MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS &RESEARCH METHODOLGY: BSRM 201

Lesson No. 1 **AUTHOR: VIKAS KATIA**

STRUCTURE

1.0 OBJECTIVES

After reading this chapter, the reader should be able to:

- Understand the Objective, Role & Scope of Research Methodology.
- Realize the characteristics of Research.
- Comprehend the Research Process.

1.1 INTRODUCTION TO RESEARCH METHODOLOGY

Research is derived from the Latin word researcher which means to look for something which is hidden. It is composed of two words Re + search, Re means fact and search means to find. Conducting research is essential to make suitable and right decisions about specific problems.

The best course of action is always dependent on a good amount of organisation may take decision to price a product on the basis of market research conducted by a agency but a house wife may decide about the price of the product on the basis of her experience in purchasing commodities. Research is the solution provides to various problems eg if the problem is that why India is an under-developed economy. Through research various features of under development will be highlighted like more dependence on agriculture, less developed industrial sector, poor infrastructure etc. and

the real course of under development will be highlighted and steps will be taken on the basis of the findings of research to get rid of under development. The questions which are generally tend to answer are what, why how, where and who Rudyard (Kipling, a noted English poet wrote a piece of poetry which explains the insight) of research beautifully. He wrote:

I put six *honest learning men,*
They thought me all knew,
Their names are what and why and when, and how and where and who,

Def: Redman & Moxy, "**Research is a systematized effort to gain new knowledge."**
Research is effectively used in economics, commerce and management. Henry Ford, the master of famous Ford Motor Company once said, "Research is fundamental to everything we do, so much so that we hardly make any significant decision without the benefit of some kind of market research. The most important managerial value of business research is that it reduces the uncertainly and risk by providing information that improves the number of ways. First it may be used to define problems or to identify opportunities to enrich management efforts. A second reason for using business research is to explain why something went wrong. Detailed information about specific mistakes or failures is frequently sought. The basic objective of seeking answers to such questions is to identify the problems areas for the business.
Different type of research studies are designed for different type of situations and problems. For example, food processing industry introducing a new juice might be interested in knowing whether golden or silver packaging would be more effective. In this situation, the problem is clearly defined and a simple experiment may be designed to get the specific answer. However in some complicated situation, the industry may be totally unaware of a problem. For example, a manufacturer may notice that employee turnover has increased too much but the plant manager may be totally ignorant of the reasons for this trend. In such cases, explanatory search may be necessary to gain insights into the nature of the problem.

1.1 Self check questions

a) From where the word 'research' has derived from?

b) Explain the managerial value of business research?

1.2 CHARACTERISTICS OF RESEARCH

1. Research is directed towards the solution of a problem. The ultimate aim is to discover cause and effect relationships between variables.
2. Research emphasizes upon the development of generalizations, principles or theories that will be helpful in predicting future occurrence.
3. Research is based upon observed experience or empirical evidence.

4. Research demands accurate observations.
5. Research involves gathering of new data from primary or secondary sources or using existing data for a new purpose.
6. Research is more often characterized by carefully designed procedures that apply strict and impartial analysis of data.
7. Research requires expertise.
8. Research strives to objective and logical, applying every possible test to validate the procedures employed, the data collected and conclusions reached. The researcher attempts to eliminate personal biasness.
9. Research involves the quest for answers to unsolved problems.
10. Research is characterized by patient and unhurried activity.
11. Research is carefully recorded and reported,
12. Research sometimes requires courage.

1.2 Self check questions

a) What is the ultimate aim of research?

b) Why Research strives to be objective and logical?

1.3 SOME KEY CONCEPTS USED IN RESEARCH

One can understand the process of research methodology and the techniques involved in it lastly of some basic concepts are known to us. The following concepts are the most important:

(1) Theory

The word Theory has also been used in different ways in the different contexts. For our purpose a theory is coherent body of general propositions used as principles or explanation of the relationships of certain observed-phenomena. A key element in the above definition is the term, proposition. So let us see what a proposition is. A proposition is a statement concerned with the relationships among concepts. It has to be based on logic. A proposition states that every want either has a certain property or stands in a certain relationship to other events.

(2) Concept

A concept is a generalized idea about some occurrences or processes. It is based on empirical events. Concepts are expressed in words that refer to various events or objects. For example, a research concept of ‘asset’ is an abstract term that may in the concrete world of reality refer to a specific machine. Concepts may be framed at different levels of action. It may be based on propositions or empirical observations. Researchers generally try to explore those concepts which are based upon empirical observations and the theorists translate the conceptualizing of reality into abstract ideas. Only when we explain how concepts are related to other concepts then we begin to construct theories. Thus, the difference between a theory and a concept is that

concept it is a single phenomenon whereas a theory is an assimilation of many concepts.

(3) Hypothesis

It means tentative or assumed statements. A hypothesis is a proposition that is empirically testable. Every research starts with framing of a hypothesis which is nothing but a tentative conclusion. We try to test it with the help of statistical tools and if, it is tested successfully, we accept the hypothesis.

1.3 Self check questions

a) What is a hypothesis?

b) Explain concept.

1.4 OBJECTIVES AND IMPORTANCE OF RESEARCH

The importance of research can be well analyzed from a famous HUDSON Maxim. "All progress is born of enquiry. Doubt is often better than over confidence for it leads to enquiry and enquiry leads to inventions."

1. To achieve the solution of problem: The main objective of every type of research is to solve the problem. The whole process of research is concentrated towards giving the solution of the problem.
2. To achieve new insights: Research is done to achieve new insights or to know or to explore more and more about problem and have the deeper knowledge of any problem.
3. To determine the frequency: Research depicts the frequency or the no. of times any specific thing/problem occurs. With research frequency of problem can be read.
4. To test the hypothesis: To solve problem one of the step is to frame the hypothesis. Hypothesis is tentative statements that provide solution to the problem. Whole of the research is concentrated towards approving/disapproving the hypothesis.
5. Basis of frame for government policies: Research provides the base for government policies & budget e.g. population policy, industrial policy, agricultural policy, fiscal & monetary policy.

1.4 Self check questions

a) Do research forms a Basis of framing government policies? True/false

b) Research is done to achieve new insights have the deeper knowledge of any problem? True/false

1.5 THE PROCESS OF THEORY BUILDING

Many times a question is asked about the generation of theories. Although this is not an

easy question to answer but still we can look at the generation of theories through abstract conceptual and empirical level exploration. At the abstract level a theory may be developed with detective reasoning by going from a general statement to a specific assertion. Deductive reasoning is a logical process of deriving a conclusion from something known to be true. For example, we know that all managers are human beings if we know that Mr. X is a manager then we can deduce that Mr. X is a human being also.

The empirical level of theory may be developed with inductive reasoning which is a logical process of establishing a general proposition on the basis of observation of particular facts. For example, if a stock broker with 20 years experience of trading in a stock exchange repeatedly notices that the price of gold rises whenever there is some disturbance in the country, the stock broker may project this empirical observation in a generalized way and build a theory that price of gold is related with the political stability in the country. It has been generally found that a theory based on either empirical observation or deductive logic may not be a perfect theory especially in the field of social sciences. These sciences deal with human beings and their behaviour is bound to have variations of human nature. In pure sciences, the relationships are fixed, so theory building is more accurate and scientific. For example, two units of oxygen and one unit of hydrogen shall always make water but a good perfect training in management may not create a manager in the real sense. So in social sciences, theory construction is often the result of a combination of deductive and inductive reasoning. We draw conclusions on the basis of our experience and then verify these conclusions, known, as Research Methodology.

1.6 THE METHODOLOGY OF RESEARCH

It will be useful for you to look at the analytical process or the methodology used by conducting scientific research. It comprises of a series of stages. Here are the following seven steps involved in the process of research:

1. Assessment of existing knowledge.
2. Formulation of concepts and propositions,
3. Statement of hypothesis.
4. Designing research to test the hypothesis.
5. Collection of empirical data.
6. Analysis and evaluation of data.
7. Explanation, statement of solutions and problems raised by the research

1.7 RESEARCH PROCESS

Research is a search for knowledge. To gain this knowledge there are number of steps involved. These steps are not mutually exclusive, nor are they separate and distinct. They do not follow each other in a specific order. However the following sequence

provides useful procedural guidelines regarding the process.

1. Formulating the research problem
2. Extensive literature survey
3. Developing the hypothesis
4. Preparing the research design
5. Determining the sample design
6. Collecting the data
7. Execution of the project
8. Analysis of data
9. Hypothesis testing
10. Generalization and Interpretation
11. Preparation of report

First five steps i.e. from setting of the problem till determining the sample design are discussed in this chapter, rest are discussed in the later chapters.

1. Formulating the Research Problem

At the first glance, it would seem fairly easy to see and pose a problem for study. But the experience of researchers is summed up in the ad-age "It is often more difficult to find and frame the problem, rather than to solve it. Problem means what the researcher wants to solve. It is the main concentration of whole of the research work."

The problem is of two types:

- Which relate to state of nature? i
- Which relate to relationship between variables?

So whole of the study research is conducted after setting the problem. The problem once set is not rigid in nature, but problem can change also. So initially the problem is set in a broad way or in a general way after defining/ redefining the problem it can be formulated in a specific way.

Essentially two steps are involved in framing the problem:

(i) Understanding the problem thoroughly
(ii) Refreshing the same into meaningful terms from analytical point of view.

There are number of sources of selecting the problem:

(a) Existing trouble
(b) Literature study
(c) Discussions
(d) Expert advice
(e) Studies already made

2. Extensive Literature Survey

Means a broad/wide survey of literature on the selected problem. So whatever the

material on such specific topic or on other related fields is available are to be surveyed by the researcher. So for this abstracting and indexing journals bibliographies are the first place to go. Academic journals, conference proceedings, government reports, books must be tapped depending on the nature of problem. It should be remembered that in survey of literature one source will lead to other. So a good library is of immense use at this stage of surveying of literature.

3. Developing the Hypothesis

Hypothesis is derived from two words hypo + thesis. Hypo means tentative or assumed, thesis means statements. So hypothesis is a set of declarative statements or sentence which is to be proved or disapproved. So after setting the problem in hypothesis framing answers to such problems are decided in advance. Hypotheses are answered to such problems. After conducting an extensive, literature survey, the researcher should able to state in clear terms the working hypothesis. These are the assumptions made in order to draw out and test its logical consequences. Every hypothesis framed should possess the following features-clarity, simplicity, declarative sentence form, capable of testing etc. The role of hypothesis is to guide the researcher by determining the area of research and to keep him on the right track. It sharpens the thinking and focuses attention on the more important fact of the problem. It also indicates the requirement of type of data.

The hypothesis can be developed by:

- Discussing it with colleagues and experts about the problem, its origin and the objective in seeking the problem.
- Examination of data and records, if available concerning the problem for possible trend, peculiarities and other clues.
- Review of similar studies or related studies.
- Personal, investigation like interviews, surveys etc.

Thus hypothesis arise as a result of prior thinking about the subject, examination of available data and material including related studies and the counsel of experts and interested parties. Working hypothesis is more useful when stated in precise and clearly defined terms.

There are two types of hypotheses framed by researcher:

- Null hypothesis
- Alternative hypothesis.

Null hypothesis is that hypothesis which researcher wants to disapprove and Alternative hypothesis is one which the researcher wants to prove.

Another important concept in testing hypothesis is level of significance. It is always certain percentage chosen with great care. It means the chances of or willingness to take the risk by accepting null hypothesis e.g. if level of significance is 5% it means that there are 5% chances of accepting wrong hypothesis and 95% level of confidence

i.e. accepting a true hypothesis.
There are various steps to test a hypothesis:

1. Making a formal statement: Null hypothesis (HJ or Alternative hypothesis (H_a).
2. To select a level of significance.
3. To select the appropriate sampling distribution.
4. To select the random sample and compute appropriate value from sample data.
5. To calculate probability and compare probabilities with the relevant table values. Statisticians have developed several tests of hypotheses. Parametric and .Non- parametric tests so by using these tests one can test hypothesis to be true/false.

4. Preparing the Research Design

Research design is preparing the blue print for action i.e. how the research will be conducted or to state the conceptual structure within which research will be conducted. The preparation of such a design facilities research to be as efficient as possible yielding maximum information. In other words, the function of research design is to provide for the collection of relevant evidence with minimal expenditure of effort, time and money. But how all these can be achieved depend on research purpose. Research purpose may be grouped into four categories:

- Exploration
- Descriptive
- Diagnosis
- Experimental

So research design differs in all cases. A flexible design which provides opportunity for considering many different aspects of problem is considered appropriate if the purpose is that of exploration. But when purpose happens to be accurate description of a situation or of an association between variables, the suitable design will be one that minimizes bias and maximizes the reliability of data collected and analyzed. There are several designs, such as experimental and non-experimental hypothesis testing.
The preparation of research design, appropriate for a particular research problem,
Involves considering action of the following:

- The means of obtaining the information.
- The availability and skills of the researcher and his staff.
- Explanation of the way in which selected means of obtaining information will be organized and the reasoning leading to the action.
- The time available for research.
- The cost factor relating to the research

TYPES OF RESEARCH DESIGN

- Exploratory Research Design
- Conclusive Research Design
  Descriptive research Design
  Experimental research Design

5. Determining the Sample Design

A sample design is a definite plan for obtaining a sample from a given population. It refers to techniques or procedure or way the researcher will select units. As whole of the units or census method is impossible.

Research Design	Exploratory	Descriptive
Overall design	Flexible	Rigid
Sampling design	Non-probability	Probability
Statistical design	No preplanned	Planned
Observational design	Unstructured	Structured
Operational design	No fixed decision	Advanced Decisions

6. Analysis of Data

The analysis of data requires a number of operations such as establishment of categories, the application of these categories to raw data through coding, tabulation and then drawing statistical inferences.

7. Preparation of the Report

Research report is one of the vital aspects of research and is considered a major constituent of the research study, for the research task remains incomplete till the report has been presented and / or written. Writing of report is the last step in a research study and requires a set of skills somewhat different from those called for in respect of the earlier stages of research.

1.7 Self check questions

a) What is a sample design?

b) Explain research design.

c) What are the two types of hypothesis?

1.8 SUMMARY

Conducting research is essential to make suitable and right decisions about specific problems. The most important managerial value of business research is that it reduces the uncertainly and risk by providing information that improves the number of ways. Research is directed towards the solution of a problem. The ultimate aim is to discover

cause and effect relationships between variables. Research strives to objective and logical, applying every possible test to validate the procedures employed, the data collected and conclusions reached.

1.9 GLOSSARY

- **Research**: It is derived from the Latin word researcher which means to look for something which is hidden.
- **Hypothesis**: It means tentative or assumed statements. A hypothesis is a proposition that is empirically testable.
- **Null Hypothesis**: is that hypothesis which researcher wants to disapprove.
- **Alternative Hypothesis** : is that hypothesis which the researcher wants to prove

1.10 SHORT ANSWER QUESTIONS

1. What do you mean by hypothesis? Explain its types.
2. What is the difference between theory and a concept?
3. What is a research problem?
4. Explain how a hypothesis can be developed?

1.11 LONG ANSWER QUESTIONS

1. Explain the various characteristics of Business Research.
2. Describe in detail the Research Process.
3. Explain the importance of research by giving real life examples.
4. Explain the various types of research designs.

1.12 ANSWERS OF SELF CHECK QUESTIONS

1.1 a) Latin
b) Reduces risks and uncertainties by providing information
1.2 a) discover cause & effect relationships
b) To reduce biasness
1.3 a) tentative statements
b) Generalized idea about some occurrences
1.4 a) True
b) True
1.7 a) definite plan for obtaining a sample from a given population
b) Blue print for action
c) Null and alternate hypothesis

1.13 SUGGESTED READINGS

- Kothari C.R., **Research Methodology: Methods and Techniques,** New Age International Publishers, New Delhi, 2nd Edition, 2006.
- Sinha, S. C. & Dhiman, A. K., **Research Methodology.**

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS &RESEARCH METHODOLGY: BSRM 201

Lesson No. 2 **AUTHOR: ARUSHI**

RESEARCH DESIGN

STRUCTURE

2.0 OBJECTIVES

- To understand three different types of research approaches which form the research design.
- To understand the relationship between three types of research approaches.
- To understand the importance of literature review in the conduct of good research.
- To understand the structure in organizing literature review and some sources of collecting it.

2.1 INTRODUCTION

A research design is simply the framework or plan for a study used as a guide in collecting and analyzing data. A research design is like a blueprint used to complete a study. It indicates the

researcher the type of research approach to apply, sampling method to be used and data collection methods to be used for fulfilling research objectives. Most importantly, the blueprint of research helps the researcher in making one of the most significant decisions about selection of proper research approach because different research approach would entail different research purpose, research questions, the precision of hypothesis formulated and data collection methods.

2.1 Self check questions
a. What is a research design?
b. Why research design is important?

2.2 IMPORTANCE OF A GOOD RESEARCH DESIGN

A proper research design is helpful in answering questions about source and type of secondary data to be used, type of research to be applied and way of collecting primary data. A proper research design is important because selection of appropriate research approach and its associated data collection method involve tactical research decisions which are constrained by time and budget availability to do research. Some of the tactical research decisions involve

- choice of measurement methods to be used,
- structure and length of questionnaire,
- type of questions to be asked and
- sampling method to be applied to choose a sample as respondent for the study.

As these decisions involve expenditure of time and money so accurate or near accurate research design is very important. Also, all these tactical decisions about research approach, sampling methodology and data collection methods should fit and complement each other as choice or application of even one wrong decision would hamper proper fulfillment of research objective. In case of such an event research approach might have to be redesigned asking for extra time and budget resources making entire research process infeasible. Thus a research design ensures that the study:
(1) will be relevant to the problem, and
(2) will use economical procedures
In the light of importance of proper research process discussed above this chapter would discuss three types of research approaches and methods of secondary data collection. The acquisition of secondary data associated with a particular research problem is very important as it becomes part of review of literature. Suitable literature review is a significant part in formulating research questions and provide possible answers in the form of hypothesis. Also, appropriate literature review helps a researcher in deciding about data collection methods and analytical approach required to analyze and interpret the research problem at hand.

2.2 Self check questions
a. What a research design includes?

2.3 TYPES OF RESEARCH APPROACH

There is never a single, standard, correct method of carrying out research. One should not wait to start research until he/she find out the proper approach, because there are many ways to tackle a problem-some good, some bad, but probably several good ways. There is no single perfect design. Rather, there are many research design frameworks. Research designs can be classified into some basic types. One very useful classification is in terms of the fundamental objective of the research: exploratory, descriptive, or causal.'

2.3.1 Exploratory Research: This type of research approach is characterized by following theories:

- *to gain insights and ideas:* exploratory study is particularly helpful in breaking broad, vague problem statements into smaller, more precise sub-problem statements, for better understanding of the problem to formulate specific hypothesis.
- *establishing priorities for further research:* exploratory research could be helpful in indicating the importance of variety of discovered alternatives. The priorities would be established because a particular hypothesis discovered in the exploratory study appeared to be promising.
- *increasing the analyst's familiarity with the problem:* a new problem can be more clearly understood by the application of exploratory research as the approach helps in clarifying concepts about the research problem because of its investigative nature.
- *Flexibility:* to transform vaguely defined problem into more precise meaning, researchers frequently change the research procedure. Thus, exploratory studies use highly flexible, unstructured, and qualitative methods which would help the researcher to get insight into general nature of the problem.

