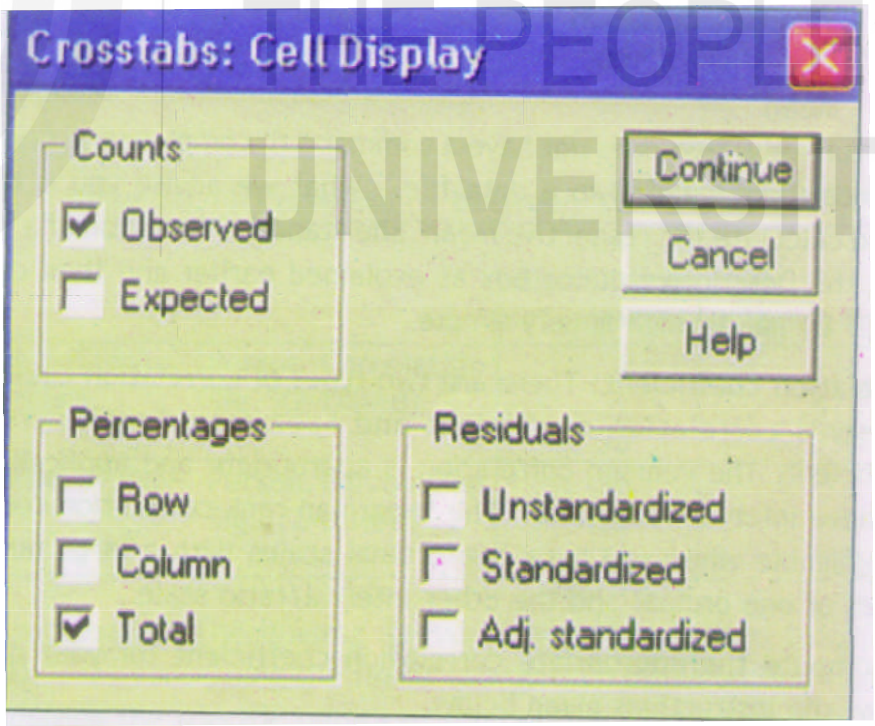


Fig. 4.24: Crosstabs: Cell Display

Check mark the appropriate (Row, Column, and/or Total) button under **percentages** area. Click **continue** button to close the Crosstabs: Cell Display dialog box. Click Ok button .on the Crosstabs dialog box, to view the output (see Table 4.9 and Table 4.10).

Table 4.9 Case Processing Summary: Output

			Cases			
Sex distribution of persons* Marital status	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
	30	100.0%	0	.0%	30	100.0%

Table 4.10 Sex distribution of persons *Marital status Crosstabutation: Output

			Marital status			
			Married	Unmarried	Widowed	Total
Sex distribution of perons	Male	Count	6	6	2	14
		% of Total	20.0%	20.0%	6.7%	46.7%
	Female	Count	7	4	5	16
		% of Total	23.3%	13.3%	16.7%	53.3%
		Count	13	10	7	30
		% of Total	43.3%	33.3%	23.3%	100.0%

Bivariate statistics

The statistical tools you often use to compare two Variables may be the coefficient of variance, correlation, and linear regression.

Coefficient of variance: As you are aware, the Coefficient of Variance (CV) is the standard deviation expressed as a percentage of the mean.

$$CV = \frac{\text{Standard deviation}}{\text{Mean}} \times 100$$

Unfortunately, SPSS does not have a command to complete the Coefficient Variance for a variable in a data file. What we advise you is that you should calculate the respective Mean and Standard deviation of a Variable using the Descriptive dialog box as explained earlier and then calculate the CV by hand which is very simple.

Correlation coefficient: There are two types of correlation coefficients: Pearson's correlation coefficient and Spearman's rank correlation coefficient. The Pearson correlation is appropriate and applicable when you have interval/ ratio data. The Spearman rank correlation coefficient is applicable when you have two ordinal scales with a large number of values or one ordinal and the other interval/ratio scale.

To compute the appropriate correlation coefficient for your data set, follow the instructions given below.

1) Select **Analyse** → **Correlate** → **Bivariate** command from the menu bar. The Bivariate Correlations dialog box appears on the screen.

2) Select the variables by shifting from left side box to the box under **Variables** area.