2.3.2 Descriptive Research: This type of research approach is characterized by following theories:

- Descriptive research is used when enough prior information is available regarding the research problem which helps the researcher in guiding the research in specific direction.
- It is based on one or more specified hypothesis that was either discovered while conducting exploratory research or is established from secondary data.
- Descriptive studies are rigid as this approach uses structured, inflexible and quantitative methods to get results of a research problem.
- This approach helps in describe characterizing of certain groups. For example which income group has preference for what kind of product?

2.3.3 Causal Research: When it is necessary to show that one variable causes or determines the values of other variables, a causal research approach must be used. Descriptive research is not sufficient, for all it can show is that two variables are related or associated. Of course, evidence of a relationship or an association is useful; otherwise, we would have no basis for even inferring that causality might be present. To go beyond this inference we must have reasonable

proof that one variable preceded the other and that there were no other causal factors that could have accounted for the relationship. Descriptive research can be used for testing hypothesis. However descriptive designs are not as satisfactory as experiments for establishing causality.

2.3 Self check Questions

a. Finding out frequency of drinking herbal tea per day is a kind of …….
b. Finding out reasons for violent behavior among school going students is a kind of …….
c. Finding out changes in changes in birth rate because of increased literacy among women is a kind of …….

Let's understand the three research approaches by illustrating an example:

A researcher would like to study the relationship between per-capita income and literacy rate in a particular territory. Now, first step would be to collect secondary data regarding these variables and their relationship. Is prior information available indicating such a relationship? Is enough data available regarding per capita income of population of region under study? Is enough education avenues like schools and colleges available? Is the researcher studying literacy at primary education level or higher education level? Availability of literature regarding such questions would help the research to build tentative but informed hypothesis regarding the problem at hand. In such an event the study would be a descriptive study whereas the lack of such information would ask for an exploratory approach towards the research problem. Suppose we had evidence that territories with high per-capita income has higher literacy rates. Are there sufficient grounds for a decision to provide means to increase per-capita income so that literacy rates can also be increased? The answer would depend first on whether data is available authenticating such a relationship in other territories as well. Second, we would have to be sure that there were no other reasons for differences in literacy rates between territories. Perhaps the increase in literacy rate is due to more awareness about benefits of education or easy access to schools and colleges in a particular region. In such a scenario causal approach towards research would be more appropriate.

Are three type of research - exploratory, descriptive, and causal - has a distinct or complementary role to play in many research studies?

In most of the research problems, three types of research that have been discussed play a complementary role to each other. For instance, a researcher wants to find reasons for increased rate of drop outs among students in higher education. The first step is to use exploratory techniques to generate all possible reasons for the problem. Exploratory research would throw numerous reasons of increased dropout rate which would help in formulation of tentative hypothesis such as: high fee structure, lack of public transport which makes accessibility difficult, gender classification among dropouts, lack of job opportunities or simply lack of interest in higher education.

Thereafter, a combination of descriptive and causal approaches is used to narrow the possible causes. Descriptive research would help in testing the tentative hypothesis and generate relationship among deduced variables from exploratory research. Causal research would help in inferring the cause and effect among the relationship and would help the researcher to conclude the reasons behind high dropouts and how to reduce them by focusing on certain specific variables.

2.4 LITERATURE REVIEW

A literature review discusses published information in a particular subject area, and sometimes information in a particular subject area within a certain time period. The purpose of literature review is to identify what knowledge and ideas have been established on a topic, and what their strengths and weaknesses are. It is an organized body of information related directly or indirectly to the research questions.

2.5 WHY TO CONDUCT LITERATURE REVIEW

- Review of literature helps in seeking information from published sources and should be able to segregate relevant information from irrelevant. It should be relevant, helpful and useful for the current research and for the reader who uses it to find established sources.
- It helps in reorganization of accessed information by critically analyzing it. The way sourced information is reorganized is dictated by research objective or the research problem that is being studied. The quantity of literature review should be good enough to help the researcher to critically analyze the literature and help in assessing the strengths and weaknesses of research problem at hand.
- It helps in identifying areas of controversy in the literature which can be taken by the researcher for current research. It helps the researcher to find contrary perspective to the established theories so that he/she can be guided to do current research, thus, helping in formulating research questions and related hypothesis.
- It helps the researcher to answer specific questions relating to data collection methods, sampling methodology, type of research approach and analytical methods, thus, providing the researcher direction in current research.

2.6 ROLE OF LITERATURE REVIEW IN THE CONDUCT OF GOOD RESEARCH

For most of the graduate and post-graduate students thesis is a piece of work which is indicative of level of their intellectual skills and abilities. A key part of thesis which illustrates this is the review of literature. For this reason literature review should provide enough evidence of the achievement of certain level of research skills and capabilities. So, following section details the areas where a researcher can enhance his/her skills for conduct of proper research through comprehensive literature review.

Prior understanding of topic and methodology: Before the initiation of research, literature review enhances a researchers' skill on selection of methodology for a particular topic. For

instance, if a research requires survey method of data collection then proper literature review regarding survey method would help the researcher to understand this methodology. This involves:

- critical appraisal of key works that advocate positive approach towards selected approach to research and
- identifying core authors and studies as examples to justify your research.

The literature will provide evidence and substance for justification of choice of a particular methodology.

Persistence: As research is an iterative process wherein a research problem can be solved by applying different research approaches, so a researcher should have the capability to be diligent in the pursuit of conducting research. Preparation of literature review helps in building such capability. Reviewing literature in concern with the research problem requires searching, studying and analyzing literature from different disciplines which requires lot of time and effort. Initial search strategies may not be sufficient thus, making a researcher to search more widely and use other sources. A thorough analysis of literature review also instills the skill of making and managing detailed records which is very helpful in doing doctoral thesis.

Justification: Why a particular topic of research is being selected? Is the selected topic worthy of research? These questions require arguments supported by existing literature. The arguments presented in favor or against regarding a particular research context should not include personal opinions or views. These should include authentic literature from reliable sources which recommends or opposes researchers' arguments.

Making prior assumptions: A careful and methodical review of existing theories and concepts helps the researcher to avoid making pre-judged assumptions about a concept. A researcher can make an important contribution to existing literature by having good knowledge of the subject. That knowledge can only be obtained through work and effort of reading and seeking out ways in which general ideas have been developed through theory and application. This requires a painstaking effort in collecting and reading published works without making baseless assumptions.

Demonstrating originality: It has been indicated that a rigorous conduct of literature would help in giving focus to a topic. Through this focusing process and attention to detail skill a researcher can contribute something new and original to existing concepts and models. All research is unique in its own way and aim of academic research is not to replicate what has already been done but to add in some way something that helps in further understanding. Literature review plays an important role in helping researchers to adopt the capability of being original. It provides research gaps, what has not been done or done in a controversial manner. Thus, a researcher always looks for such gaps and try to fill them by his/her own work.

2.6 Self check questions

a. What is the purpose of conducting literature review?

2.7 STRUCTURE OF LITERATURE REVIEW

A properly structured literature review should contain at least three basic elements: an introduction or background information section; the body of the review containing the discussion of sources; and, finally, a conclusion and/or recommendations section to end the paper.

Introduction: gives a quick idea of the topic of the literature review, such as the central theme or organizational pattern.

Body: contains your discussion of sources and is organized chronologically, thematically, or methodologically.

Conclusions/Recommendations: contains discussion about what has been drawn from reviewing literature so far.

Organizing the body

This section discusses three ways of organizing the sources and content of literature review.

1. *Chronological:* this method involves writing about the literature collected according to when they were published. For instance, if a researcher is conducting study about impact of per capita income on literacy rate then the literature collected should concern with the growth in per capita income over a selected time period, changes in literacy over that time period and relationship between two variables. The literature collected should be arranged according to the date of publication for the variables under consideration. A better method would be to arrange the literature in three different sections corresponding to the three above indicated concepts and then structure literature to each concept chronologically.
2. *Thematic:* Thematic reviews of literature are organized around a topic or issue, rather than the progression of time. For instance in above example, the literature regarding per capita income might also include literature of related topics such as family background, geographic dispersion or family size. In thematic method the studies are still organized chronologically but material regarding concept development is emphasized more than its time progression.
3. *Methodological:* A methodological approach differs from the two above in that the focusing factor usually does not have to do with the content of the material. Instead, it focuses on the "methods" of the researcher or writer. This method influences either the types of documents in the review or the way in which these documents are discussed. For instance, in the per capita income example if methodological method is used then literature might be organized by methods of calculating per capita income or geographic variation in per capita income etc.

2.8 SOURCES OF LITERATURE

One of the quickest and cheapest ways to discover hypotheses is in the work of others, through a literature search. The search may involve conceptual literature, trade literature, or, quite often, published statistics.

- Exploratory insights into this problem could easily and cheaply be gained by analyzing published data and trade literature. Such an analysis would quickly indicate whether the problem was an industry problem or a firm problem. Without the analysis of secondary data as a guide, there is a great danger of researching the wrong "why."
- It is important to remember that in a literature search, the major emphasis is on the discovery of ideas and tentative explanations of the phenomenon. Thus the analyst must be alert to the hypotheses that can be derived from available material, both published and the company's internal records.
- The secondary data can also gathered from the knowledge and experience of those familiar with the general subject being investigated. Thus the respondents must be chosen because of the likelihood that they will offer the contributions sought. It is a waste of time to interview those who have little competence or little relevant experience.
- An intensive study of selected cases of the phenomenon under investigation can also be a key part of literature review. This method is particularly useful in situations which reflect sudden changes or the order in which events have occurred over a period of time.
- Newspapers and Periodicals on a number of important current socio-economic problems can be obtained from the numerical data collected and published by some reputed magazines, periodicals, and newspapers like eastern economist, economic times, the financial express, Indian journal of economics, commerce, capital, transport, statesman's yearbook and the times of India year book etc.
- International Publications of a number of foreign governments and international agencies provide valuable statistical information on a variety of important economic and current topics.

2.9 SUMMARY

A research design is the blueprint for a study that guides the collection and analysis of data. Just as different blueprints reflect differing degrees of detail, research designs vary in their specificity. Some are very detailed and involve the investigation of specific "if-then" relationships, while others simply provide a picture of the overall situation. Exploratory research is basically "general picture" research. It is quite useful in becoming familiar with a phenomenon, in clarifying concepts, in developing but not testing "if-then" statements, and in establishing priorities for further research. Descriptive studies are anything but flexible. Rather, they are rigid in requiring a precise specification of who, what, when, where, why, and how of the research.

When confronted by a new problem, the researcher's first attempts at data collection should logically focus on secondary data. Secondary data are statistics gathered for some other purpose, in contrast to primary data, which are collected for the purpose at hand. Such data which guides a researcher in various steps of research and also indicate the research gaps in previous studies form part of literature review. A literature review can be just a simple summary of the sources, but it usually has an organizational pattern and combines both summary and synthesis.

2.10 GLOSSARY

- **Research design:** is a blueprint of research.
- **Descriptive research:** is inflexible, rigid and quantitative in nature used to generally find frequency of an event
- **Exploratory research:** is flexible and qualitative in nature used to generate tentative hypothesis about the phenomenon under study.
- **Causal Research:** shows that one variable is the cause or effect of the other variable.
- **Literature review:** A literature review discusses published information in a particular subject area, and sometimes information in a particular subject area within a certain time period.

2.11 SHORT ANSWER QUESTIONS

1. What is Literature review? What is the importance of good literature review?
2. What is a research design? Is a research design necessary to conduct a study? Explain its importance.
3. List the various sources of literature review.

2.12 LONG ANSWER QUESTIONS

1. What is a research design? Is a research design necessary to conduct a study?
2. What are the different types of research designs? What is the basic purpose of each?
3. How does a meticulous literature review helps researcher in building certain skills helpful in doing research?
4. Explain the structure of literature review

2.13 ANSWER TO SELF CHECK QUESTIONS

2.1 a. framework or plan for a study used as a guide in collecting and analyzing data.
b. helps the researcher in making one of the most significant decisions about selection of proper research approach.

2.2 a. choice of measurement methods, structure of questionnaire, sampling method to be applied to choose a sample.

2.3 a. Descriptive research
b. Exploratory research
c. Causal Research

2.6 a. enhances a researchers knowledge & selection of methodology for a particular topic

2.14 SUGGESTED READINGS

- Zikmund, William G.; Business Research Methods, Thomson – South Western, Bangalore, 2006, 5th Indian Reprint.

- Cooper, Donald R. and Schindler, Pamela S.; Business Research Methods, Tata McGraw Hill, New Delhi, 2007, 9th Edition.

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS & RESEARCH METHODOLGY: BSRM 201

Lesson No. 3 **Author: PARMOD K AGGARWAL**

SAMPLING DESIGNS

STRUCTURE

3.0 OBJECTIVES

After reading this chapter, the reader should be able to-

- Understand the importance and need for Sampling.
- Differentiate between Census and Sampling methods.
- Identify the different types of Sampling Methods.

3.1 INTRODUCTION

Sampling is defined as the process of learning about population on the basis of sample drawn from it. In it, a part of universe is studied who represent the whole population because it includes all the characteristics of whole universe. There are various probability and non probability sampling methods which are used to collect the data from sample of items selected from population and conclusions are drawn from them. These are known as sampling techniques. For example, if some-one wants to purchase a carton of apples, he will examine only one or two from the

whole lot and on that basis he will examine only one or two from the whole lot and on that basis he will decide whether to purchase the carton or not. We use sampling because of following reasons.

1. **Economy or Reduced Cost**: The sampling method is economical. In the sampling techniques, there is less cost of data collection, administration, transport, training and man hours spent. Collecting data from 2000 or 3000 farmers costs less as compared to 10 lakh farmers covering the entire universe.
2. **Large Scope**: For collecting information highly technically trained personnel with scientific equipment are required. That is why we can say **that it has larger** scope.
3. **Scientific Approach:** The sampling technique is scientific in approach as it is based Oil random sampling. This technique is based on the theory of probability and on certain laws, (a) law of statistical Regularity (b) law of Inertia of large numbers (c) law of Persistence (d) law of Optimization (e) law of validity. Sampling can also ascertain the extent of sampling error and degree of reliability and thus this technique has scientific approach.
4. **Greater Accuracy :** The sampling often permits a higher level of accuracy due to following reasons (a) detailed information can be obtained from a small group (b) Qualified person can be appointed and trained (c) relatively less data can be handled easily.
5. **Detailed Enquiry:** In sampling, the number of units and the area of study are small. Therefore, it is possible to take a detailed and intensive study as is done in case of social, economic and business studies.
6. **Reliability:** In case of sampling results are more reliable because in this it is possible 10 determine the extent of sampling errors and the degree of reliability on the basis of probability.
7. **Less Time**: In sampling only representative units are approached and thus it saves time and man hours. Moreover the processing, editing and analyzing the data also consumes less time.
8. **Administrative Convenience**: Sampling requires small administrative set up involving less personnel including trained investigators. They can be conveniently managed and thus it leads to administrative convenience.
9. **Indispensability**: The sampling technique of collecting information is indispensable in type of universes. Infinite universe, hypothetical universe and universe liable to be destroyed through testing are the examples where any other method cannot be applied so we left with choice of only using sampling techniques.

3.1 Self check questions

a) Explain the concept of a universe or population in statistical terms.

b) List and briefly explain three laws on which the scientific approach of sampling is based.

3.2 BASIC CONCEPTS

1. Universe or Population

In statistics, universe or population means an aggregate of Items about which we obtain information. A universe or population means the entire field under Investigation about which knowledge is sought. For example, if we want to collect information about the average monthly expenditure of all the 3,000 students of a college, then the entire aggregate of 3,000 students will be termed as Universe or Population. A population can be of two kinds (i) Finite and (ii) Infinite. In a finite population, number of items is definite such as, number of students or teachers in a college. On the other hand, an infinite population has infinite number of items eg. number of stars in the sky, number of water drops in an ocean, number of leaves on a tree or number of hairs on the head.

2. Sample

A part of population is called sample. In other words, selected or sorted units from the population are known as a sample. In fact, a sample is that part of the population which we select for the purpose of Investigation. For example, if an investigator selects 200 students from 2000 students of a college who represent all of them, then these 200 students will be termed as a sample. Thus, sample means some units selected out of it population which represent it.

3.2 Self check questions

a) Why sampling technique is preferred over census method?

3.3 CENSUS AND SAMPLE METHODS

There are two methods to collect statistical data:

(1) Census Method

(2) Sample Method

3.3.1 CENSUS METHOD

Census method is that method in which information or data is collected from each and every unit of the population relating to the problem under investigation and conclusions are drawn on their basis. This method is also called as Complete Enumeration Method. For example, suppose some information (like Monthly Expenditure. Average Height, Average Weight etc) is to be collected regarding 3000 students of a college, For that purpose if we collect data by inquiring each and every student of the college then this method will be called as Census method. In this example, the whole college i.e. all 3000 students will be considered as a population and every student as an individual will be called the unit of the population. Population census in India is conducted after every ten years by using census method.

Merits

(i) **Reliable and Accurate Data:** Data obtained by census method have more

reliability and accuracy because in this method data are collected by contacting each and every unit of the universe.

(ii) **Extensive Information:** This method gives detailed information about each unit of the universe. For example, Indian population census does not only provide the knowledge about the number of persons but also information about their age, occupation, income, education, marital status etc.

(iii) **Suitability:** This method is more suitable for the population with limited scope and diverse characteristic. Use of this method is also appropriate where intensive study is desired.

Demerits

(i) **More Expensive:** Census method is an expensive one. More money is needed for it as information is collected from each unit of the population. This is why this method is used by Government mostly for very important issues like Census etc.

(ii) **More Time:** This method involves much time for data collection because data are collected from each and every unit of the population. This results in delay in making statistical inferences.

(iii) **More Labour:** This method of data collection also involves very much labour. For this the enumerators in a large number are required.

(iv) **Not Suitable for Specific Problems**: This method is not suitable relating to certain specific problems and infinite population. For example if the population is infinite or items of the population are perishable or very complex type, then the census method is not suitable.

3.3.2 SAMPLING METHOD

Sampling method is that method in which data is collected from the sample of items selected from population and conclusions are drawn from them. For example if a study is to be made regarding the monthly expenditure of 3000 students of a college, then instead of collecting information from each student of the college, if we collect information by selecting some students like 100, then this will be called Sampling method. On the basis of sampling method, it is possible to study the monthly expenditure of all the students of the college. Sampling method has three main stages (i) to select a sample (ii) to collect information from it and (iii) to make inference regarding the population.

3.3.3 IMPORTANCE OF SAMPLING METHOD

In modern times sampling method is an important and popular method of statistical inquiry. Besides economic and business world, this method is widely used in daily life. For example, a housewife comes to know of the cooking of the whole lot of rice by observing two- three grains only A doctor tests the blood of a patient by examining one or two drops of blood only. In the same way, we learn about the quality of a commodity while buying the items of daily use like wheat, rice, pulses,

etc. by observing the sample or specimen. In factories, statistical quality controller inspects the quality of items by examining a few units produced teacher gets the knowledge about the efficacy of his teaching by putting questions to a few students. In reality, there is scarcely any area left where sampling method is not used.

Merits

(i) **Saving of Time and Money**: Sampling method is less expensive. It saves money and labour because only a few units of the population are studied.

(ii) **Saving of Time**: In sampling method, data can be collected more quickly as these are obtained from some items, of the universe. Thus much time is saved.

(iii) **Intensive Study**: As number of items is less in sampling method, they can be intensively studied.

(iv) **Organizational Convenience**: In this method, research work can be organized and executed more conveniently. More skilled and competent investigators can be appointed.

(v) **More Reliable Results**: If sample is selected in such a manner as it represents totally the universe, then the results derived from it will be more accurate and reliable.

(vi) **More Scientific**: Sampling method is more scientific because data can be inquired with other samples.

(vii) **Only Method:** In some fields where inquiry by census method is impossible, then in such situation, sampling method alone is more appropriate. If the population is infinite or too widespread or of perishable nature, then sampling method is used in such cases.

Demerits

(i) **Less Accurate:** Sampling method has less accuracy because rather than making inquiry about each unit of the universe, partial inquiry or inquiry relating to some selected units only is made.

(ii) **Wrong Conclusions:** If method of selecting a sample is not unbiased or proper caution has not been taken, then results are definitely misleading.

(iii) **Less Reliable :** Compared to census method, there is more likelihood of the bias of the investigator, which makes the results less reliable

(iv) **Need of Specified Knowledge:** This is a complex method as specialized knowledge is required to select a sample.

(v) **Not Suitable:** If all units of a population are different from one another, then sampling method will not prove to be much useful.

3.3.4 DIFFERENCE BETWEEN CENSUS AND SAMPLE METHOD

The main difference between the census method and the sampling method arc as follows:

(i) **Scope:** In census method, all items relating to a universe axe investigated whereas in sampling method only a few items are inquired.

(ii) **Cost:** Census method is expensive from the point of view of time, money and labour whereas sampling method economizes on them.

(iii) **Field of Investigation :** Census method is used in investigations with limited field whereas sampling method is used for investigations with large field

(iv) **Homogeneity:** Census method is useful where units of the population are heterogeneous whereas sampling method proves more useful where population units are homogenous.

(v) **Type of Universe:** In such fields where study of each and every unit of the universe is necessary, census method is more appropriate. On the contrary, when population is infinite or vast or liable to be destroyed as a result of complete enumeration, then sampling method is considered to be more appropriate.

3.3 Self check questions

a) What are the demerits of using census technique?

b) What are the benefits of using sample technique instead of census?

c) Which technique is useful where units are homogeneous?

3.4 SAMPLING METHODS

The method of selecting a sample out of a given population is called sampling. In other words, sampling denotes the selection of a part of the aggregate statistical material with a view to obtaining information about the whole. Nowadays, there are various methods of selecting a sample from a population in accordance with various needs.

1. Probability Sampling Methods
   (1) Simple Random Sampling
   (2) Stratified Random Sampling
   (3) Systematic Random Sampling
   (4) Multistage Random Sampling
   (5) Cluster Sampling
2. Non-Probability Sampling Methods:
   (1) Judgement Sampling
   (2) Quota Sampling
   (3) Convenience Sampling
   (4) Extensive Sampling

3.4.1 PROBABILITY SAMPLING METHODS

Probability sampling methods are such methods of selecting a sample from the population in which all units of the universe are given equal chances of being included in the sample. There are various variants of probability sampling methods, which are given below:

1. Simple Random Sampling

Simple random sampling is that method in which each item of the universe has an equal chance of being selected in the sample. Which item will be included in the sample and which not, such decision is not made by an investigator on his will but selection of the units is left on chance. According to random sampling, there are two methods of selecting a random sample:

(i) **Lottery Method**: In this method, each unit of the population is named or numbered which is marked on separate piece of paper. Such chits are folded and put into some urn or bag. Thereafter as many chits are made selected by some person as many units are to be included in a sample,

(ii) **Tables of Random Numbers**: Some experts have constructed random number tables. These tables help in selection of a sample. Of all such various tables, Tippett's Tables are most famous and are in use. Tippett has constructed a four digit table of 10,400 numbers by using numbers as many as 41,600. In this method, first of all, all the items of a population are written serially. There after by making use of Tippett's tables, in accordance with the size of the sample, numbers are selected. The selection of a sample with the help of Tippett's table can be made clear by an example :

An Extract of Tippett's Table

2952	6641	3992	9792	7979	5911
3170	5524	4167	9525	1545	1396
7203	4356	1300	2693	2370	7483
3408	2762	3563	6107	6913	7691
0560	5246	1112	9025	6008	8127

For example, 12 units are to be chosen out of 5000 units. With Tippett's table, to decide such units, firstly 5000 units will be serially ordered from 1 to 5000 and then from Tippett's table, 12 numbers will be chosen from the beginning which is less than 5000. These 12 numbers are follows:

2952	3992	3170
4356	1300	2693
2370	3408	2762
4167	1545	1396

The items of such serial numbers will be included in the sample. If units of the population are less than 100, then 4 digit random numbers will be made compact into two digit numbers, and then such two digit numbers will be selected. Like as to select 6 units out of 60 units, the units with serial numbers 29, 39, 31, 41, 15 and 13 will be included in the sample.

Merits

(i) There is no possibility of personal prejudice in this method. In other words, this method is free from personal bias.

(ii) Under this method, every unit of the universe gets equal chance of being selected.

(iii) The use of this method saves time, money and labour.

Demerits

(i) If sample size is small, then sample is not adequately represented.
(ii) If universe is very small, then this method is not suitable.
(iii) If some items of the universe are so important that their inclusion in the sample is very essential, then this method will not be appropriate.
(iv) This method will not be appropriate when population has units with diverse characteristics.

2. Stratified Random Sampling

This method is used when units of the universe are heterogeneous rather than homogenous. Under this method, first of all, units of the population are divided into different strata in accordance with their characteristics. Thereafter by using random sampling, sample items are selected from each stratum. For example, if 150 students are to be selected out of 1500 students of a college, then firstly the college students will be divided into three groups on the basis of Arts, Commerce and Science. Suppose there are 500, 700, 300 students respectively in three faculties and 10% sample is to be taken, then on the basis of random sampling 50, 70, and 30 students respectively will be selected by using random sampling. Thus this method assumes equal representation to each class or group and all the units of the universe get equal chance of being selected in the sample.

Merits

(i) There is more likelihood of representation of units in this method
(ii) Comparative study on the basis of facts at different strata is possible under this method.
(iii) This method has more accuracy

Demerits

(i) This method has limited scope because this method can be adopted only when the population and its different strata are known.
(ii) There can be the possibility of prejudice if the population is not properly stratified.
(iii) If the population is too small in size, it is difficult to stratify it.

3. Systematic Random Sampling

In this method, all the items of the universe are systematically arranged and numbered and then sample units are selected at equal intervals. For example, if 5 out of 50 students are to be selected for a sample, then 50 students would be numbered and systematically arranged. One item of the first 10 would be selected at random. Subsequently, every 10th item from the selected number will be selected to frame a sample. If the first selected number is 5th item, then the subsequent numbers would be 15th, 25th, 35th and 45th.

Merits

(i) It is a simple method. Samples can be easily obtained by it.

(ii) This method involves very little time in sample selection and results are almost accurate.

Demerits

(i) In this method each unit does not stand the equal chances of being selected because only the first unit is selected on random sampling basis.

(ii) If all the units are different in characteristics, then results will not be appropriate.

4. Multistage Random Sampling

When sampling procedure passes through many stages, then it is known as multi-stage random sampling. In this method, firstly the entire universe or population is divided into stages or sub stages. From the each stage some units are selected on random sampling basis. Thereafter, these units are subdivided and on the basis of random sampling again some subunits are selected. Thus, this goes on with sub-division further and selection on. For example, for the purpose of a study regarding Adult Education in Punjab State, first some districts will be selected on random basis. Thereafter out of the selected districts, some tehsils and out of tehsils, some villages or towns may be thus selected, further out of the villages or towns, some neighbourhood, or wards and out of the wards, some households will be selected from whom the inquiry will be made concerning the problem at hand.

Merits

(i) This is the best method of studying a universe or population on regional basis.

(ii) This method is suitable for those problems where decisions on the basis of sample alone cannot be taken.

Demerits

(i) This method requires many tests to correctly estimate the level of accuracy which involves a lot of time and labour.

(ii) In this method, level of estimated accuracy level is pre-decided which does not seem logical.

5. Cluster Sampling

In this method, simply the universe is divided into many groups called cluster and out of which a few clusters are selected on random basis and then the clusters are completely enumerated. This method is usually applied in industries like as in pharmaceutical industry, a machine produces medicines/tablets in the batches of hundred each, then for quality inspection, a few randomly selected batches are examined.

3.4.2 NON-PROBABILITY SAMPLING METHODS

Non-probability sampling methods are those methods in which selection of units is made on the basis of convenience or judgement of the investigator rather than on the basis of probability or chance. In such methods, selection of units is made in accordance with the specific objectives and convenience of the investigator.

1. Judgement Sampling

Under this method, the selection of the sample items depends exclusively on the judgement of the investigator. In other words, the investigator exercises his judgement in the choice and includes those items in the sample which he thinks are most typical of the universe with regard to the characteristics under study. For example, if a sample of 20 students is to be selected from a class of 80 students for analyzing the spending habits of the 80 students, the investigator would select 20 students, who in his opinion are representative of the class.

Merits

(i) This method is less expensive.
(ii) This method is very simple and easy.
(iii) This method is useful in those fields where almost similar units exist or some units- are too important to be left out of the sample.

Demerits

(i) There is greater chance of the investigator's own prejudice in this method.
(ii) This method is not very accurate and reliable.

2. Quota Sampling

In this method, the investigators are assigned definite quotas according to some criteria. They are instructed to obtain the required number to fill in each quota. The investigators select the individuals (i.e. sample items) to collect information on their personal judgements within the quotas. When all or a part of the whole quota is not available or approachable, the quota is completed by supplementing new respondents. Quota sampling is a type of judgement sampling.

Merits

(i) In this method, there is greater chance of important units being included.
(ii) Statistical inquiry is more organized in this method on account of the units of the quotas being fixed.

Demerits

(i) Possibility of prejudice will exist.
(ii) There is greater likelihood of sampling error in this method.

3. Convenience Sampling

In this type of non-probability sampling, the choice of the sample is left completely to the convenience of the investigator. The investigator obtains a sample according to his convenience. For example, a book publisher selects some teachers conveniently on the basis of the list of the teachers from the college prospectus and gets feedback from them regarding his publication. This method is less expensive and simple but is unscientific and unreliable. This method results in more dependence on the enumerators. This method is appropriate for sample selection where the universe or population is not clearly defined or list of the units is not available or sample units are not clear in themselves.

4. Extensive Sampling

In this method, sample size is taken almost as big as the population itself like 90% the section of the population. Only those units are left out for which data collection is very difficult or almost impossible. Due to very large sample size, the method has greater level of accuracy. Intensive study of the problem becomes possible but this method involves heavy resources at disposal.

3.4 Self check questions

a) What are the methods of selecting a random sample?
b) In which sampling technique, investigators are instructed to obtain required number to fill the quota
c) Stratified sampling is done in case of ____
d) In which sampling, every nth item will be selected to frame a sample
e) In which technique, sample size is taken almost as big as population?

3.5 SAMPLING AND NON-SAMPLING ERRORS

The choice of a sample though may be made with utmost care, involves certain errors which, may be classified into two types: (1) Sampling Errors, and (2) Non-Sampling Errors. These errors may occur in the collection, processing and analysis of data.

1. Sampling Errors

Sampling errors are those which arise due to the method of sampling. Sampling errors arise primarily due to the following reasons:

(1) Faulty selection of the sampling method.
(2) Faulty demarcation of sampling units.
(3) Variability of the population which has different characteristics.

2. Non-Sampling Errors

Non-sampling errors are those which creep in due to human factors which always vary from one investigator to another. These errors arise due to any of the following factors

(1) Faulty planning.
(2) Faulty selection of the sample units.
(3) Lack of trained and experienced staff which collect the data.
(4) Negligence and non-response on the part of the respondent.
(5) Errors in compilation.
(6) Errors due to wrong statistical measures.
(7) Framing of a wrong questionnaire.
(8) Incomplete investigation of the sample survey.

3.5 Self check questions

a) Explain variability in population?
b) Which errors arise due to human factor?

3.6 SUMMARY

There are several research designs and the researcher must decide in advance of collection and analysis of data as to which design would prove to be more appropriate for his research project. He must give due weight to various points such as the type of universe and its nature, the objective of his study, the resource list, desired standard of accuracy etc while deciding on a particular research design.

3.7 GLOSSARY

- **Research Design**: The arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure.
- **Variable**: An entity which can be measured, or which can take on different quantitative values.
- **Treatments**: The different conditions under which experimental and control groups are put are usually referred to as treatments.
- **Experiments**: The process of examining the truth of a statistical hypothesis reacting to some research problem is known as an experiment.
- **Error**: A deviation from accuracy or correctness.
- **Sampling**: Items selected at random from a population and used to test hypotheses about the population.

3.8 SHORT ANSWER QUESTIONS

1. Explain the difference between stratified and cluster sampling?
2. Explain the concept of sample and census.
3. What is the importance of sampling.

3.9 LONG ANSWER QUESTIONS

1. Define sample. What are requirements of Good Sample?
2. Discuss the methods of probability sampling.
3. Explain the methods of non-probability sampling.
4. Distinguish between random sampling and non random sampling.
5. Explain various types of sampling and non sampling errors.

3.10 ANSWERS OF SELF CHECK QUESTIONS

3.1 a) A universe terms is an aggregate of items about which information is sought.

b) Three laws are : Law of Statistical Regularity, Law of Inertia of Large Numbers, Law of Persistence

3.2 a) economy or reduced costs

3.3 a) more expensive and time taking

b) Saving of time and money for large populations

c) Sampling

3.4 a) Lottery method and tables of random numbers

b) quota sampling

c) heterogenous

d) systematic random sampling

e) extensive sampling

3.5 a) population have different characterstics

b) non sampling error

3.10 SUGGESTED READINGS

- Statistical Methods by S.P. Gupta.
- Mathematical Statistics by S.C. Gupta.
- Statistics by Kapoor and Saxena.
- Statistical Analysis by T.L. Kausal.

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS & RESEARCH METHODOLGY: BSRM 201

Lesson No. 4 **AUTHOR: SHILPI GOYAL**

SCALING TECHNIQUES

STRUCTURE

4.0 OBJECTIVES

After reading this chapter, the reader should be able to :

- Understand the importance of scaling techniques in research and measurement.
- Outline the major scale construction techniques and understand the commonly used scaling techniques.

4.1 INTRODUCTION

In research we quite often face measurement problem, especially when the concepts to be measured are complex and abstract and we do not possess the standardized measurement tools. Alternatively, we can say that while measuring attitudes and opinions, we face the problem of their valid measurement. Similar problem may be faced by a researcher while measuring physical or institutional concepts, though in a lesser degree. Thus there was the need to develop the techniques which may enable us to measure abstract concepts more accurately. This brought in the study of scaling techniques.

4.2 MEANING OP SCALING

Scaling can be defined as "procedure for the assignment of numbers (or other symbols) to a property of objects in order to impart some of the characteristics of numbers to the properties in question."

Scaling describes the procedures of assigning numbers to various degrees of opinion, attitude and other concepts. This can be done in two ways :

i. Making a judgement about some characteristic of an individual and then placing him directly on a scale that has been defined in terms of that characteristic and
ii. Constructing questionnaires in such a way that the score of individual's responses assigns him a place on a scale.

The scale is a continuum, consisting of the highest point, and the lowest point along with several intermediate points between these two extreme points. These scale point positions are so related to each other that when the first point happens to be the highest point, the second point indicates a higher degree in terms of a given characteristic as compared to the third point and the third point indicates a higher degree as compared to the fourth and so on. Numbers for measuring the distinctions of degree in the attitudes/ opinions are, thus, assigned to individuals corresponding to their scale positions. Hence the term scaling is applied to the procedures for attempting to determine quantitative measures of subjective abstract concepts.

4.3 PRIMARY SCALES OF MEASUREMENT

There are four primary scales of measurement: nominal, ordinal, interval and ratio.

4.3.1 Nominal Scale

A nominal scale is a figurative labeling scheme in which the numbers serve only as labels or tags for identifying and classifying objects. When a nominal scale is used for the purpose of identification, there is strict one-to-one correspondence between the number and the objects. Each number is assigned to only one object and each object has only one number assigned to it. For example, the numbers assigned to respondents in a study constitute a nominal scale. When used for classification purposes, the nominally scaled numbers serve as labels for classes or categories. All objects in the same class have the same number and no two classes have the same number. The numbers in a nominal scale do not reflect the amount of characteristic possessed by the objects. The only permissible operation on the numbers in a nominal scale is counting. Only a limited number of statistics, which are based on frequency counts, such as, percentages, mode, chi-square, and binomial tests etc. are permissible.

4.3.2 Ordinal Scale

An ordinal scale is a ranking scale in which numbers are assigned to objects to indicate the relative extent to which the objects possess some characteristic. An ordinal scale allows a researcher to determine whether an object has more or less of a characteristic than some other object, but not how much more or less. Thus an ordinal scale indicates relative position, not the magnitude of the difference between the objects. Common examples of ordinal scales include quality rankings, ranking of teams in a tournament, socioeconomic class, and occupational status etc. in addition to the counting operation allowable for nominal scale data, ordinal scale permits the use of statistics such as median, percentile, quartile, rank-order correlation etc.

4.3.3 Interval Scale

In an interval scale, numerically equal distances on the scale represent equal values in the characteristic being measured. An interval scale contains all the information of an ordinal scale but it also allows a researcher to compare the differences between objects. The difference between any two scale values is identical to the difference between any other two adjacent values of an interval scale. There is a constant or equal interval between scale values. The difference between 1 and 2 is the same as the difference between 2 and 3, which is the same as the difference between 5 and 6. A common example in everyday life is a temperature scale.

In an interval scale, the location of the zero point is not fixed. Both the zero point and the units of measurement are arbitrary. The statistical techniques that may be used on interval scale data include all of those that can be applied to nominal and ordinal data in addition to the arithmetic mean, standard deviation, product-moment correlations etc.

4.3.4 Ratio Scale

A ratio scale possesses all the properties of the nominal, ordinal, and interval scales,

and, in addition, an absolute zero point. Thus, a ratio scale allows a researcher to identify or classify objects, rank order the objects, and compare intervals or differences. It is also meaningful to compute ratios of scale values. Common examples of ratio scales include height, weight, age etc. all statistical techniques can be applied to ratio data. These include specialized statistics such as geometric mean, harmonic mean, and coefficient of variation.