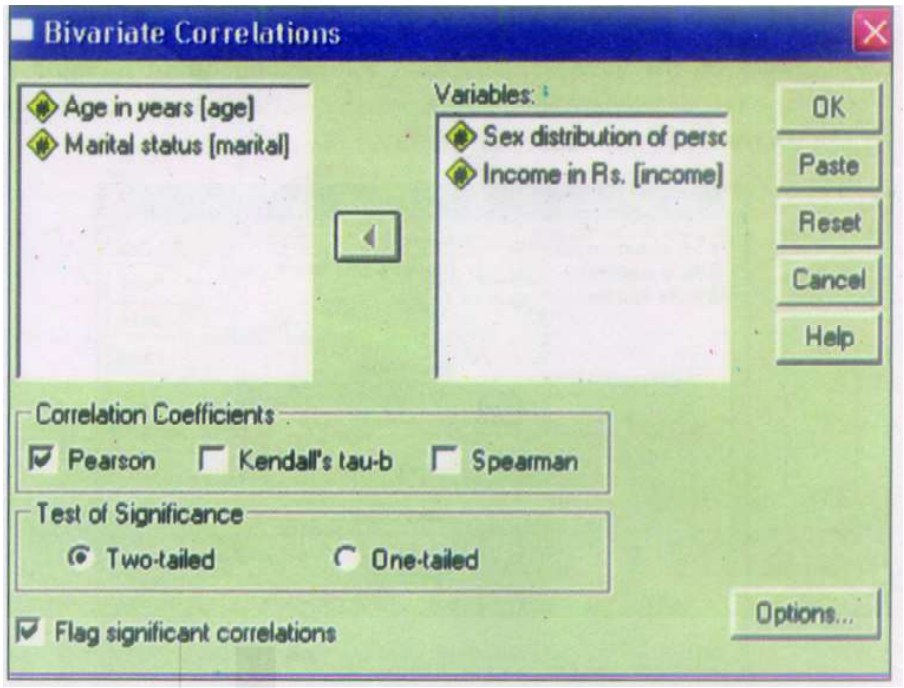


Fig. 4.25: Bivariate Correlation

3) Check mark the appropriate (Pearson, and/or Spearman) button under **Correlation Coefficients** area to select the type of correlation coefficient.

4) If you want the associated tests of significance, check mark the appropriate (Two-tailed or one-tailed) button under Test of Significance area. You will see the results in the Output Viewer Window.

Correlations

Table 4.11 Correlations between Two Variables

		Sex distribution of persons	Income in Rs.
Sex distribution of persons	**Person Correlation Sig. (2-tailed) N**	1.30	-.136 .472 30
Income in Rs.	**Person Correlation Sig. (2-tailed) N**	-.136 .472 30	1.30

Linear regression: The linear regression technique is used to: (a) test the hypotheses concerning the linear relationship between two variables (b) estimating the specific nature of relationship; and (c) to predict the values of dependent variable when you know the values of independent variable. To run the linear regression procedure follow the steps given below.

1) Select **Analyse** → **Regression** → **Linear** command from the menu bar. The Linear Regression dialog box appears on the screen.

2) Dick on the variable name that will be **dependent variable** in the left side box. Shift the dependent variable to the box under Dependent area using arrow tab.

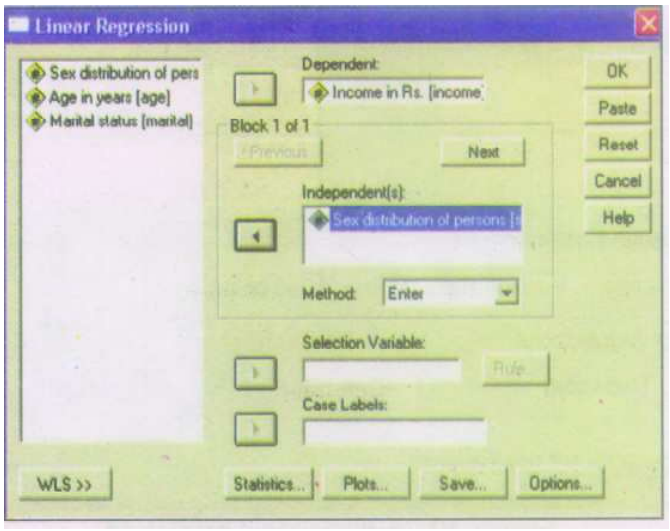


Fig. 4.26: Linear Regression

3) Click on the variable name that will be independent in the left side box. Shift the independent variable to the box under **InDependent(s)** area using arrow tab.

4) Click **OK** button. You will see the results in the Output Viewer Window.

Observe that the output consists of four points: (a) a table of variables used in regression, (b) a model summary, (c) an ANOVA table; and (d) a table of coefficients. You may be interested in a portion of the output. We will explain how to select a partial output in another Unit on use of SP5S in report writing.

Regression

Table 4.12 Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	Sex distribution of persons		Enter

a) All requested variables entered
b) Dependent Variable: Income in Rs.

Table 4.13 Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of theEstimate
1	.136a	.019	-.016	194413.110

a) Predictors: (Constant), Sex distribution of persons

Table 4.14 ANOVA

	Model	Sum of Squares	df	Mean Square	F	Sig.
1)	Regression Residual Total	2.01E+10 1.06E+12 1.08E+12	1 28 29	2.005E+10 3.780E+10	.531	.472a

a) Predictors: (Constant), Sex distribution of persons

b) Dependent Variable: Income in Rs.

Table 4.15 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1)	(Constant)	314275.0	114722.6	-.136	2.739 -	.011
	Sex distribution of persons	-51825.0	71147.913		.728	.472

a) Dependent Variable: Income in Rs.

4.6 MULTIVARIATE ANALYSIS

Sometimes, you may be interested in exploring complex relationships involving more than two variables. In this section, you will learn how to cross tabulate your data for the analysis of relationships between more than two variables. Similarly, how to compute multiple regression analysis has also been covered.

Elaboration of cross-tables

In the earlier section, you have learned the cross-tabulation of two variables. You can introduce a third variable by sub grouping one of the two variables. This can be done by introducing a variable as control variable. A control variable decomposes the data into sub-groups based on the categories of the control variable. To add a control variable for your cross-tabulation, follow the steps given below.

1) Select **Analyse → Descriptive Statistics → Crosstabs** command from the menu bar. The **Crosstabs** dialog box appears on the screen as shown on the next page.

2) Click on the variable in the source list that will form the row(s) of the table. Shift this variable to the box under Row(s) using arrow key.

3) Similarly, shift the variable that will form the columns of the table from source list to the box under **Column(s)** using arrow key.

4) Click on the variable which will act as control variable I (this variable splits the variable selected at step-3 into sub groups). Shift the control variable to the box under **Layer 1 of 1** using arrow key.