4.3 Self check questions

a. Give some examples of ratio scales.
b. Which scale possesses all the properties of the nominal, ordinal, and interval scales, and, in addition, an absolute zero point
c. Which scale serves only as labels or tags for identifying and classifying objects

4.4 SCALING TECHNIQUES

4.4.1 Rating Scales

These scales are also called categorical scales. The rating scales involve qualitative description of a limited number of aspects of a thing or of traits of a person. When rating scales are used, an object can be judged in absolute terms against some specified criteria, i.e. the properties of the objects are judged without reference to other similar objects. These ratings may be in such forms as "like-dislike", "above average, average, below average", or other classifications with more categories such as "like very much-like somewhat-neutral- dislike somewhat-dislike very much" and so on. In practice, three to seven point scales are generally used because more points on a scale increase the sensitivity for measurement.

Rating scale may be either a graphic rating scale or an itemized rating scale.

i. **The Graphic Rating Scale** is quite simple and is commonly used in practice. In this, the various points are usually put along the line to form a continuum and the rater indicates his rating by simply making a mark at the appropriate point on a line that runs from one extreme to the other. Scale points with brief descriptions may be indicated along the line, their function being to assist the rater in performing his job. This type of scale has several limitations. The respondents may check at almost any position along the line which may make the analysis difficult. The meanings of the terms like "very much" and "somewhat" may depend upon respondent's frame of reference,
ii. **The Itemized Rating Scale**, also known as numerical scale, presents a series of statements from which a respondent selects one as best reflecting his evaluation. These statements are ordered progressively in terms of more or less of some property. The chief merit of this type of scale is that it provides

more information and meaning to the rater, and thereby increases reliability. But this form is relatively difficult to develop and the statements may not say exactly what the respondent would like to express.

4.4.2 Ranking Scales

These scales are also called comparative scales. Under ranking scales, relative judgements are made against other similar objects. The respondents under this method directly compare two or more objects and make choices among them. Here are two generally used approaches of ranking scales:

i. **Method of paired comparisons:** under it, the respondent can express his attitude by making a choice between two objects.

ii. **Method of Rank Order**: under this method of comparative scaling, the respondents are asked to rank their choices. This method is easier and faster than the method of paired comparison. For example, with 10 items, it takes 45 pair comparisons to complete the task, whereas the method of rank order simply requires ranking of 10 items.

4.4 Self check questions

a. In which scale, various points are usually put along the line to form and the rater indicates his rating by making a mark at a point on a line that runs from one extreme to the other.

b. Ranking scales are also called _____

4.5 SCALE CONSTRUCTION TECHNIQUES

The main techniques by which scales can be developed are:

i. **Arbitrary Approach**: It is an approach where the scale is developed on ad hoc basis. This is the most widely used approach. It is presumed that such scales measure the concepts for which they have been designed, although there is little evidence to support such an assumption.

ii. **Consensus Approach**: Here a panel of judges evaluates the items chosen for inclusion in the instrument in terms of whether they are relevant to the topic area and unambiguous in implication.

iii. **Item analysis Approach**: Under this, a number of individual items are developed into a test which is given to a number of respondents. After administering the test, total scores are calculated for everyone. Individual items are then analyzed to determine which items discriminate between persons or objects with high total scores and those with low scores.

iv. **Cumulative Scales**: They are chosen on the basis of their conforming to

some ranking of items with ascending and descending discriminating power.

v. **Factor Scales**: They may be constructed on the basis of inter correlations of items which indicate that a common factor accounts for the relationship between items. This relationship is typically measured through factor analysis method.

4.6 ARBITRARY SCALES

Arbitrary scales are developed on ad hoc basis and are designed largely through the researcher's own subjective selection of items. The researcher first collects a few statements or items which he believes are unambiguous and appropriate to a given topic. Some of these are selected for inclusion in the meaningful instrument and then people are asked to check in a list the statements with which they agree.

The chief merit of these scales is that they can be developed very easily, quickly and with relatively less expense. They can also be designed to be highly specific and adequate. At the same time, there are some limitations. We do not have any objective evidence that such scales measure the concepts for which they have been developed.

4.7 DIFFERENTIAL SCALES (THURSTONE -TYPE SCALES)

The name of L.L.Thurstone is associated with differential scales which have been developed using consensus scale approach. Under such an approach the selection of items is made by a panel of judges who evaluate the items in terms of whether they are relevant to the topic area and unambiguous in implication. The detailed procedure is described as under:

(a) The researcher gathers a large number of statements, usually twenty or more that express various points of view toward a group, institution, idea, or practice.

(b) These statements are then submitted to a panel of judges, each of whom arranges them in eleven groups or piles ranging from one extreme to another in position. Each of the judges is requested to place those statements in the first pile that he considers most unfavourable to the issue, in the second pile, those statements that are next most unfavourable and so on till in the eleventh pile, he puts the statements that he considers to be most favourable.

(c) This sorting by each judge yields a composite position for each of the items. In case of marked disagreement between the judges in assigning a position to an item, that item is discarded.

(d) For items that are retained, each is given its median scale value between one and eleven as established by the panel.

(e) A final selection of statements is then made. For this purpose a sample of statements whose median scores are spread evenly from one extreme to the other is taken. The statements so selected, constitute the final scale to be

administered to the respondents. The position of each statement on the scale is the same as determined by the judges.

After developing the scale, the respondents are asked to check the statements with which they agree. Such scales are considered most appropriate and reliable when used for measuring a single attitude. But an important deterrent to their use is the cost and effort required to develop them. Another weakness of such scales is that the values assigned to various statements by the judges may reflect their own attitudes. The method is not completely objective; it involves ultimately subjective decision process.

4.8 SUMMATED SCALES (OR LIKERT -TYPE SCALES)

Summated scales (or Likert-type scales) are developed by utilizing the item analysis approach wherein a particular item is evaluated on the basis of how well it discriminates between those persons whose total score is high and those whose score is low. Those items or statements that best meet this sort of discrimination test are included in the final instrument.

Thus, summated scales consist of a number of statements which express either a favourable or unfavourable attitude towards the given object to which the respondent is asked to react. The respondent indicates his agreement or disagreement with each statement in the instrument. Each response is given a numerical score, indicating its favourableness or unfavourableness, and the scores are totaled to measure the respondent's attitude.

In a Likert scale, the respondent is asked to respond to each of the statements in terms of several degrees, usually five of agreement or disagreement. For example, when asked to express opinion whether one considers his job quite pleasant, the respondent may respond in any one of the following ways: (i) strongly agree, (ii) agree, (iii) undecided, (iv) disagree, (v) strongly disagree. At one extreme of the scale there is strong agreement with the given statement and at the other, strong disagreement and between them lie intermediate points. Each point on the scale carries a score. Response indicating the least favourable degree of job satisfaction is given the least score (say 1) and the most favourable is given the highest score (say 5).the same thing is done in respect of each and every statement in the instrument. This way the instrument yields a total score for each respondent, which would then measure the respondent's favourableness toward the given point of view. If the instrument consists of, say 30 statements, the following would be the score values:

30 x 5 = 150 Most favourable response possible

30 x 3 = 90 A neutral attitude

30 x 1 = 30 Most unfavourable attitude

The scores for any individual would fall between 30 and 150. If the score happens to be above 90, it shows favourable opinion to the given point of view, a score of below 90 would mean unfavourable opinion and a score of exactly 90 would be suggestive of a

neutral attitude.

Procedure

The procedure for developing a Likert-type scale is as follows:

i. The researcher collects a large number of statements which are relevant to the attitude being studied and each of the statements express definite favourableness or unfavourableness to a particular point of view or the attitude and the number of both types of the statements is approximately equal.
ii. After the statements have been gathered, a trial test is administered to a number of subjects.
iii. The response to various statements are scored in such a way that a response indicative of most favourable attitude is given the highest score of 5 and that with the most unfavourable attitude is given the lowest score, say, of 1.
iv. Then the total score of each respondent is obtained by adding his scores that he received for separate statements.
v. The next step is to array these total scores and find out those statements which have a high discriminatory power. For this purpose, the researcher may select some part of the highest and the lowest total scores say the top 25 percent and the bottom 25 percent. These two extreme groups are interpreted to represent the most favourable and the least favourable attitudes and are used as criterion groups by which to evaluate individual statements.
vi. Only those statements that correlate with the total test are retained in the final instrument and all others are discarded.

Advantages

(a) It is relatively easy to construct.
(b) Likert-type scale is considered more reliable because in this respondents answer each statement included in the instrument.
(c) This scale permits the use of statements that are not manifestly related to the attitude being studied.
(d) It can be easily used in respondent-centered and stimulus-centered studies. It means we can study how responses differ between people and how responses differ between stimuli.
(e) Likert-type scale takes much less time to construct.

Limitations

(a) With this scale, we can simply examine whether respondents are more or less favourable to a topic, but we cannot tell how much more or less they are.
(b) The total score of an individual respondent has little clear meaning since a

given total score can be secured by a variety of answer patterns.

(c) There remains a possibility that people may answer according to what they think they should feel rather than how they do feel.

4.8 Self check questions

a. Likert scales are also called _____

b. With likert scale, one can examine whether respondents are more or less favourable to a topic, but cannot tell how much more or less they are. True/false

4.9 CUMULATIVE SCALES

Cumulative scales or Louis Guttman's scalogram analysis consist of series of statements to which a respondent expresses his agreement or disagreement. The special feature of this type of scale is that the statements in it form a cumulative series. It means that the statements are related to one another in such a way that an individual, who replies favourably to say, item 3, also replies favourably to items 2 and 1, and one who replies favourably to item 4 also replies favourably to items 3, 2, and 1 and so on.
The individual score is worked out by counting the number of points concerning the number of statements he answers favourably. If one knows this total score, one can estimate as to how a respondent has answered individual statements constituting cumulative scales. The major scale of this type of cumulative scales is the Guttmann's scalogram.
The technique developed by Louis Guttmann is known as scalogram analysis or scale analysis. It refers to the procedure for determining whether a set of items forms a uni-dimensional scale. Under this technique, 'the respondents are asked to indicate in respect of each item whether they agree or disagree with it, and if these items form a unidimensional scale, the response pattern will be as under:

Item Number				**Respondent**
4	**3**	**2**	**1**	**Score**
X	X	X	X	4
	X	X	X	3
		X	X	2
			X	**1**
				0
X= Agree				

A score of 4 means that the respondent is in agreement with all the statements which

indicates the most favourable attitude. But a score of 3 would mean that the respondent is not in agreement with item 4, but he agrees to all the other items. In the same way, one can interpret other values of the respondent's scores.

4.10 FACTOR SCALES

Factor scales are developed through factor analysis or on the basis of inter correlations of items which indicate that a common factor accounts for the relationship between items. Factor scales are particularly "useful in uncovering latent attitude dimensions and approach scaling through the concept of multiple-dimension attribute space". More specifically, the two problems viz., how to deal appropriately with the universe of content which is multidimensional and how to uncover underlying dimensions, which have not been identified, are dealt with through factor scales. The important factor scales based on factor analysis are Semantic Differential and Multidimensional Scaling.

4.10.1 Semantic Differential Scale

This scale was developed by Charles E. Osgood, G.J.Suci and P.H.Tannenbaum, is an attempt to measure the psychological meanings of an object to an individual. This scaling consists of a set of bipolar rating scales, usually of 7 points, by which one or more respondents' rate one or more concepts on each scale item. For instance, the S.D. scale items for analyzing candidates for leadership position may be shown as:

(E	Successful	Unsuccessful
(P	Severe	Lenient
(P	Heavy	Light
(E	Progressive	Regressive
(P	Strong	Weak
(A	Active	Passive
(A	Fast	Slow
(E	True	False
(E	Sociable	Unsociable

4.10.2 Multidimensional Scaling

It is a relatively more complicated scaling device, but with this sort of scaling, one can scale objects, individuals or both with a minimum of information.MDS can be characterized as a set of procedures for portraying perceptual or affective dimensions of substantive interest. MDS is used when all the variables in a study are to be analyzed simultaneously and all such variables happen to be independent. The underlying assumption in MDS is that people perceive a set of objects as being more or less similar to one another on a number of dimensions instead of only one. Through MDS technique, one can represent geometrically the locations and interrelationships among a set of points.

4.11Self check questions

a. Which scale consists of a set of bipolar rating scales, usually of 7 points

4.11 SUMMARY

The chapter focused on the importance of scaling in measurement and research. The researchers faced the problem of valid measurement while measuring complex and abstract concepts and while measuring attitudes. This leads to the development of scaling techniques. Various scale construction techniques and the commonly used scaling techniques have been discussed.

4.12 GLOSSARY

- **Scaling** - procedure for the assignment of numbers to a property of objects in order to impart some of the characteristics of numbers to the properties in question
- **Nominal Scale** - a scale whose numbers serve only as labels or tags for identifying and classifying objects with a strict one-to-one correspondence between the numbers and the objects.
- **Ordinal Scale** - a ranking scale in which numbers are assigned to objects to indicate the relative extent to which some characteristic is possessed.
- **Interval Scale** - a scale in which the numbers are used to rate objects such that numerically equal distances on the scale represent equal distances in the characteristics being measured.
- **Ratio Scale** - it allows the researcher to identify, rank order the objects and compare intervals or differences.

4.13 SHORT ANSWER QUESTIONS

1. Define the term scaling. Discuss the various scaling techniques used in the measurement of attitudes.
2. Discuss the Likert Scale and Thurstone Scale with one example each. Which one of the two scales is better and why?
3. What is semantic differential scaling?

4.14 LONG ANSWER QUESTIONS

1. What do you mean by scaling? State some scale construction techniques.

2. What is a likert scale? What is the procedure of development of a likert scale. Give its merits and demerits.

3. Explain ranking and rating scales in detail.

4.15 ANSWERS TO SELF CHECK QUESTIONS

4.3 a. Height, age, weight

b. Ratio scale

c. Nominal scale

4.4 a. Graphic rating scale

b. Comparative scales

4.8 a. summated scales

b. True

4.10 a. semantic differential scale

4.16 SUGGESTED READINGS

- Kothari, C. R., **Research Methodology : Methods and Techniques,** New Age International Publishers, New Delhi, 2nd Edition, 2006.
- Malhotra, N. K., **Marketing Research : An Applied Orientation,** Pearson Education, New Delhi, 1st Edition, 2003.

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS &RESEARCH METHODOLGY: BSRM 201

Lesson No. 5 **AUTHOR: ARUSHI**

TOOLS AND TECHNIQUES OF DATA COLLECTION

5.0 OBJECTIVES

The readers of this chapter should be able to understand

- Types of data and its usage
- The primary and secondary sources of data and some fundamental ways of collecting data

5.1 INTRODUCTION

Statistics is concerned with understanding the real world through the information that we derive from classification and measurement of data. Its distinctive characteristic is that it deals with variability and uncertainty which is everywhere. This gives it its fundamental role and is the reason why its tentacles reach into every corner of the scientific enterprise. Variability and

uncertainty are two sides of the same coin and, together, are the hallmark of a statistical problem. It is this which gives statistics right of entry to almost every sphere of human endeavor.

For instance, if somebody states to a garment manufacturer that 'December was colder than usual this year' then such statement leads to lot of variability and uncertainty. This does not tell the meaning of usual. Was this years' December winter was compare with last year or previous ten years. Because variability in winter data would change with change in time period. Also, can manufacturer reach to a conclusion from such a statement that next year winters would be harsh in December and the company should manufacture more of sweaters? To derive correct information and make a decision accurately manufacturer has to rely on proper and reliable collection of data and apply relevant statistical tools to reach a conclusion. Thus, statistics involves extraction of information from collected data.

Similarly, consider a statement 'the store is facing stiff competition from mom and pop stores'. Was the store a small sized store or a hyper mart? What was its location: residential, commercial or suburbs? What was the targeted population: high income, low income, youngsters or families? Should increased competition lead the store to increase promotion, lower its prices or shut shop? This indicates that every decision is bereft with variability and uncertainty. To deal with this aspect a manager relies on information which in turn is derived from data. Now, a manager confronts with huge amount of data on daily basis. How data is differentiated from information or relevant information which helps in decision making is extracted from huge amount of available data? Data is a combination of noise and information and statistics facilitates to segregate noise and information and obtain information for manager or decision maker. Thus, most of the businesses are affected by uncertainty and variability. So statistics is important to almost all business decisions.

5.2 Types of Data

Information in raw or unorganized form that refers to conditions or ideas is called as data. Data does not directly affect the concerned subject/party or person. The meaningful aspect of the data extracted by using certain statistical technique is called information. Information is accurate and presented within a context in an organized format. Information affects decision making. For example, a good monsoon leading to higher yield in crops and thereby increased income for farmers might be data for an urbanized dweller, but it can be information for a FMCG manager for whom a farmer is a prospective customer.

Data can be broadly categorized as qualitative and quantitative. Higher the rainfall better would be crop yield and higher would be the income of farmers. But how much rainfall would lead to how much increase in crop yield and it would result in how much increase in income that would allow farmers to spend on disposables. The former data is qualitative in nature as it approximates the data whereas latter data is quantitative in nature as it is definite and can be quantified and verified.

It is been emphasized in previous sections that informed and accurate decision depends on how statistics is used to segregate information from data by minimizing the impact of noise.

The first step in the entire process is collection of data regarding concerned problem. So, one of the prominent aspects in statistical analysis is understanding of type of data to be collected. Different types of data can be used and interpreted in different manner. For instance, data regarding a students' roll number in the class, his grade in subject of statistics and his marks in the subject are different from each other and are analyzed differently. All such data should not be analyzed the same way statistically because the entities represented by numbers are different. For this reason, researchers need to know the level of data measurement represented by numbers being analyzed. Four common levels of data measurement have been discussed:

5.2.1 Nominal data

Numbers that are used only to classify or categorize represent nominal data. For example, jersey number of players, car registration numbers, rolls number of students etc. Such data does not provide a value statement. Student with roll number one does not indicate his/her superiority over roll number ten student in any way. Many demographic questions in surveys result in data that are nominal because the questions are used for classification purpose only. For instance, educational qualification asked by a researcher in a survey assigned '1' to graduate, '2' to post-graduate and '3' to above post graduate. The assignment of 3 does not in any way is used to imply that such a respondent is better or more qualified than a respondent with assignment '1'. This is done just for classification purposes and is done to differentiate a respondent from another. Few statistical techniques are applicable for such data such as chi-square and some other non-parametric tests. PAN card numbers, bank account numbers, ZIP codes, etc. are some other examples of nominal data. Nominal data is sometimes also used to represent qualitative data.

5.2.2 Ordinal data

Ordinal data involve ranking data in a particular order. The order can be either ascending or descending. For instance, if in a class of statistics course a student gets A grade and other B and another student C grade, then it can be interpreted that A grade student has got more marks then the students who have got B and C grade. Similarly student with B grade has scored more than student with C grade but less than student with A grade. But ordinal data does not indicate the interval between data. Student who has got A grade is better than student with B grade but he has scored how much higher than other is not told by ordinal data. In various surveys five point Likert scale is used to measure a variable. For instance, consider a restaurant which wants to find out the customers' response towards taste of the meal measured on a five point Likert scale designed from Very Good, Good, Average, Bad and Very Bad. If scale is arranged in such a way then the data to be measured is arranged in an order where Very Good indicates higher score than Very Bad. But the scale does not indicate the difference between the ordered data i.e. the quantitative difference between very good and good is not indicated in the ordinal data. Various statistical tools are applicable for ordinal data such as mean, median to describe the data. But mode cannot be applied as mode indicates only highest frequency occurring data which is applicable mostly for nominal data.

5.2.3 Interval data

Interval data involves assigning quantitative differences between two consecutive numbers. The differences represented between consecutive numbers are equal i.e. interval data have equal intervals. For instance in the above example of restaurant if Very Good is assigned '1', Good as '2', Average as '3', Bad as '4' and Very Bad as '5' then the difference between two consecutive value is one which is equal between every interval. Interval data can also be applied in the case of students' grade illustration. If corresponding to each student grade is given their marks then the difference between marks of students under study can be calculated showing the difference between two data values. Thus, roll numbers indicate nominal data, grades ordinal data and marks interval data. This kind of data is most appropriate for various statistical techniques such as regression, t test, analysis of Variance etc. as it is easy to study quantitative values than qualitative values.

5.2.4 Ratio data

Ratio data have all the properties of interval data but it has an absolute zero and ratio of two numbers is calculated with certain meaning. The notion of zero means that base value is fixed and it cannot be arbitrarily assigned. For example if a person has weight of 50 kgs and another has 100 kgs, then it can be interpreted by using ratio that second persons' weight is two times more than the first person. In this example, calculated ratio has a meaning whereas if in case of interval data ratio is calculated it would result in meaningless results. For example, in the restaurant example if one respondent gives taste of the meal as '1' indicating very good and another respondent gives '2' indicating good and then ratio of their responses is found i.e. 1 / 2 or 2/1. The ratio value does not have any meaning and cannot be interpreted. Other examples of ratio data involve production cycle time, race time by athletes, number of shoes sold by two different stores etc. As ratio data is a metric data so this data can be analyzed by any of the statistical techniques.

5.2 Self check questions

1. Classify each of the following as nominal, ordinal, interval or ratio data:
   (i) Time required to produce a tyre
   (ii) Liters of milk a family drinks per month
   (iii) Ranking of four courses in the university designated as excellent, good, satisfactory and poor.
   (iv) Age of employees
   (v) An employees' identification number
   (vi) Response time of an emergency unit.

5.3 DATA COLLECTION

This chapter in the beginning discussed the importance of collection of data for accurate decision making by applying statistics. The collected data should pertain to the research problem at hand.

Before collecting data through primary sources like surveys it is important to get data through secondary sources like published articles to understand the research problem.

5.3.1 Secondary data

Secondary data is the data which got its origin for some other purpose. For instance, if research problem at hand was to find reasons for decrease in sales of washing machine which requires demographic data of consumers in a particular geographic area. But another study was already conducted involving determining sources of finance consumers use to buy washing machines from bank records. For this study the demographic data had been collected. So, this demographic data becomes a source of secondary data for research problem at hand.

Advantages of Secondary Data

The most significant advantages of secondary data are the cost and time economies they offer the researcher. If the required information is available as secondary data then collection of it should take no more than a few days and would involve little cost. On the other hand, a field survey would involve time and cost for creating, testing, delivering, collecting data; data coded, punched, and tabulated. With secondary data expenses have been incurred by the original source of the information and do not need to be borne by the user. Expenses are shared by the users of commercial sources of secondary data, but even here the user's costs will be much less than they would be if the firm collected the same information itself. The time and cost economies prompt the general caution: Do not bypass secondary data. Begin with secondary data, and only when the secondary data are exhausted or show diminishing returns, proceed to primary data. Thus, secondary data would typically

- help to better state the problem under investigation
- suggest improved methods or data for better coming to grips with the problem
- provide comparative data by which primary data can be more insightfully interpreted.

Disadvantages of Secondary Data

When confronted by a new problem, the researcher's first attempts at data collection should logically focus on secondary data. Secondary data are statistics gathered for some other purpose, in contrast to primary data, which are collected for the purpose at hand. Certain disadvantages associated with secondary data are discussed below.

- Secondary data possess significant cost and time advantages and it is only when their pursuit shows diminishing returns and the problem is not yet resolved that the researcher should proceed to primary data.
- Since secondary data are collected for other purposes, it will be rare when they fit perfectly the problem as defined. In some cases, the fit will be so poor as to render them completely inappropriate.
- It is not uncommon for secondary data to be expressed in units different from those deemed most appropriate for the project. Size of retail establishment, for instance, can be

expressed in terms of gross sales, profits, square feet, and number of employees. Consumer income can be expressed by individual, family, households, and spending unit.

- Finally, secondary data quite often lack publication currency. The time from data collection to data publication is often long, sometimes as much as three years, for example, as with much government census data. While census data have great value while current, this value diminishes rapidly with time, as many marketing decisions require current, rather than historical, information.

5.3.2 Types of Secondary Data

There are a number of ways by which secondary data can be classified. One of the most useful is by source, which immediately suggests the classification of internal and external data.

Internal data are those found within the organization for whom the research is being done, while external data are those obtained from outside sources. The sales and cost data compiled in the normal accounting cycle represent promising internal secondary data for many research problems. This is particularly true when the problem is one of evaluating past marketing strategy or of assessing the firm's competitive position in the industry. It is less helpful in future directed decisions, such as evaluating a new product or a new advertising campaign. Two of the most significant advantages associated with internal secondary data are their ready availability and low cost.

The ***external*** sources can be further split into those that regularly publish statistics and make them available to the user at no charge, and those organizations that sell their services to various users. The many standardized marketing information services that are available are another important source of secondary data for the marketing researcher. These services are available at some cost to the user and in this respect is more expensive source of secondary data than published information. This section reviews some of the main types and some of the main sources of standardized marketing information service data.

Industry Services are services available to the consumer goods manufacturer than to the industrial goods supplier. The consumer goods services are also much older than the industrial goods services. For instance, whereas the Nielsen Retail Index dates from 1934, the industry information services were born in the 1960s. This means that the industrial goods services are still evolving in terms of the type of information being collected and how it is made available to users.

Consumer Services A number of standardized marketing information services directly involve consumers and their behavior. Some are concerned with purchase or consumption behavior, some with viewing and reading habits while still others are used for a variety of purposes

5.3.3 Types of Primary Data

Demographic/Socioeconomic Characteristics: One type of primary data of great interest to marketers is the subject's demographic and socioeconomic characteristics, such as age,

education, occupation, marital status, sex, income, or social class. These variables are used to cross classify the collected data and in some way make sense of it. We might be interested, for instance, in determining whether people's attitudes toward ecology and pollution are related to their level of formal education. Alternatively, a common question asked by marketers is whether the consumption of a particular product is related in any way to a person's or family's age, education, income, and so on.

Attitudes/Opinions: Attitude is one of the more important notions in the marketing literature, since it is generally felt that attitudes are related to behavior. Obviously, when an individual likes a product he will be more inclined to buy it than when he does not like it; when he likes one brand more than another, he will tend to buy the preferred brand. Attitudes may be said to be the forerunners of behavior. Thus, marketers are often interested in people's attitudes toward the product itself, their overall attitudes with respect to specific brands, and their attitudes toward specific aspects or features possessed by several brands.

Awareness/Knowledge: Awareness/knowledge as used in marketing research refers to what respondents do and do not know about some object or phenomenon. For instance, a problem of considerable importance is the effectiveness of magazine ads. One measure of effectiveness is the product awareness generated by the advertisement. Awareness and knowledge are also used interchangeably when marketers speak of product awareness. Marketing researchers are often interested in determining whether the respondent is aware of the product, its features, where it is available.

Behavior: Behavior concerns what subjects have done or are doing. Most typically in marketing this means purchase and use behavior. It takes place under specific circumstances, at a particular time, and involves one or more actors or participants. The focus on behavior then involves a description of the activity with respect to the various components.

5.3.4 Basic Means of Obtaining Primary Data

The researcher attempting to collect primary data has a number of choices to make among the means that will be used. The primary decision is whether to employ communication or observation. Communication involves questioning respondents to secure the desired information, using a data collection instrument called a questionnaire. The questions may be oral or in writing, and the responses may also be given in either form. Observation does not involve questioning. Rather, it means that the situation of interest is checked and the relevant facts, actions, or behaviors recorded. Choosing a primary method of data collection implies a number of supplementary decisions. For example, should we administer questionnaires by mail, over the telephone, or in person? Should the purpose of the study be disguised or remain undisguised? Should the answers be open ended or should the respondent be asked to choose from a limited set of alternatives? A decision with respect to method of administration, say, has serious implications regarding the degree of structure that must be imposed on the questionnaire. Depending on structure and purpose of questionnaire following communication methods have been discussed.

Survey method involving a set of questions is a structured and undisguised form of data collection. Such method does not involve asking ambiguous questions as purpose is not to study hidden or latent behaviour. Questions pertaining to number of magazines being read in a month, amount of money spend on eating out, quality of services used are some of the examples that fall in this category of data collection. The method is used extensively as it is both time and cost effective. Surveys are widely used by market researchers to determine the preferences and attitudes of consumers. The results can be used for a variety of purposes from helping to determine the target market for an advertising campaign to modifying a candidates' platform in an election campaign. For example, a television network might conduct a survey to profile characteristics of owners of luxury automobiles including what they watch on television and at what times. This information can be very useful in formulating an advertisement theme for an automobile company. Some basic points to consider regarding questionnaire design are as follows:

- Questionnaire should be kept as short as possible to encourage respondents to complete it.
- Questions should be simple and clearly worded to enable respondents to answer quickly, correctly and without ambiguity.
- Open ended questions are useful in providing free expression but are time consuming and difficult to analyze.
- Leading questions should be avoided as they lead respondent to answer in a particular way.
- It is useful to pre-test a questionnaire.

Personal interviews allow the researcher to handle complex issues more effectively. Cooperation from respondents is more as talking to someone one-on-one helps in rapport and confidence building. The method helps in finding out additional details that might not emerge from initial responses. Issues which are sensitive such as incidents of drug abuse or issues relating to personality traits are better dealt by using personal interview method. Unfortunately, individual interviewing can be quite expensive and may be intimidating to some who are not comfortable sharing details with a researcher.

Focus Groups overcome the drawbacks associated with personal interview. Under this research format, a group of respondents (generally numbering 8-12) are guided through discussion by a moderator. The power of focus groups as a research tool rests with the environment created by the interaction of the participants. In well-run sessions, members of the group are stimulated to respond by the comments and the support of others in the group. In this way, the depth of information offered by a respondent may be much greater than that obtained through individual interviews. However, focus groups can be costly to conduct especially if participants must be

paid. Also, a respondent may get influenced by opinions of others. In case of sensitive subjects participants may be hesitant to share with others.

Projective techniques are unstructured and disguised forms of collecting behavioural data. This method involves presentation of ambiguous and unstructured object or activity that a respondent is asked to respond. Word Association, Sentence completion and picture interpretation test are certain projective techniques to reveal hidden feelings and opinions with which respondents might be unaware of. The communication method of data collection has the general advantages of versatility, speed, and cost, while observational data are typically more objective and accurate. *Versatility* is the ability of a technique to collect information on the many types of primary data of interest to marketers. A respondent's demographic and socioeconomic characteristics, the individual's attitudes and opinions, awareness and knowledge, intentions, the motivation underlying the individual's actions, and even the person's behavior may all be ascertained by the communication method. Observation is limited in scope to information about behavior and certain demographic and socioeconomic characteristics. But there are certain limitations to these observations. Observations are limited to present behavior, for example. A person's past behavior cannot be observed. Nor person's intentions as to future behavior can be observed.

The *speed and cost* advantages of the communication method are closely intertwined. Communication is a faster means of data collection than observation because it provides a greater degree of control over data gathering activities. The researcher is not forced to wait for events to occur with the communication method as she or he is with the observation method. In some cases, it is impossible to predict the occurrence of the event precisely enough to observe it. For still other behavior, the time interval can be substantial. For instance, an observer checking for brand purchased most frequently in one of several appliance categories might have to wait a long time to make any observations at all. Much of the time the observer would be idle.

Observation method might suffer from disadvantages of limited time, scope and cost but this method has advantage of being more objective and accurate. This is because the observational method is independent of the respondent's unwillingness or inability to provide the information desired. For example, respondents are often reluctant to cooperate whenever their replies would be embarrassing, humiliating, or would in some way place them in an unfavorable light. Observation typical produces more objective data than communication. The interview represents a social interaction situation. Thus the replies of the person being questioned are conditioned by the individual's perceptions of the interviewer. The same is true of the interviewer, although the interviewer's selection and training affords the researcher a greater degree of control over these perceptions than those of the interviewee. With observation, though, the subject's perceptions play less of a role. Sometimes people are not even aware that they are being observed. This removes the opportunity for them to tell the interviewer what they think the interviewer wants to hear, or to give socially acceptable responses.

5.3.5 Questionnaire

Following are the basic principles for drafting the questionnaire:

1. **Covering Letter**: The person conducting the survey must introduce himself and make the aims and objectives of the enquiry clear to the informant. A personal letter can be enclosed indicating he purposes and aims of enquiry. The informant should be taken into confidence. He should be assured that his answers will be kept confidential. A self - addressed and stamped envelope should be enclosed for the convenience of the informant to return the questionnaire
2. **Number of Questions**: Minimum number of questions based on the objectives and scope of enquiry only should be asked. More the number of questions, lesser the possibility of good and proper response. Fifteen to twenty five questions should be sufficient for making the required enquiry. Lengthy questions should preferably be divided into simple parts, and irrelevant questions should be avoided.
3. **Personal Questions**: Personal questions like asking about his addictions should be avoided. The Informant may not desire to answer such questions which may disclose his confidential, private or personal information. Questions affecting the sentiments for the informants should not be asked.
4. **The questions should be simple and clear** : The language of the questions should be easy to understand
5. **The questions should be arranged logically:** It helps in classification and tabulation of data. It is not logical to ask a man his income before asking him whether he is employed or not. There should be a proper sequence of the questions.
6. **Instructions to the Informants:** Clear and definite instructions for filling in the questionnaire and address where completed questionnaire should be sent must be given.
7. **The questions should be divided and subdivided under different heads and subheads** : The question should be divided and subdivided under proper heads and subheads and should be properly numbered for the convenience of the informant and the investigator
8. **Multiple Choice Questions :** Questions should be framed in such a way that the answers are factual or objective and the informant should be able to give the answers simply by using a tick mark in the blank

   Which of the following languages you use most for writing? (Put a tick mark)

   | | English | | Hindi | | Punjabi | | Urdu | | Any other

9. **Simple Alternative Questions.** (Yes/No) As far as possible the question should be framed in such a way that they are answerable in 'Yes' or 'No' or ' Right' or 'Wrong' e.g.
   - Are you married? Yes/No
   - Are you employed? Yes/No
10. **Specific Information Questions**: We get specific answers to certain types of

questions. These questions are simple and direct.

- In which class do you read?
- How many brothers have you?
- What is your mother tongue?
- What is your father?

11. **Open Question**: Open question makes the informant free to give any reply he chooses. Such questions are difficult to tabulate and increase labor in statistics work and should be minimum.
    Example:

    (a) Suggest the measures to solve the problems of poor students in University of Delhi
    (b) How will you solve the wage problem in your industry?

12. **Relevant Question**: The question should be directly related to the point under enquiry for which the data is being collected.
13. **Avoidance of Leading Questions**: As far as possible leading questions should be avoided. Why do you like 'Taj Mahal Tea'? Instead of such simple question, two questions can be framed for enquiry, namely.
14. **Attractive Layout**: The questionnaire should be made to look as attractive as possible, keeping in view the possible answer to the questions of schedule, sufficient space should be provided.

5.3 Self check questions True/False

a. Begin with primary data and check the collected data with secondary references.
b. Secondary data provides comparative data by which primary data can be more insightfully interpreted.
c. Observation method of collecting primary data is more cost effective than communication method.
d. Consumers' behavior towards fast food products can be better evaluated by using survey method than observation method.

5.4 SUMMARY

Statistics is an important decision-making tool in business and is used in virtually all areas of business. The main objective of this field is to provide a decision maker relevant information from available huge amount of data. This chapter has discussed nominal, ordinal, interval and ratio data by illustrating few examples. Also, two branches of statistics viz. descriptive and inferential statistics have been discussed. Both types of statistics should be applied before reaching a conclusion. Descriptive statistics only make decisions about the group for which data has been collected whereas inferential statistics helps to make decisions about larger set of population if sample of subjects belong to selected population.

5.5 GLOSSARY

- **Secondary data:** is already published data which was collected for a purpose, but it can be used for current research problem. The ways of collecting secondary data can be both internal and external.
- **Primary data:** is the data which needs to be collected regarding attitudes, behavior and intentions of customers or respondents regarding research problem on hand. The ways of collecting primary data involves communication and observation method.

5.6 SHORT ANSWER QUESTIONS

1. What is the difference between primary and secondary data?
2. What distinguishes internal secondary data from external secondary data?
3. What are the cautions that you should take before using secondary data?

5.7 LONG ANSWER QUESTIONS

1. Explain the advantages and disadvantages of primary and secondary data?
2. Explain the various sources of collecting primary and secondary data?
3. What is a questionnaire? What points you need to consider while drafting a questionnaire?

5.8 ANSWERS TO SELF CHECK QUESTIONS

5.2 a. Ratio
b. Ratio
c. Ordinal
d. Interval
e. Nominal
f. Ratio
5.3 a. False
b. True
c. False
d. True

5.9 SUGGESTED READINGS

- Black, K., *Business Statistics For Contemporary Decision Making*, Fifth Edition, Wiley India,
- Keller, G., *Statistics for Management,* First India Reprint 2009, Cengage Learning India Private Limited.
- Donald R. Cooper & Pamela S. Schindler, *Business Research Methods,* Tata McGraw-Hill Publishing Co. Ltd., New Delhi, 9th Edition.
- S.P. Gupta, *Business Statistics*, Sultan Chand, New Delhi, 2006.

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS &RESEARCH METHODOLGY: BSRM 201

LESSON No. 6 **AUTHOR – SAHIL RAJ**

INTRODUCTION TO STATISTICS: STATISTICS AND BUSINESS RESEARCH

STRUCTURE

6.0 OBJECTIVES

The purpose of this lesson is:

- Introduction of the concept of Statistics
- Familiarize the reader with the Techniques of Statistics
- Application of Statistics in Business and Research

6.1 INTRODUCTION

The word Statistics' is very popularly used in practice, ft conveys a variety of meaning to people many which are inaccurate or, at the very least, misleading. The average person conceives of 'statistics' as column of figures, zigzag graphs or tables like statistics of production, consumption etc. In addition to meaning numerical facts, ’statistics' also refers to a subject , as well as symbols, formulae and theorems and 'accounting' refers to principles and methods as well as to accounts, balance sheets and income statements. In this sense 'statistics' is a body of methods for obtaining and analyzing data in order to base decisions on them. It is a branch of scientific methods used in dealing with phenomena that can be described numerically either by count or

by measurement. Thus, the word statistics refers either to quantitative information or to a method of dealing with quantitative information.