5) For computing the row/column/total percentages in the table,

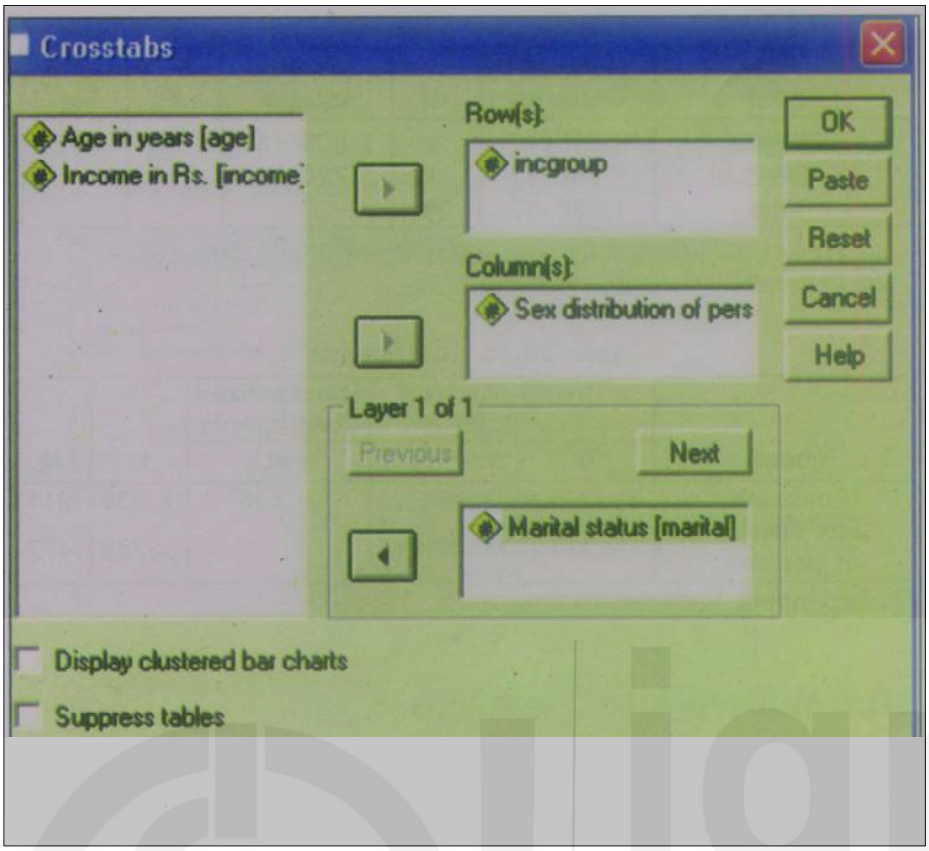


Fig. 4.27: Linear Regression Crosstabs; Cell Display for Computing the Row/ Column/Percentages

Click on **Cells...**tab. The **Cross tabs: Cell Display** dialog box appears. on the screen. Click the appropriate (Row, Column, and/or Total) button under **Percentages** area. Click **Continue** button to close the Cross tabs: Cell Display dialog box.

6) Click **OK** button to close the Crosstabs dialog box.

Crosstabs

Table 4.16 Case Processing Summary

Income group * Sex distribution of persons * Marital status	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
	30	100.0%	0	.0%	30	100.0%

Let's look at the table 4.17; Income Group and Sex Distribution of Persons' Marital Status Crosstabulation. You will find crosstabulation of marital status by income group of married, unmarried and widowed persons with their sex distribution. This is a good example that you can try and use in your own research work. Not only this but you can also use the other examples given in this unit for your research project.

Table 4.17 Income Group and Sex Distribution Persons' Marital Status Crosstabulation of

Marital status				Sex distribution of persons		Total
				Male	Female	
Married	Income Group	Less than Rs. 50000	Count % of Total	1 7.7%	2 15.4%	3 23.1%
		Rs. 100001-200000	Count % of Total	0 .0%	3 23.1%	3 23.1%
		Rs. 200001-400000	Count % of Total	3 23.1%	1 7.7%	4 30.8%
		More than Rs. 400000	Count % of Total	2 15.4%	1 1.1%	3 23.1%
	Total		Count % of Total	6 46.2%	1 53.8%	13 100.0%
Unmarried	Income group	Less than Rs. 50000	Count % of Total	0 .0%	1 10.0%'	1 10.0%
		Rs. 50001100000	Count % of Total	1 10.0%	1 10.0%	2 20.0%
		Rs. 100001-200000	Count % of Total	3 30.0%	1 10.0%	4 40.0%
		Rs. 200001-400000	Count % of Total	1 10.0%	1 10.0%	2 20.0%
		More than Rs. 400000	Count % of Total	1 10.0%	0 .0%	1 10.0%
	Total		Count % of Total	6 60.0%	4 40.0%	10 100.0%
Widowed	Income	Less than Rs. 50000	Count % of Total	0 .0%	2 28.6%	2 28.6%
		Rs. 100001-200000	Count % of Total	1 14.3%	0 .0%	1 14.3%
		Rs. 200001-400000	Count % of Total	1 14.3%	2 28.6%	3 42.9%
		More than Rs. 400000	Count % of Total	0 .0%	1 14.3%	1 14.3%
	Total		Count % of Total	2 28.6%	5 71.4%	7 100.3%

Multiple regression

In two variable linear regression you have used one dependent variable and one independent variable. The multivariate regression is used to investigate the relationship between two or more independent variables on a single dependent variable. The procedure for computing the statistics for multiple regression is the same as that for two variable linear regression explained earlier, except that you have more than one variable under Independent(s) area in the Linear Regression dialog box.

1) **Analyse → Regression-Linear** command from the menu bar. The **Linear Regression** dialog box appears on the screen.

2) Click on the variable name that will be a dependent variable in the left side box. Shift this variable to the box under **Dependent** area using arrow tab.

3) Click on the variable name that will be independent in the left side box. Shift this variable to the box under **Independent(s)** area. Follow this step until all the desired independent variables are selected.