The methods by which statistical data are analyzed are called statistical methods, although the term is sometimes used more loosely to cover the subject 'statistics' as a whole. Statistical methods are applicable to a very large extend in the field of economics, sociology, anthropology, business, agriculture and education. Statistical methods are also used by government bodies, private business firms and research agencies as an indispensable aid in forecasting, controlling and exploring.

Thus statistics may be defined as the science of collection, organization, presentation, analysis and interpretation of numerical data.

According to the above definition, there are five stages in a statistical investigation.

1. **Collection:** Collection of data constitutes the first step in a statistical investigation. Utmost care must be exercised in collecting data because they form the foundation of statistical analysis. If data are faulty, the conclusions drawn can never be reliable.
2. **Organizations**: Data collected from published sources are generally in organized form. However large mass of figures that are collected from a survey frequency needs organization. The first step is editing. The collected data must be edited very carefully so that the omission, inconsistencies, irrelevant answers and computation may be corrected. After the edition the next step is to classify some characteristics possessed by the items constituting the data. The last step tabulation which means to arrange the data in columns and rows so that there is clarity in the data.
3. **Presentation:** After the data have been collected and organized they are ready for the presentation. Data presented in an orderly manner facilitates statistical analysis.
4. **Analysis:** After collecting, organizing and presentation the next step is analysis. Methods used in analyzing the presented data are numerous ranging from simple observation of data to complicated, sophisticated and highly mathematical techniques.
5. **Interpretation:** The last stage in statistical investigation is interpretation i.e. drawing conclusions from the data. Correct interpretation will lead to a valid conclusion of study and thus aid in decision making.

According to Prof. Horace Secrist who defined statistics as follows -

"By Statistics we mean aggregates of facts affected to a marked extend by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a

predetermined purpose and placed in relation to each other"

This definition clearly points out certain characteristics which numerical data possess in order that they may be called statistics. These are as follows:

1. **Statistics are aggregates of facts:** Single and isolated figures are not statistics for the simple reason that such figures are unrelated and cannot be compared.
2. **Statistics are affected to a marked extend by multiplicity of causes:** Facts are affected to a considerable extend by a number of forces operating together.
3. **Statistics are numerically expressed:** All statistics are numerical statements of facts. Qualitative statements do not constitute statistics. The reason is that these statements are vague and one cannot make out anything.
4. **Statistics are enumerated according to reasonable standards of accuracy:** Facts and figures about any phenomenon can be derived in two ways, by actual counting and measuring or by estimate. Estimates cannot be as precise as actual counts. However in some cases 100 per cent accuracy is difficult to attain. Hence in many statistical studies reasonable standards of accuracy should be attained.
5. **Statistics are collected for a predetermined purpose:** The purpose of collecting data must be decided in advance. The purpose should be specific and well-defined. A general statement of purpose is not enough.
6. **Statistics are collected in a systematic manner:** Before collecting statistics a suitable plan of data collection should be prepared and the work carried out in a systematic manner.
7. **Statistics should be placed in relation to each other:** If numerical facts are to be called statistics, they should be comparable. Statistical data are often compared period wise or region wise.

6.1 Self check questions

a) Define statistics
b) Statistics is an aggregate of facts and are numerically expressed. True/False

6.2 ORIGIN AND GROWTH OF STATISTICS

It may be interesting to point out that statistics is not a new disciple but as old as the human itself. Though, in the present usage, the word "statistics" is barely a century old, it has been in use for a much longer period. According to Greek historian, in 1400 B.C. a census of all lands in Egypt was taken. Similar reports on the ancient Chinese and Romans are also available. The word "statistics" comes from Italian word

"Statista" meaning "Statesman" or the German word "statistic" which means a political state. It was first used by Professor Gottfried Achenwall, a professor in Marborough in 1749 to refer to the subject-matter as a whole. Achenwall defined statistics as the "the political science of the several countries". The word 'statistics' appeared for the first time in the famous book, Elements of Universal Erudition by Baran J.F. Von Biefeld, translated by W.Hooper M.D.(3 Vols..London, 1770). One its chapter is entitled as Statistics' and contains a definition of the subject as the "the science that teaches us what is the political arrangement of all the modem States of the known world".

There are four important reasons explaining why managers should have knowledge and understanding of Statistics:

1. To be able to properly present and describe information.
2. To be able to use information obtained from samples to draw conclusions about large populations.
3. To be able to make improvements in processes.
4. To be able to make reliable forecasts.

Managers need to apply statistical techniques to aid in their decision making.

6.3 STATISTICAL TECHNIQUES

These statistical techniques can be classified into two broad categories:

1. Descriptive Statistics
2. Inferential Statistics.

Descriptive Statistics: Descriptive Statistics includes any treatment designed to describe or summarize the given data, bringing out their important features. These statistics do not go beyond this. This means no attempt is made to infer anything that pertains to more than the data themselves. Thus if someone compiles the necessary data and reports that during the year 2006-2007, there were 3000 public limited companies in India of which say, 1500 earned profit and 1500 sustained losses, his study belongs to the domain of descriptive statistics. He may further calculate the average profit earned per company and also the average loss sustained per company. Methods used in descriptive statistics may be called as descriptive methods. Under descriptive statistics there are certain methods e.g. frequency distribution, measures of central tendency, that is, averages, measures of dispersion and skewness. Thus descriptive statistics includes-

- Collect
- Organize
- Summarize
- Display
- Analyze

Inferential Statistics: Although descriptive Statistics is an important branch of statistics and it continues to be so, its recent growth indicates a shift in emphasis towards the methods of Statistical inference. The methods of Statistical inference are required: predict the demand for a product for a specified year. Inferential Statistics are also important while comparing the effectiveness of a given medicine in the treatment of any disease while determining the nature and extent of relationship between the variables comes under the preview of Inferential Statistics. Thus Inferential Statistics include following steps-

- Predict and forecast values of population parameters
- Test hypotheses about values of population parameters
- Make decisions

6.3 Self check questions

a) What type of Statistics includes any treatment designed to describe or summarize the given data
b) What are the different types of statistical techniques?

6.4 STATISTICS AND BUSINESS RESEARCH

With the increase in competition in globalize world the decision making in business enterprise is becoming very complex. Moreover the decision once taken has far reaching implications on the enterprise. So it is very important for any manager to take rational as well as quick decisions. Management nowadays is a very specialized job and specialized managers are doing this specialized job. For example Marketing manager is given the job for marketing a product and financial manager is given the job for managing finance of the enterprise. Moreover modern business firms are working in great deal of uncertainty concerning future operations and managers are taking decision in anticipation of demand. So a manager has taken practical decisions without any chance for the method of trial and error. Manager has to analyze the situation carefully and apply certain statistical tools to reach to a decision.

Business activities can be divided into:

1. Marketing
2. Production
3. Finance
4. Banking
5. Purchase
6. Accounting
7. Control
8. Credit
9. Personnel

10. Research and Development

1. **Marketing:** Statistical analysis is frequently used in providing information for marketing decisions. In the field of marketing, it is necessary first to find out what can be sold and evolve a suitable strategy so that goods reach the ultimate consumer. A skilful analysis of data on population, purchasing power, habits, competition, transportation cost etc should precede any attempt to establish the market. The analysis may reveal that in certain areas where one thought of big markets potential, there hardly exists any scope.
2. **Production:** In the field of production, statistical methods play a very important role. The decision about what to produce, how much to produce, when to produce, for whom to produce is based largely on facts. Statistical tools help immensely in quality control, inventory levels and dealing the labour problems.
3. **Finance:** The financial managers in discharging their finance function efficiently depend on the statistical analysis of facts and figures. Financial forecasting, breakeven analysis and investment decisions under uncertainty are part of their activities. Managers take the help of various models. These models involve application of several statistical concepts.
4. **Banking:** Banking institutions have found it necessary to establish departments within their organizations for the purpose of gathering and analyzing information, not only regarding their own operations, but on general economic conditions and every line of business in which they might be interested. Probably the banks, more than any other individual business, feel the direct effect of the conditions in every type of business and need to be constantly informed as to the trends in every line of activity. Thus the banks use the objective analysis furnished by statistics and then temper their decisions on the basis of qualitative information.
5. **Purchase**: The purchase department in discharging its function makes use of statistical data to frame suitable purchase policies such as from where to buy, how much to buy, at what time to buy and at what price to buy.
6. **Accounting**: Statistical methods are also employed in accounting. In particular the auditing functions make frequent application of statistical sampling and estimation procedures.
7. **Control**: The management control process combines statistical and accounting methods in making the overall budget for the coming year including sales, material, labour and other costs.
8. **Credit**: The credit department performs statistical analysis to determine how much credit to extend to various customers.
9. **Personnel**: The personnel department frames personnel policies based on

facts. It makes the statistical studies of wage rates, incentive plans, cost of living etc. Such studies help the personnel department in the process of manpower planning.

10. **Research and Development**: Many organizations have R&D departments which primarily concerned with finding out how existing products can be improved; what new product lines can be added and how the optimal use of resources made. In the absence of factual data it is impossible to carry fruitful research and development.

Thus with the help of statistical tools in every above mentioned activity, manager can reach to an accurate decision. For example, a marketing researcher uses data of consumer buying habits to develop new products. A production manager uses quality control data to decide when to make decisions in production process. Statistical tables and charts are frequently used by sales managers to present numerical facts of sales. The techniques of time series analysis and forecasting enable the managers to predict the future demand. Similarly the concepts of central tendency are used by managers. For example the most widely used concept of mean or simply average is one of the simplest tool used by managers. The correlation and regression which not only tell about the relation between two or more variables but also tells about the extent of relation between the variables are important tools used in management decisions.

Similarly statistics is widely used in research process. The need for statistics is one of measurement and comparison. Research needs to validate or disprove processes, and uses statistics to determine if the current approach is worthwhile and accurate. Various statistical methods are used for research process. Starting from making an assumptions regarding particular phenomenon known as hypothesis in statistical terms to collection of data and finally applying various tests in statistics to verify that phenomenon. These tests vary from: chi square, z test, t test, ANOVA one way, ANOVA two way etc. The detailed explanation of these tests will be given in the following chapters.

6.5 LIMITATIONS OF STATISTICS

- There are certain concepts where statistics cannot be used. This is because these concepts are not amenable or measurement. For example beauty, intelligence, courage.
- Statistics reveal the average behavior, the normal or the general trend. An application of the 'average' concept if applied to an individual or a particular situation may lead to a wrong conclusion and sometimes may be disastrous.
- Data collected for a particular purpose may not be relevant or useful in other situations.
- Statistics is not 100 per cent precise as is Mathematics or Accountancy.
- In Statistical surveys, sampling is generally used as it is not physically possible to cover all the units or elements comprising the universe. The

results may not be appropriate as far as the universe is concerned.

- At times, relationship between two or more variables is studied in Statistics, but such a relationship does not indicate 'cause and effect' relationship. In such cases it is the user who has to interpret the results carefully.
- A major limitation of Statistics is that it does not reveal all pertaining to a certain phenomenon. There is some background information that Statistics does not cover. Similarly there are some other aspects related to the problem on hand, which are also not covered. The user of Statistics has to be well informed and should interpret Statistics keeping in mind all other aspects having relevance on the given problem.

6.5 Self check questions

a) Statistics is not 100 per cent precise. True/false

6.6 GLOSSARY

- **Correlation** - Statistical term which tells about relation between variables
- **Regression** - Statistical term which tells about extent of relation between variables
- **Inventory** - Items that are stored for future use.
- **Statistic** - German word which means Political state.

6.7 SHORT ANSWER QUESTIONS

1. What do you understand by the term 'Statistics'?
2. Explain the different types of statistical techniques.

6.8 LONG ANSWER QUESTIONS

1. Discuss the application of Statistics in business and research?
2. What are the limitations of Statistics?

6.8 ANSWER OF SELF CHECK QUESTIONS

6.1 a) Aggregates of facts affected to a marked extend by multiplicity of causes, numerically expressed, collected in a systematic manner for a predetermined purpose

b) True

6.3 a) Descriptive statistics

b) Descriptive and inferential statistics

6.5 a) True

6.10 SUGGESTED READINGS

- Levin & Rubin, 2007, **Statistics for Management,** Pearson Education, New Delhi, 7th Edition.
- Gupta, S. P., 2008, **Statistical Methods,** Sultan Chand 86 Sons, New Delhi, 14th Edition.

- Gupta, S. C., 2007, **Statistical Methods,** Sultan Chand & Sons, New Delhi, 33rd Edition.
- Beri, G. C., 2003, **Statistics for Management,** Tata McGraw-Hill, New Delhi.

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS &RESEARCH METHODOLGY: BSRM 201

Lesson No. 7 **AUTHOR: KAJAL**

MEASURES OF CENTRAL TENDENCY

STRUCTURE

7.0 Objectives

The objective of this lesson is to have a general understanding of measures of central tendency and Statistics as applicable to Business Management and its use and relevance in areas of Management Research.

7.1 Introduction

When dealing with numerical information, good data analysis involves not only presenting the data and observing what the data are trying to convey but also computing and summarizing the key features and analyzing the findings. In any analysis a variety of descriptive measures representing the properties of central tendency, variation and shape may be used to summarize the major features of the data. Most sets of data show a distinct tendency to group or cluster about a certain central point. Thus, for any particular set of data it usually becomes possible to select some typical value to describe the entire set. Such a descriptive typical value is a measure of central tendency. This chapter describes three measures of central tendency viz. mean, median and mode.

7.2 Measures of Central Tendency

When dealing with numerical information, good data analysis involves not only presenting the data and observing what the data are trying to convey but also computing and summarizing the key features and analyzing the findings. In any analysis a variety of descriptive measures representing the properties of central tendency, variation and shape may be used to summarize the major features of the data. Most sets of data show a distinct tendency to group or cluster about a certain central point. Thus, for any particular set of data it usually becomes possible to select some typical value to describe the entire set. Such a descriptive typical value is a measure of central tendency. This chapter describes three measures of central tendency viz. mean, median and mode.

7.3 Mean

The most popular and widely used measure of representing the entire data by one value is what most laymen call an 'average' and what the statisticians call the' arithmetic mean'. Its value is obtained by adding together all the items and by dividing this total by the number of items.

(i) Calculation of Mean: Individual Series

The process of computing mean in case of ***individual observations*** (where frequencies are not given) is very simple. Add together the various values of the variable and divide the total by the number of items. Symbolically:

$$\text{X bar} = (X1 + X2 + X3 + \ldots\ldots\ldots\ldots..+Xn) / N$$

Here X bar = Arithmetic Mean, N = Number of observations.

The formula involves two steps in calculating mean:

(i) Add together all the values of the variable X and obtain the total.

(ii) Divide this total by the number of the observations, i.e., N.

Example: The following data gives the monthly income of 10 employees in an office

Income (Rs.) 1780, 1760, 1690, 1750, 1840, 1920, 1100, 1810, 1050, 1950

Mean = (1780 + 1760 + 1690 + 1750 + 1840 + 1920 + 1100 + 1810 + 1050 + 1950) / 10

= 1665

(ii) Calculation of Mean: Discrete Series

The formula for computing mean is: $\overline{X} = \Sigma fX / N$

where, f = Frequency; X = The variable in question; N = Total number of observations, i.e. Σf

Steps:

- Multiply the frequency of each row with the variable and obtain total ΣfX.
- Divide the total obtained in previous step by the number of observations i.e., total frequency.

Example: From the following data of the marks obtained by 60 students calculate the arithmetic mean:

Table 10

Marks (X)	20	30	40	50	60	70
No. of students (f)	8	12	20	10	6	4
fX	160	360	800	500	360	280

N = 60 ΣfX = 2460

$\overline{X} = \Sigma fX / N = 2460/60 = 41$

(iii) Calculation of Mean – Continuous Series

The formula for computing mean is

$$\overline{X} = \Sigma fm / N$$

where m = mid-point of various classes; f = frequency of each class; N = total frequency

Steps:

- Obtain the mid-point of each class and denote it by m
- Multiply these mid-points by the respective frequency of each class and obtain the total Σfm

- Divide the total obtained in step (i) by the sum of the frequency *i.e.,* N.

Example: From the following data compute arithmetic mean by direct method:

Table 11						
Marks	0-10	10-20	20-30	30-40	40-50	50-60
Mid-point, m	5	15	25	35	45	55
No. of students, f	5	10	25	30	20	10
fm	25	150	625	1050	900	550

N=100, $\Sigma fm = 3300$

$\overline{X} = \Sigma fm/N = 3300/100 = 33$

7.3 Self-check questions

a) State whether the statements are true or false-

The data 6, 4, 3, 8, 9, 12, 13, 9 has mean 9.

b) Arithmetic mean is a positional value.

7.4 Median

The median by definition refers to the middle value in a distribution. In case of median one-half of the items in the distribution have a value the size of the median value or smaller and one-half have a value the size of the median value or larger. The median is just the 50th percentile value below which 50 per cent of the values in the sample fall. It splits the observation into two halves. As distinct from the arithmetic mean which is calculated from the value of every item in the series, the median is what is called a positional average. The term 'position' refers to the place of a value in a series. The place of the median in a series is such that an equal number of items lie on either side of it.

(i) Calculation of Median: Individual Series

- Arrange the data in ascending or descending order of magnitude.
- In a group composed of an odd number of values such as 7, add 1 to the total number of values and divide by 2. Thus, 7 + 1 be 8 which divided by 2 gives 4-the number of the value starting at either end of the numerically arranged groups will be the median value. In a large group of 199 items the middle value would be 100^{th} item.

Thus, Median = size of (N + 1) /2th item.

Example: Income of five employees is Rs. 900, 950, 1020, 1200 and 1280. After arranging in ascending order: 900, 950, 1020, 1200 and 1280 it is found that the median would be 1020.

For the above example the calculation of median was simple because of odd number of observations. When an even number of observations are listed, there is no single middle position value and the median is taken to be the arithmetic mean of two middlemost items. For example, if in the above case we are given the income of six employees as 900, 950, 1020, 1200, 1280, 1300, the median income would be: (1020 + 1200)/2 = 1110.

(ii) Computation of Median – Discrete series

Steps:

- Arrange the data in ascending or descending order of magnitude.
- Find out the cumulative frequencies.
- Apply the formula: Median = size of (N + 1) /2 th item
- Now look at the cumulative frequency column and find that total which is either equal to (N + 1)/2 or next higher to that and determine the value of the variable corresponding to it. That gives the value of median.