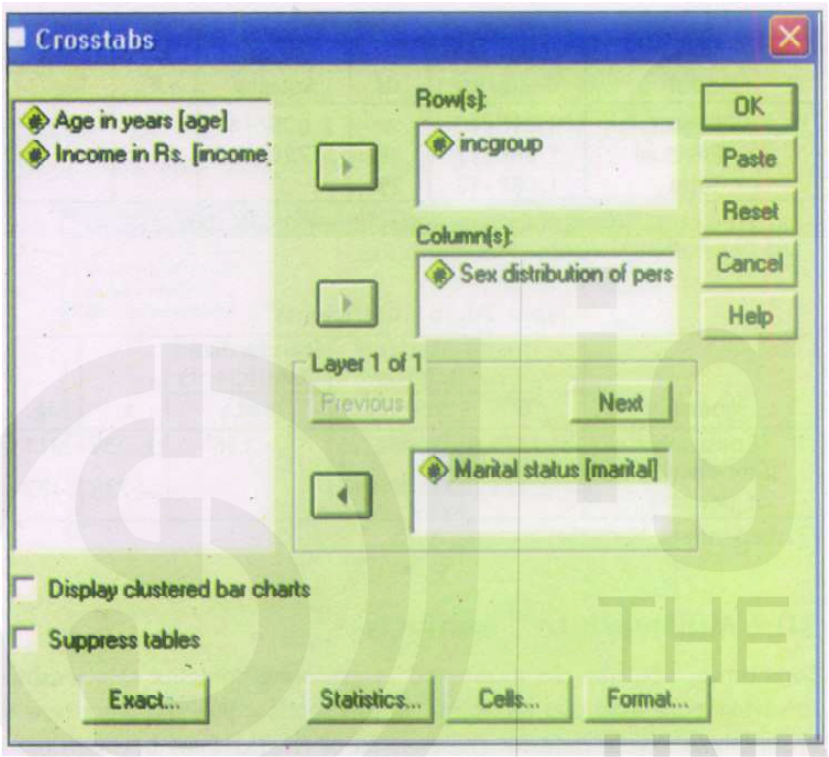


Fig. 4.28: Linear Regression Dialog Box

4) Click **OK** button. You will see the results in the Output Viewer Window.

Regression

Table 4.18 Variables Entered/Removed

Model	Variables Entered	VariablesRemoved	Method
1)	Marital status, Sex distribution of persons, Age in yeas Enter		

a) All requested variables entered
b) Dependent Variable: Income in Rs.

Table 4.19 Model Summary

Model	R	R Square	Adjusted R Sqaure	Std. Error of the Estimate
1	.168a	.028	-.084	200774.054

Table 4.20 ANOVA[b]

Model	Sum of Squares	df	Mean Square	F	Sig.
1) Regression	3.03E+10	3	1.010E+10	.250	.860[a]
Residual	1.05E+12	26	4.031E+10		
Total	1.08E+12	29			

a) **Predictors:** (Constant), Marital status, Sex distribution of persons, Age in years left side box. Shift this variable to the box under Test Variable(s) area.

b) **Dependent Variable:** Income in Rs.

Table 4.21 Coefficients

Model	Unstandardzed Coefficients		Standardzed Coefficients		
	B	Std. Error	Beta	t	Sig.
1) (Constant)	362849.3	153394.2		2.365 -	.026
Sex distribution of persons	-44008.8	75727.227	-.116	.581	.566
Age in years	-680.236	2431.316	-.056	-.280 -	.782
Marital status	-18212.7	46773.389	-.076	.389	.700

a) Dependent Variable: Income in Rs.

4.7 TESTS OF SIGNIFICANCE

In this section, you will learn one sample t-test for a mean, two sample t-test for the equality of means, and chi-square test for independence.

One sample t-test

The one sample t-test compares the sample means to a population mean, using t-distribution as the standard of comparison.

1) Select **Analyse → Compare Means → One-Sample TTest** command from the menu bar. The **One Sample T Test** dialog box appears on the screen.

2) Click on the variable name you want to perform the t-test on the left side box. Shift this variable to the box under **Test Variable(s)**.

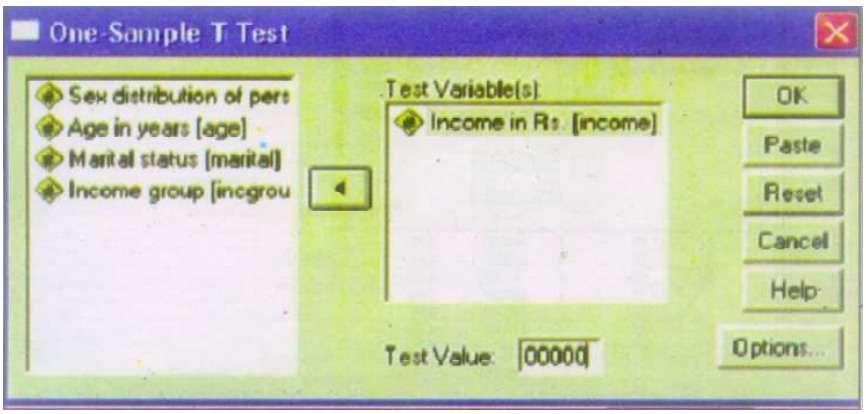


Fig. 4.29: One Sample T Test

3) Suppose you have selected Income Variable for the t-test and hypothesis takes the population mean income as Rs.200000. Type 200000 in the test box next to **Test Value**.

4) Click **OK** button. The following is the output you will see in Output Viewer Window.

T-Test

Table 4.22 One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Income in Rs.	30	234810.00	192833.232	35206.370

Table 4.23 One Sample Test with Test Value

	Test Value = 200000					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Income in Rs.	.989	29	.331	34810.00	-37195.11	106815.11

Two sample T-Test

Sometimes, our data analysis focuses on two distinct groups within a single population or we may want to compare the two populations in terms of their respective means. The two-sample t-test for the equality of means will help you in this.

1) Select **Analyse → Compare Means → Independent-Samples T Test...** command from the menu bar. T**he Independent-Samples T** dialog box appears on the screen.

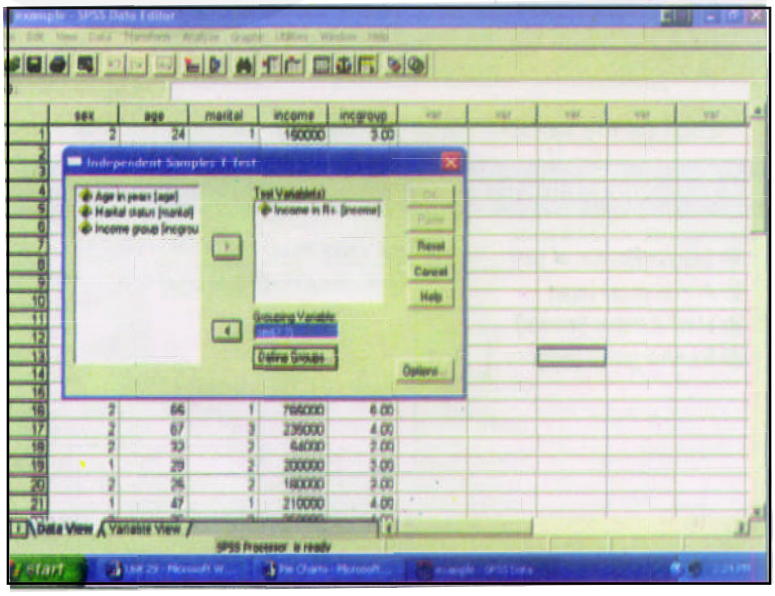


Fig. 4.30: Example SPSS Data Editor: Income Variable

2) Click on the Variable name in the left side box that you want to include for the t-test. Suppose you want to include Income Variable, shift this variable to the right side box under **Test Variable(s)** using arrow key.

3) Click on the variable you want to group. Suppose you want to test the income mean difference between males and females, click on sex variable. Shift this variable to the box under **Grouping Variable** using arrow key.

4) Click on the **Define Groups…**button. The **Define Group** dialog box appears on the screen. Type "T in the **Group 1** box. Type '2' in the **Group 2** box. Click **Continue** button to close the Define Groups dialog box.