Example:

Table 12

Income (Rs.)	1000	1500	800	2000	2500	1800
No. of persons, f	24	26	16	20	6	30
Cumulative frequency	16	40	66	96	116	122

Median = (122 + 1)/2 = 61.5th item

Size of 61.5th item = 1500

(iii) Computation of Median – Continuous series

Determine the particular class in which the value of median lies. Use N/2 as the rank of the median and not (N + 1)/2. In a continuous frequency distribution all the frequencies lose their individuality. The effort now is not to find the value of one specific item but to find a particular point on a curve-that one which will have 50 per cent of frequencies on one side of it and 50 percent of the frequencies on the other. It will be wrong to use the rule. Hence it is N /2 which will divide the area of curve into two parts and as such we should use N/2 instead of (N + 1)/2, in

continuous series. After ascertaining the class in which median lies, the following formula is used for determining the exact value of median.

$$\text{Median} = L + \frac{(N/2 - c.f.)}{f} * i$$

L = lower limit of the median class

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

f = frequency of the median class

i = class interval of the median class

Example:

Table 13

Marks	Frequency, f	c.f.
5-10	7	7
10-15	15	22
15-20	24	46
20-25	31	77
25-30	42	119
30-35	30	149
35-40	26	175
40-45	15	190
45-50	10	200

Median = size of N/2 th item

= 200/2 = 100th item

So, Median lies in class 25-30

Median = L + [(N/2 –c.f.) / f] * i

= 25 + [(100-77) / 42] * 5 = 27.74

7.4 Self-check questions

a) State whether the statement is true or false-

The median is always one of the numbers in a data.

b) An incomplete distribution is given below:

Variable	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency 10	20	?	40	?	25	15	

It is given that median value=35 and total frequency = 170. Find out missing frequency.

7.5 Mode

The mode is the value in the set of data that appears most frequently. Unlike the arithmetic mean, the mode is not affected by the occurrence of any extreme values. However, the mode is used only for descriptive purposes, because it is more variable from sample to sample than other measures of central tendency.

(i) Calculation of Mode – Individual observations

For determining mode count the number of times the various values repeat themselves and the value occurring maximum number of times is the modal value.

Example: The data for a sample of 1-year total percentage returns achieved by domestic general stock funds whose marketing fees are paid from fund assets is presented as:

32.2, 29.5, 29.9, 32.4, 30.5, 30.1, 32.1, 35.2, 10.0, 20.6, 28.6, 30.5, 38.0, 33.0, 29.4, 37.1, 28.6

Compute the mode.

Solution: From the data it can be easily computed that 28.6 and 30.5 occur twice in the series. These two values occur the maximum number of times as compared to other values. Thus, the data has two modes – 28.6 and 30.5. Such data are described as bimodal.

(ii) Calculation of Mode – Continuous Series

The value of mode for a continuous series is determined by applying the formula:

$\text{Mode} = L + [(f_1 - f_0) / (2f_1 - f_0 - f_2)] * i$

L = lower limit of the modal class

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

Where mode is ill-defined, its value may be ascertained by the following formula based upon the relationship between mean, median and mode:

$$\text{Mode} = 3*\text{Median} - 2*\text{Mean}$$

Example: Calculate mode from the following table:

Table 14			
Marks	**No. of students**	**Marks**	**No. of students**
Above 0	80	Above 60	28
Above 10	77	Above 70	16
Above 20	72	Above 80	10
Above 30	65	Above 90	8
Above 40	55	Above 100	0
Above 50	43		

Solution: Since cumulative frequency is given in the distribution, thus it is first converted into a simple frequency distribution.

Marks	**No. of students**	**Marks**	**No. of students**
0-10	3	60-70	12
10-20	5	70-80	6
20-30	7	80-90	2
30-40	10	90-100	8
40-50	12		
50-60	15		

By inspection the modal class is 50-60

Mode = $L + [(f_1 – f_0) / (2f_1 – f_0 – f_2)] * i$

L = 50, f_1 = 15, f_0 = 12, f_2 = 12, i = 10

Substituting the values, we get

Mode = 50 + [(15-12) / (2*15 – 12 – 12)] * 10

= 50 + (3/6) * 10

= 55

7.5 Self-check questions

a) State whether the statement is true or false-

The mode is always one of the numbers in a data

b) The value that occurs the most frequently in a data set is termed as___.

7.6 GLOSSARY

- **Arithmetic Mean** : The mean is the most commonly-used type of average and is often referred to simply as the average. The term mean" or "arithmetic mean is preferred in mathematics and statistics to distinguish it from other averages such as the median and the mode.
- **Median** : In probability theory and statistics, a median is described as the number separating the higher half of a sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one.
- **Mode** : In statistics, the mode is the value that occurs the most frequently in a data set or a probability distribution. Like the statistical mean and the median, the mode is a way of capturing important information about a random variable or a population in a single quantity.

7.7 SHORT ANSWER QUESTIONS

1. From the following data compute arithmetic mean by direct method:

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	5	10	25	30	20	10

2. State whether the statements are True or False. Also explain. Mean, median and mode may be the same for some data.
3. Explain the relationship between mean, median and mode.
4. The runs scored by 11 players in the cricket match are as follows:

7, 16, 121, 51, 101, 81, 1, 16, 9, 11, 16

Find the median of the data.

7.8 LONG ANSWER QUESTIONS

1. Calculate mean, median and mode from the following frequency distribution:

Variable	**Frequency**	**Variable**	**Frequency**
10-13	8	25-28	54
13-16	15	28-31	36
16-19	27	31-34	18
19-22	51	34-37	9
22-25	75	37-40	7

2. Calculate median and mode of the data given below. Using them find arithmetic mean

Marks	10	20	30	40	50	60
No. of Students	8	23	45	65	75	80

3. The median and mode of the following wage distribution are known to be Rs.33.5 and Rs.34 respectively. Three frequency values from the table are however missing. Find these missing values.

Wages (Rs.)	**No. of Workers**
0-10	4
10-20	16
20-30	**o**
30-40	?
40-50	9
50-60	6
60-70	4

Total frequency is 230

4. Distinguish between mean, median and mode.

7.9 ANSWERS OF SELF CHECK QUESTIONS

7.3 a) False. Mean is 8

b) False.

7.4 a) True

b) 35,25

7.5 a) True

b) Mode

7.10 SUGGESTED READINGS

- G. C. Beri, **Business Statistics,** Tata McGraw-Hill Publishing Co. Ltd., New Delhi, 2nd Edition.
- J. K. Sharma, **Business Statistics**, Pearson Education, New Delhi, 3rd Reprint, 2005.

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS &RESEARCH METHODOLGY: BSRM 201

Lesson No. 8 **AUTHOR: KAJAL**

MEASURES OF DISPERSION

Structure

8.0 OBJECTIVES

After reading this chapter, the reader should be able to :

- Understand the importance and significance of Dispersion.
- Analyze the different methods of Dispersion.
- Practically apply & evaluate the concepts of consistency and variation.

8.1 INTRODUCTION

As we have seen in the previous chapters, the mean, median and mode are just measures of central tendency and do not indicate the extent of variability in a distribution. It is necessary to describe the variability or the dispersion of the observations. In 2 or more distributions, the central value may be same but still there can be wide disparities in the formation of the distribution.

Dispersion has been defined by various statisticians as follows :

Dispersion is the measure of variation of the items."

- A. L. Bowley

The degree to which the numerical data tends to spread about an average value is called the variation of dispersion of the data."

Spiegel

Dispersion or spread is the degree of the scatter or variation of the variable about a central value."
— Brooks & Dick

8.2 SIGNIFICANCE OF MEASURING DISPERSION

- Measures of variation point out as to how far an average is representative of the mass. When dispersion is small, the average is a typical value in the sense that it closely represents the individual value and it is reliable in the sense that it is a good estimate of the average in the corresponding universe. On the other hand, when dispersion is large, the average is not so typical, and unless the sample is very large, the average maybe quite unreliable.
- Another purpose of measuring dispersion is to determine nature and cause of variation in order to control the variation itself. In social sciences a special problem is requiring the measurement of variability is the measurement of inequality of the distribution of income or wealth etc.
- Measures of dispersion enable a comparison to be made of 2 or more series with regard to their variability. The study of variation may also be looked upon as a means of determining uniformity of consistency. A high degree of variation would mean little uniformity or consistency whereas a low degree of variation would mean great uniformity or consistency.
- Many powerful analytical tools in statistics such as correlation analysis, testing of hypothesis, analysis of variance, Statistical Quality control etc are based on measures of variation of one kind or the other.

8.3 DESIRABLE CHARACTERISTICS OF MEASURE OF DISPERSION

Features	Desirable Characteristics
Understanding	It should be easily understood.
Utilizing available information	It should utilize all the values of the observations in the data
Ease of calculations	It should be easy to calculate
Sampling variations	If samples are taken from a population, the values calculated from various samples should not vary much from each other. If it is so, then the measure is considered reliable.
Mathematical Manipulation	It should be amenable to mathematical manipulations so that further meaningful conclusions about the data can be made.

Merits

1. They indicate the dispersal character of the statistical series.
2. They speak of the reliability or dependability of the average value of a series.
3. They enable the statisticians in making comparison between 2 or more statistical series with regard to the character of their uniformity or consistency or equitability.
4. They enable one in controlling the variability of a phenomenon ubder the purview.
5. They facilitate in making further statistical analysis of the series through the devices like coefficient of Skewness, coefficient of Kurtosis, etc

Demerits

1. They are liable to misinterpretation and wrong generalization by a statistician of biased character.
2. They are liable to yield inappropriate results as there are different methods of calculating dispersion.
3. Excepting one or two, most of the methods of dispersion involve complicated processes of computation.

They by themselves cannot give any idea about the symmetrical or skewed character of a series

8.3 Self-check questions-

a) Which of the following are methods under measures of dispersion?
   a. Standard deviation
   b. Mean deviation
   c. Range
   d. All of the above

b) Which of the following is not a characteristic of a good measure of dispersion?
   a. It should be rigidly defined
   b. It should be based on extreme values
   c. It should be capable of further mathematical treatment and statistical analysis
   d. None of the above

8.4 Measures of Dispersion for Ungrouped Data

The measures of dispersion for ungrouped data that have been discussed in this section are: range, interquartile range, mean absolute deviation, variance and standard deviation.

8.4.1 Range

Range represents the difference between largest and smallest value in the raw data. An advantage of range is that it is easy to calculate. But as it uses only two values from entire data ignoring the importance of other values in the data, so this method tends to relevant for small data rather than large data. Range method of finding variability in the data is mostly used in cases such as finding range of height of students in a class, difference between costliest and cheapest product, in quality assurance to construct control charts etc.

For instance, for the following data set:

4 3 0 5 2 9 4 5

Range = Highest value – Smallest value

= 9 – 0 = 9

8.4.1 Self-check questions

a) The range of 10 20 30 40 is _____.

b) The range represents _____.
   a. The lowest number
   b. The highest number
   c. The middle number

d. The difference between the lowest and highest number

8.4.2 Inter-quartile Range

Interquartile range is the difference between first and third quartile values of data set. It is the range of the middle 50% of the data and is determined by computing the value of difference between third and first quarter. To understand interquartile range lets first observe the concept of quartiles.

Quartiles are measures of central tendency that divide the entire data into four subgroups. The first quarter Q1 divides the lowest 25% of data with the upper 75% of data. The third quartile on the other hand divides the 75% of data from the last quarter i.e. from 25% of remaining data. Lastly, second quarter divides the entire data into two equal halves and is equivalent to median. The following example illustrates the calculation first and third quartiles.

Data set: 106 109 114 116 121 122 125 129

1. Arrange the data in an order: either ascending or descending. In the above example the data is already set in an ascending order.
2. Number of values in the data set (N) = 8
3. The quartile location indicated by i = (N+1)/4
   = (8+1)/4 = 2.25
4. First quartile (Q1) = average of second and third numbers
   = (109 + 114)/2
   = 111.5

This indicates that one fourth or 25% of given data (in this case two values) are smaller than 111.5.

Similarly for third quarter i = 3(N+1)/4
= 3(8+1)/4
= 6.75

So, Q3 = average of sixth and seventh numbers
= (122 + 125)/2 = 123.5

This indicates that three-fourths or 75% of data (in this case six values) are less than 123.50.

Now coming back to interquartile range, its calculation is interpreted by using following example:

Data set: 6 2 4 9 1 3 5

Step 1: arrange the data in an order: in this case by arranging in ascending order we get

1 2 3 4 5 6 9

Step 2: First quarter i = (7+1)/4 = 2

So, first quarter is the average of second and third values

Q1 = (2+3)/2 = 2.5

This implies that one fourth data would be less than 2.5

Step 3: for third quarter i = 3(7+1)/4 = 7.5

So third quarter is the average of seventh and eight values

Q3 = (5+6)/2 = 6.5

This implies that three-fourth of data is less than 6.5

Interquartile range is = third quarter – first quarter

= Q3 – Q1

= 6.5 – 2.5

= 4.0

This implies that 50% of data spans a range of 4.0

8.4.2 Self-check questions

a) 25^{th} percentile is equal to _____.
(Ist quartile / 25 the quartile / 24 the quartile / 2nd quartile)

b) State true or false –

Upper quartile is the lowest value of top 25% of items.

8.4.3 Mean Absolute Deviation (MAD), Variance and Standard Deviation

The calculation of these three measures of dispersion involves similar process. For example, for the data set 5 9 16 17 18, the sum of deviation from the mean (in this case 13) would be: (5-13) + (9-13) + (16-13) + (17-13) + (18-13) = 0

Thus, total deviation of the data set is zero. So, does it mean that data does not show any variability? But observation of data set shows that values differ from each other. This problem is rectified by using Mean absolute deviation or variance method.

In case of ***MAD*** take ignore the negative sign and consider all the values as positive and then find its average

Thus, MAD = sum of absolute value of difference between values and mean / no. of observations

For the above data set: MAD = (8 + 4 + 3 + 4 + 5) / 5

= 4.8

Variance involves squaring the deviation values and then finding its average

Variance = sum of square of deviation values / no. of observations

= (64 + 16 + 9 + 16 + 25) / 5

= 26

Standard deviation is the square root of variance. It has a lot of importance in conjunction with advanced analysis like inferential statistics. Almost no analysis complete without the usage of standard deviation concept.

Thus, Standard deviation = square root of variance

= square root of (26)

= 5.1

One of the important applications of standard deviation is to infer that a given data is normally distributed or not, as this information is of prime importance in application of parametric statistics. For data to be normally distributed 68% of it should have a deviation of one from mean on both sides of the curve or 95% of data should have a deviation of 2 or 99.7% of data should have deviation of three. For instance, the average

price of petrol across a state is Rs.80.00 with a standard deviation of Rs.1.00. To check this sample of prices of petrol were taken from various parts of the state and is checked for its normal distribution. Now, from all the collected data 68% of it falls within limit of one standard deviation from the mean then it can be inferred that data is normally distributed. This implies that it is correct to infer that 68% of values representing price of petrol vary from Rs.79.00 to Rs.81.00 and this information can be used further for application of other statistical tools.

8.4.3 Self-check questions

a) **The square of standard deviation is _____.**
   a. Square deviation
   b. Mean square deviation
   c. Variance
   d. None of the above

b) **The numerical value of a standard deviation can never be ___**

8.4.4 Coefficient of Variation

Coefficient of variation is the ratio of standard deviation to the mean expressed in percentage.

CV = (standard deviation / mean) * (100)

It is a relative comparison of standard deviation with respect to mean. It is more useful in cases of those data sets which have different means as in the case of time series data set. For instance, five week average price of stock A is 57, 68, 64, 71, 62

Then to calculate its coefficient of variation mean = 64.40 and standard deviation = 4.84

So, CV = (4.84/64.40) * 100 = 7.5%.

Now suppose, five week price of stock B is 12, 17, 8 15 13

For its CV mean = 13 and standard deviation = 3.03

Thus, CV = (3.03/13) * 100 = 23.3%

By looking at only standard deviation values, stock A is riskier than B. But average price of stock A is also high which is almost three times higher than stock B. so, it would be erroneous to interpret only on the basis of standard deviation. Coefficient of variation reveals the risk of stock in terms of size of standard deviation relative to size of average price in percentage terms. In case of stock A a standard deviation of 4.84 with respect to average price of 64.40 looks to be small whereas a standard deviation with a size of 3.03 with respect to 13 looks to be high. This is indicated by calculation of coefficient of variation. Thus, stock B is riskier than A.

Example 1: A sample of 12 accounting firms reveals the following number of professional per office:

7 10 9 14 11 8 5 12 8 3 13 6

(i) determine the mean absolute deviation, variance and standard deviation

(ii) determine the interquartile range

Solution:

(i)

Data values	Deviation from the mean	Absolute value of deviation from mean	Square values
7	7-8.83 = -1.83	1.83	3.34
10	10-8.83 = 1.17	1.17	1.36
9	0.17	0.17	0.02
14	5.17	5.17	26.72
11	2.17	2.17	4.70
8	-0.83	0.83	0.68
5	-3.83	3.83	14.66
12	3.17	3.17	10.04
8	-0.83	0.83	0.68
3	-5.83	5.863	33.98
13	4.17	4.17	17.38
6	-2.83	2.83	8.01
Mean = 8.83		***MAD = 2.66***	***Variance = 11.06***

Standard deviation = square root of variance

= square root of (11.06)

. = 3.32

(ii) Interquartile range

1. Arrange the data in an order: either ascending or descending. In the above example the data is already set in an ascending order.

Data set:

3 5 6 7 8 8 9 10 11 12 13 14

For Q1 i = (12 + 1) / 4

= 3.25

Q1 = (6+7)/2

= 6.5

For Q3 i = 3(12+1)/4

= 9.75

Q3 = (11 + 12) / 2

= 11.5

Thus inter quartile range = Q3-Q1

= 11.5 – 6.5 = **5.0**

8.4.4 Self-check questions

a) Determine the first, second and third quartiles of the following data:

10.5 14.7 15.3 17.7 15.9 12.2 10.0 14.1 13.9 18.5

13.9 15.1 14.7

b) Compute the interquartile for the following data:

5 8 14 6 21 11 9 10 18 2

c) Find MAD, variance and standard deviation of following data:

4 5 3 6 5 6 5 6

8.5 Measures of Dispersion for Grouped Data

As discussed grouped data is the organized data which has been organized into frequency distributions. Two measures of dispersion for grouped data that have been discussed in this chapter are: standard deviation and variance. Standard deviation is the square root of variance.