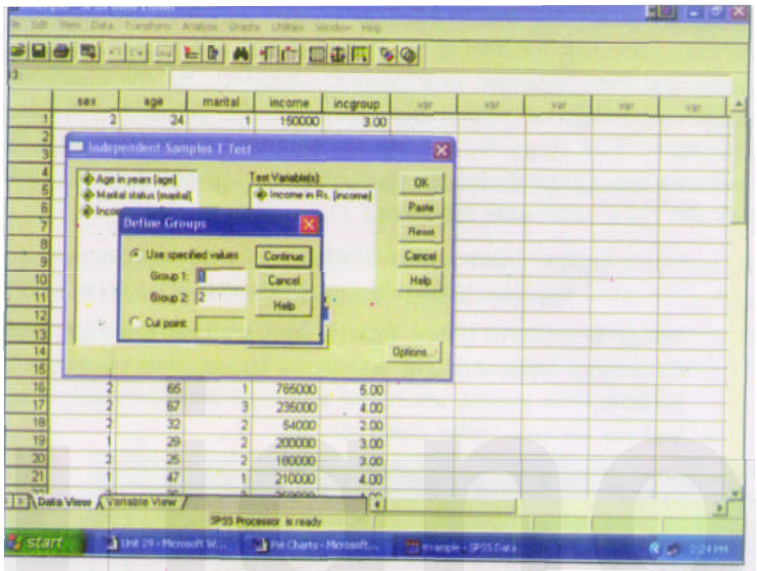


Fig. 4.31: Example SPSS Data Editor: Grouping Variable

5) Click **OK** button to close the Independent-Samples T Test dialog box. The following table presents the output.

T-test

Table 4.24 Group Statistics

	Sex distribution of persons	N	Mean	Std. Deviation	Std. Error Mean
Income in Rs.	**Male**	14	262450.00	176610.187	47201.058
	Female	16	210625.00	208616.994	52154.248

Table 4.25 Independent Samples Test

			Income in Rs.	
			Equal variances assumed	**Equal variances not assumed**
Levene's Test for Equality of Variancest-test for Equality of Means	F		.118	
	Sig,		.734	
	t		.728	.737
	df		28	27.978
	Sig. (2-taited)		.472	.467
	Mean Difference		51825.00	51825.00
	Std. Error Difference		71147.913	70342.061
	95% Confidence Interval of the Difference	Lower	-93914.893	-92269.306
		Upper	197564.893	195919.306

Chi-square test for independence

The chi-square test for testing the independence of attributes is for the categorical data arranged in a cross tabulation. The chi-square test appears as an option within the procedure for generating a cross-tabulation. The steps for chi-square test are repeat of steps for generating cross-tabulation with the addition of Expected counts under cells and chi-square under Statistics.

1) Select **Analyse → Descriptive Statistics → Crosstabs** command from the menu bar. The cross tabs dialog box appears on the screen.

2) Select a variable under **Row(s)** and another variable under **Column(s)**. Suppose you have selected Income level (a categorical variable with income levels low and high) variable under Row(s) and sex variable under Column(s).

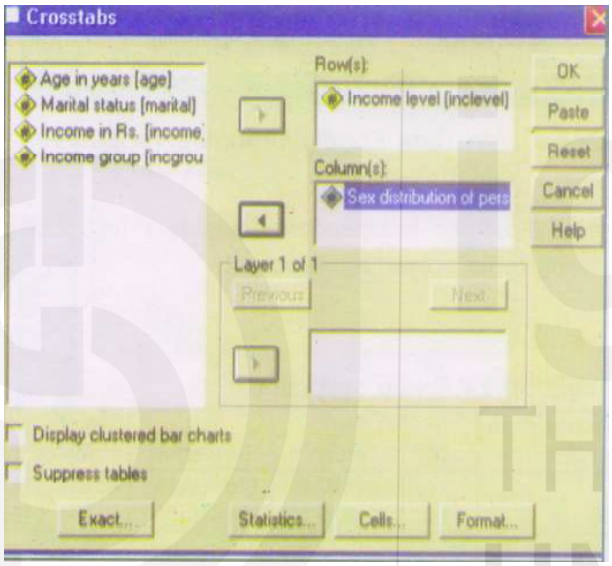


Fig. 4.32: Crosstables: Incomes Levels

3) Click on the Statistics...button in the Cross tabs dialog box. The Cross tabs:Statistics dialog box appears on the screen. Check mark the Chi-square button. Click Continue button to close the Cross tabs: Statistics dialog box.

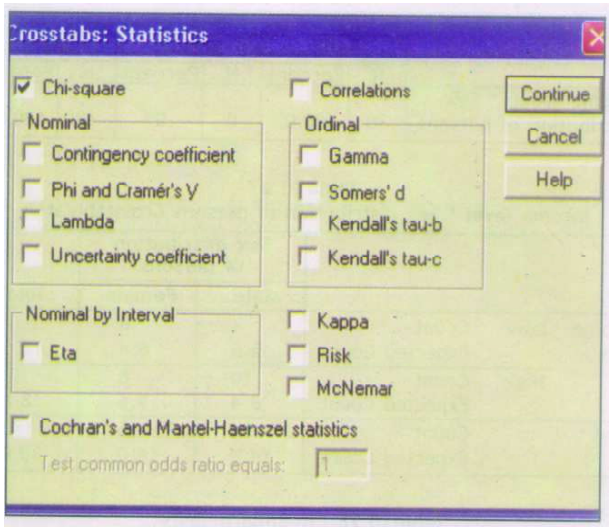


Fig. 4.33: Crosstabs: Statistics

4) Click on the **Cells** button in the Cross tabs dialog box. The **Crosstabs: Cell Display** dialog box appears on the screen. Check mark both Observed (if not already check marked) box and Expected box under Counts. Click **Continue** button to close **Cross tabs: Cell Display** dialog box.

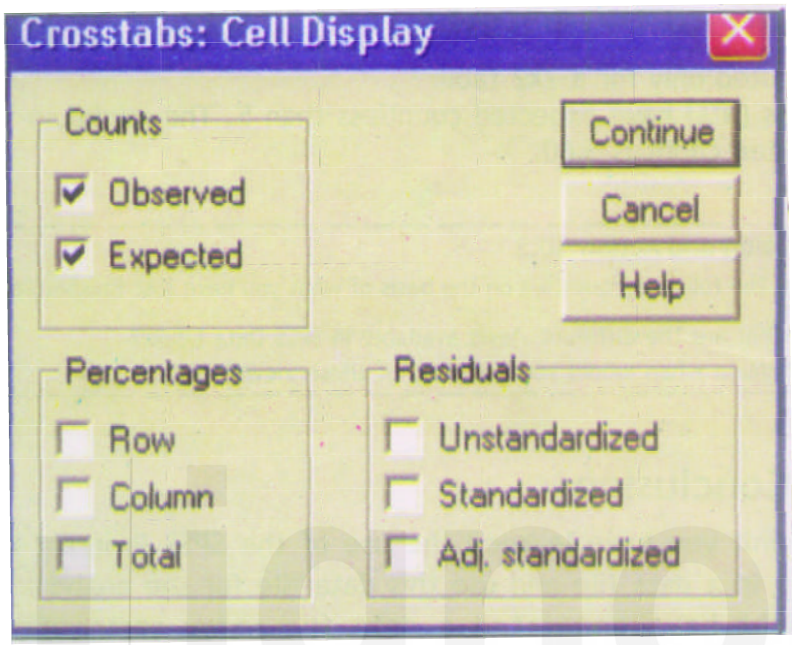


Fig. 4.34: Crosstabs: Cell Display

5) Click OK button in the Cross tabs dialog box. The output generated is shown below.

Crosstabs

Table 4.26 Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Income group * Sex distribution of persons	30	100.0%	0	.0%	30	100%

Table 4.27 Income level * Sex distribution of persons Crosstabulation

			Sex distribution of persons		
			Male	Female	Total
Income level	Low	Count	4	8	12
		Expected Count	5.6	6.4	12.0
	High	Count	10	8	18
		Expected Count	8.4	9.6	18.0
Total		Count	14	16	30
		Expected Count	14.0	16.0	30.0

Table4.28 Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)	Exact Sig. (1-sided)
Pearson Chi-Square	1-429	1	.232		
Continuity Correction[a]	.675	1	.411		
Likelihood Ratio	1.448	1	.229		
Fisher's Exact Test				.284	.206
Linear-by-Linear Association	1.381	1	.240		
N of Valid Cases	30				

a) Computed only for a 2×2 table

b) 0 cells (.0%) have expected countless than 5. The minimum expected count is 5.60.

Reflection and Action 4.3

Answer the following question on the basis of what you have just finished reading.

- What are the different views available in SPSS Data Editor?
- Explain when would you use each of these views?

4.8 CONCLUSION

In this unit you have learned the use of the SPSS Program to enter the data in a data file and use this data file for the analysis of data. You might have generated a data file using some other data base programs such as Excel. It is very easy to convert such data filers into a SPSS data file.

This unit provides an introduction to the SPSS. You can do a range of statistical analyses from simple cross tabulation to more complex statistical techniques, depending upon the individual researcher's requirement. However, we have tried to explain only simple commands and statistical tools, which are more popular in social research. We will leave it to the student to try and learn the full range of features in SPSS.

Suggested Reading

Nie, N. H., C.H. Hull, J. G. Jenkins, K. Steinbrenner and D. H. Bent 1979. *Statistical Package for the Social Sciences*. McGraw Hill: New York.