The formula for variance is:

$\sigma^2 = \sum f(m-\mu)^2 / n$

or $\sigma^2 = [\sum fm^2 - \{(\sum fm)^2/n\}] / n$

where f = frequency

m = class midpoint

n = total frequencies of population

μ = grouped mean for the population

Example 2:

Compute the mean, mode, variance and standard deviation of the following data:

Class Interval	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
Frequency	6	22	35	29	16	8	4	2

Solution:

Class Interval	**Frequency, f**	**Mid-point, m**	**fm**	**m – mean**	**(m-mean)²**	**f*(m-mean)²**	**fm²**
10-15	6	12.5	75	-13.16	173.19	1039.14	937.50
15-20	22	17.5	385	-8.16	66.59	1464.98	6737.50
20-25	35	22.5	787	-3.16	9.99	349.65	17718.75
25-30	29	27.5	797.5	1.84	3.39	98.31	21931.25
30-35	16	32.5	520	6.84	46.79	748.64	16900
35-40	8	37.5	300	11.84	140.19	1121.52	11250
40-45	4	42.5	170	16.84	283.59	1134.36	7225
45-50	2	47.5	95	21.84	476.99	953.98	4512
	$\sum f = 122 = n$		$\sum fm =$ 3130			Sum = 6910.58	Sum = 87212.50

Mean $= \sum fm / \sum f$

$= 3130 / 12 = 25.66$

Variance $= \sum f(m-\mu)^2 / n$

$= 6910.58 / 122 = 57.11$

Standard deviation = square root of (57.11)

= 7.56

By using other formula

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

variance = $\{87212.5 - (3130)^2/122\} / 122$

= 57.11

Standard deviation = square root of (57.11)

= 7.56

8.5 Self-check questions

a) Determine the variance and standard deviation for the following data

Class	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	7	11	18	13	6	4

8.6 GLOSSARY

- Dispersion : The degree to which the numerical data tends to spread about an average value is called the variation of dispersion of the data
- Range : It is defined as the difference between the value of the smallest item and the value of the largest item included in the distribution.
- Inter-Quartile Range : The difference between the upper quartile and the lower quartile is referred to as Inter-Quartile Range
- Mean Deviation : (also called Average Deviation) is defined as the Arithmetic Mean of the absolute deviation of all the values from their Median (or Mean or Mode).
- Standard Deviation : is the square root of the arithmetic mean of the squares of all the deviations, the deviations being measured from the arithmetic mean of the observations
- Coefficient of Variation : The ratio of Standard Deviation to Arithmetic Mean, expressed in percentage is called as Coefficient of Variation.

8.7 SHORT ANSWER QUESTIONS

1. Why is dispersion important?
2. What are the different types of dispersion?
3. What is the difference between range, mean deviation, and variance?
4. What is the coefficient of variation?

8.8 LONG ANSWER QUESTIONS

1. Discuss the advantages and disadvantages of different measures of dispersion.
2. How can dispersion be used to compare the spread of two datasets?
3. How can dispersion be used to identify outliers in a dataset?

8.9 ANSWERS OF SELF CHECK QUESTIONS

8.3 a) D

b) D

8.4.1 a) 30

b) A

8.4.2 a) Answer – A 1st Quartile.

b) True

8.4.3 a) Answer: c

b) Answer: negative

8.4.4 a) 13.05, 14.7, 15.6

b) IQR = 9.25

c) Variance = 1.14

8.5 a) 185.69, 13.62

8.10 SUGGESTED READINGS

- Gupta, S. P.; Statistical Methods, Sultan Chand & Sons, New Delhi, 2007.
- Srivastava, T. N. & Rego, Shailaja; Statistics for Management, Tata McGraw Hill, New Delhi, 2007.
- Aggarwal, B. M..; Business Statistics, Sultan Chand & Sons, New Delhi, 2007.
- Sharma, J. K.; Business Statistics, Pearson Education, New Delhi, 2004.

MBA-Distance Education (First year)

Semester-2

Lesson No. 9 **AUTHOR : CHETNA SHARMA**

SKEWNESS AND KURTOSIS : CONCEPTS AND MEASURES

STRUCTURE

1. OBJECTIVES

After going through this lesson, the reader will be able to :

1. Distinguish between a symmetrical and a skewed distribution;
2. Compute various coefficients to measure the extent of skewness in a Distribution;
3. Distinguish between platykurtic, mesokurtic and leptokurtic distributions; and

Compute the coefficient of kurtosis.

9.1 INTRODUCTION

Ir. this Unit you will learn various techniques to distinguish between various shapes of a frequency distribution. This is the final Unit with regard to the summarization of univariate data. This Unit will make you familiar with the concept of skewness and kurtosis. The need to study these concepts arises from the fact that the measures of central tendency and dispersion fail to describe a distribution completely. It is possible to have frequency distributions which differ widely in their nature and composition and yet may have same central tendency and dispersion.

Thus there is need to supplement the measures of central tendency and dispersion. Consequently, in this lesson, we shall discuss two such measures, viz, measures of skewness and kurtosis.

9.2 CONCEPT OF SKEWNESS

The skewness of a distribution is defined as the lack of symmetry. In a symmetrical distribution, the Mean, Median and Mode are equal to each other and the ordinate at mean divides the distribution into two equal parts such that one part is mirror image of the other (Fig. 9.1). If some observations, of very high (low) magnitude, are added to such a distribution, its right (1eft) tail gets elongated.

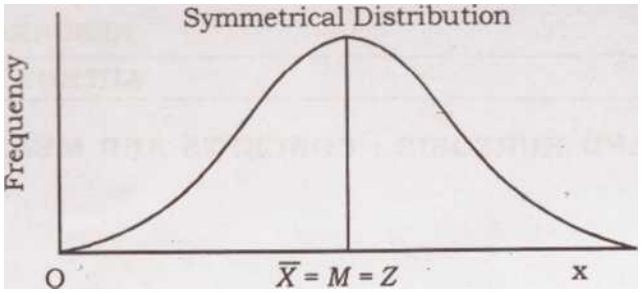


Figure 9.1

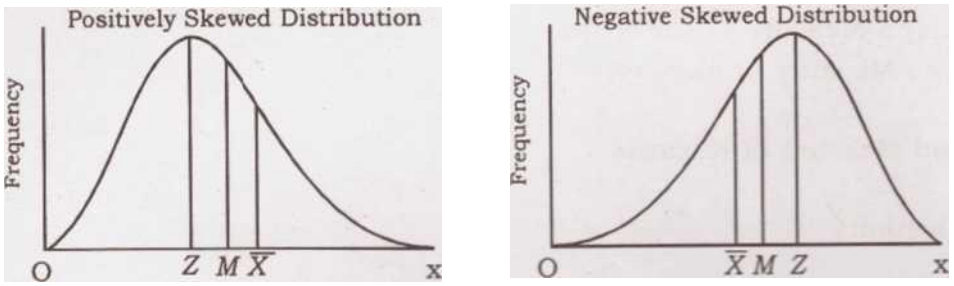


Figure 9.2

These observations are also known as extreme observations. The presence of extreme observations on the right hand side of a distribution makes it positively skewed and the three averages, viz., mean, median and mode, will no longer be equal. We shall in fact have Mean > Median > Mode when a distribution is positively skewed. On the other hand, the presence of extreme observations to the left hand side of a distribution make it negatively skewed and the relationship between mean, median and mode is: Mean < Median < Mode. In Fig. 9.2 we depict the shapes of positively skewed and negatively skewed distributions.

The direction and extent of skewness can be measured in various ways. We shall discuss four measures of skewness in this lesson.

9.2.1 KARL PEARSON'S MEASURE OF SKEWNESS

In Fig. 9.2 you noticed that the mean, median and mode are not equal in a skewed distribution. The Karl Pearson's measure of skewness is based upon the divergence of mean from mob in a skewed distribution.

Since Mean = Mode in a symmetrical distribution, (Mean - Mode) can be taken as an absolute measure of skewness. The absolute measure of skewness for a distribution depends upon the unit of measurement. For example, if the mean =2.45 meter and mode = 2.14 meter, then absolute measure of skewness will be .2.45 meter - 2.14 meter = 0.31 meter. For the same distribution, if we change the unit of measurement to centimeters, the absolute measure of skewness is 245 centimeter - 214 centimeter = 31 centimeter. In order to avoid such a problem Karl Pearson takes a relative measure of skewness.

A relative measure, independent of the units of measurement, is defined as the Karl Pearson's Coefficient of Skewness (S_k) given by _ Mean-Mode

^{k} *Zd.*

The sign of Sk gives the direction and its magnitude gives the extent of skewness. If Sk

> 0, the distribution is positively skewed, and if Sk < 0 it is negatively skewed. So far we have seen that S_t is strategically dependent upon mode. If mode is not defined for a distribution we cannot find Sk. But empirical relation between mean, median and mode states that, for a moderately symmetrical distribution, we have

Mean - Mode = 3 (Mean - Median)

Hence. Karl Pearson s coefficient of skewness is defined in terms of median as :

„ 3(Mean-Median) s d

Example 9.1 : Compute the Karl Pearson's coefficient of skewness from the following data :

Heights (in inches)	Number of Persons
58	10
59	18
60	30
61	42
62	35
63	28
64	16
65	8

Table for the computation of Mean and S.D.

Height (X)	u = X - 61	No. of Persons [J]	/«	/«3
58	-3	10	-30	90
59	-2	18	-36	72
60	-1	30	-30	30
61	0	42	0	0
62	1	35	35	35
63	2	28	56	112
64	3	16	48	144
65	4	8	32	128
Total		187	75	611

Mean = 61+ -^- = 61.4 187

$$\sqrt{\frac{611}{187}-\left(\frac{75}{187}\right)^{2}}$$

To find mode, we note that height is a continuous variable. It is assumed that the height has been measured under the approximation that a measurement on height that is, e.g., greater than 58 but less than 58.5 is taken as 58 inches while a measurement greater than or equal to 58.5 but less than 59 is taken as 59 inches.

Thus the given data can be written as :

Heights (in inches)	Number of Persons
57.5 - 58.5	10
58.5 - 59.5	18
59.5 - 60.5	30
60.5 - 61.5	42
61.5 - 62.5	35
62.5 - 63.5	28
63.5 - 64.5	16
64.5 - 65.5	8

By inspection, the modal class is 60.5 - 61.5. Thus, we have l_m = 60.5, A_1=42-30=12, A_2=42-35=7 and h= 1.

Mode = 60.5 + ^^*1=61.13

Hence, the Karl Pearson's coefficient of skewness :

= 61.4-61.13 k 1.76

Thus, the distribution is positively skewed.

9.2.2 BOWLEY'S MEASURE OF SKEWNESS

This measure is based on quartiles. For a symmetrical distribution, it is seen that Q,, and Q_3 are equidistant from median. Thus, (Q_3 - Md) - (Md - Q,) can be taken as an absolute measure of skewness.

A relative measure of skewness, known as Bowley's coefficient (SQ) is given below :

$$\frac{(Q_3 - M_d) - (M_d - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2M_d + Q_1}{Q_3 - Q_1}$$

9.2.3 KELLY'S MEASURE OF SKEWNESS

Bowley's measure of skewness is based on the middle 50% of the observations because it leaves 25% of the observations on each extreme of the distribution. As an improvement over Bowley's measure, Kelly has suggested a measure based on P_{10} and, P_{90} so that only 10% of the observations on each extreme are ignored.

Kelly's coefficient of skewness, denoted by Sp is given below :

$$S_p = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

Note that P^ = Md (median). It may be noted here that although the coefficient S_K, S_Q and S_p are not comparable, however, in the absence of skewness, each of them will be equal to zero.

9.3 MOMENTS

The r,h moment about mean of a distribution, denoted by (a_r is given by:

H-= - I-£/ (*,-*)'
' N i=1

Thus, r,h moment about mean is the mean of the r lh power of deviations of observations from their arithmetic mean. In particular,

if r= 0, we have n = u N i=i -*)° = 1,

if r = 1, we have (a1 = /(*, - *)' =0,

1

if r = 2, we have |i, = — 'L/,(Xl -X) =c2,
N 1=1

1 n — 3

if r = 3, we have u, = —'I fl(Xl -X) and so on.
J N i=i

These moments are also known as central moments.

In addition to the above, we can define raw moments as moments about any arbitrary mean. Let A denote an arbitrary mean, then uth moment about A is defined as

r= 0, 1,2, 3, ...
N i=i

When A = 0, we get various moments about origin.

9.3.1 MOMENT MEASURE OF SKEWNESS

The moment measure of skewness is based on the property that, for a symmetrical distribution, all odd ordered central moments are equal to zero. We note that n, = 0, for every distribution, therefore, the lowest order moment that can provide an absolute measure of skewness is M_3- Further, a coefficient of skewness, independent of the units of measurement,
is given by :

a, = ^4- = ±7p7 = Yi, where p. and y
G

are defined as the first beta and first gamma coefficients respectively.

9.4 CONCEPT AND MEASURE OF KURTOSIS

Kurtosis is another measure of the shape of a distribution. Whereas skewness measures the lack of symmetry of the frequency curve of a distribution, kurtosis is a measure of the relative peakedness of its frequency curve. Various frequency curves can be divided into three categories depending upon the shape of their peak. The three shapes are termed as Leptokurtic, Mesokurtic and Platykurtic as shown in Fig. 9.3.

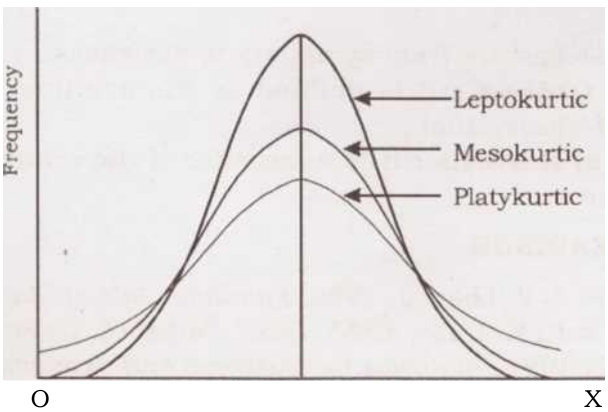


Figure 9.3

A measure of kurtosis is given by : $P2 - 2 \frac{Mi}{r2}$

This is a coefficient given by Karl Pearson. The value of $P_2 = 3$ for a mesokurtic cum. When $P_2 > 3$, the curve is more peaked than the mesokurtic curve and is termed as leptokurtic. Similarly, when $p_2 < 3$, the curve is less peaked than the mesokurtic curve and is called as platykurtic curve.

9.4 Self- check questions

a) A distribution with a kurtosis less than 3 is termed as___.

b) What is the kurtosis of a normal distribution?

c) What is the skewness of a normal distribution?

d) If the distribution is negatively skewed, then the mean is less than the mode

9.5 SUMMARY

In this lesson you have learned about the measures of skewness and kurtosis. These two concepts are used to get an idea about the shape of the frequency curve of a distribution. Skewness is a measure of the lack of symmetry whereas kurtosis is a measure of the relative peakedness of the top of a frequency curve.

9.6 GLOSSARY

- Skewness : Departure from symmetry is skewness.
- Moment of Order r : It is defined as the arithmetic mean of the rth power of deviations of observations.
- Coefficient of Kurtosis : It is a measure of the relative peakedness of the top of a frequency curve.

9.7 SHORT ANSWER QUESTIONS

1. Compute the Karl Pearson s coefficient of skewness from the following data :

Daily Income (Rs.) :	0-20	20-40	40-60	60-80	80-100
No. of Families :	13	25	27	19	16

2. The following measures were computed for a frequency distribution :
   Mean = 50, coefficient of Variation = 35% and Karl Pearson's Coefficient of Skewness = - 0.25. Compute Standard Deviation, Mode and Median of the distribution.
3. Compute the Moment coefficient of skewness from the following data :

Marks Obtained :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency :	6	12	22	24	16	12	8

9.8 LONG ANSWER QUESTIONS

1 The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Examine the skewness and kurtosis of the distribution.

2 Compute the first four central moments from the following data. Also find the two beta coefficients.

Value :	5	10	15	20	25	30	35
Frequency :	8	15	20	32	23	17	5

3 The first four moments of a distribution are 1,4, 10 and 46 respectively. Compute the moment coefficients of skewness and kurtosis and comment upon the nature of the distribution.

9.9 ANSWERS OF SELF CHECK QUESTIONS

9.4 a) Answer – Platykurtic

b) Answer – 3

c) Answer – 0

d) Answer – true.

9.10 SUGGESTED READINGS

- Elhance, D. N. & V. Lhance, 1988, Fundamentals of Statistics, Kitab Mahal, Allahabad.
- Nagar, A. L. & R. K. Dass, 1983, Basic Statistics, Oxford University Press, Delhi .
- *Mansfield, E., 1991, Statistics for Business and Economics : Methods and Applications, W.W. Norton and Co.*
- Yule, G U. & M. G Kendall, 1991, An Introduction to the Theory of Statistics, Universal Books, Delhi.

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS &RESEARCH METHODOLGY: BSRM 201

Lesson No. 10 **AUTHOR: KAJAL**

CORRELATION ANALYSIS

Structure

10.0 Objective

To understand

- the techniques of evaluating degree and direction of relationship between two variables.
- the cause and effect linkage.
- the regression line and how to use it in forecasting.
- the goodness of fit of regression model by understanding the meaning of standard error of estimate and coefficient of determination.

10.1 Introduction

Correlation is degree of association between two variables. It indicates the extent to which the variation in the scores on one variable results in a corresponding variation in the scores on the second variable. For example if correlation analysis indicates that marks in Graduate Management Aptitude Test (GMAT) are

positively associated with performance in MBA then higher the marks in GMAT higher marks will be scored in MBA.

In this chapter we shall be considering only linear relationship between two variables implying that if two variables are correlated and scores of these variables are plotted on the graph then they will follow a straight line. Such a line is called as *regression line*. A strong correlation indicates that there is only a small amount of error and most of the points lie close to the regression line; a weak correlation indicates that there is a lot of error and the points are more scattered. In the second case it could be concluded that a linear relationship is not a good model for the considered data.

10.1 Self-check questions

a) What is a simple correlation?

10.2 Measurement of relationship

The importance of correlation being an important measure of finding relationship between two variables is signified by its additional appropriateness than covariance.

The simplest and easiest way to deduce whether two variables are associated is to find whether if one variable deviates from its mean in a particular direction then does other variable also deviates in a similar fashion. For instance if one variable is X and its mean is X-bar and other variable is Y and its mean is Y-bar then their deviation from respective means is denoted by (X-X bar) and (Y-Y bar). To understand what these deviations infer lets study the following example. The data in table 1 represents five observations of sales and number of advertisements watched.

Table 1

	1	**2**	**3**	**4**	**5**	**Mean**	**S.D.**
No. of advertisements watched (x)	5	4	4	6	8	5.4	1.67
Sales (y)	8	9	10	13	15	11.0	2.92

Covariance (x,y) = ∑(x-x bar)*(y-y bar)/n-1

By applying this formula for the above data covariance = 4.25

A positive covariance indicates that deviation of both variables is in the same direction and if one increases then other variable also increases and vice-versa. On the other hand, a negative covariance indicates that if one variable increases then other decreases. But one problem with covariance as a measure of relationship between two variables is that it is not a standardized measure i.e. value changes with change in scale of measurement. For instance, if advertisement expenditure is measured in terms of dollars rather than in rupees then covariance changes.

To rectify this anomaly of lack of standardization, correlation is used. Covariance measure is standardized by dividing the deviations of two variables with their standard deviations. Thus, standardized covariance is known as correlation coefficient (r).

r = covariance (x,y) / {S.D.(x) * S.D.(y)}

$$= \frac{\sum(x\text{-}x\ bar)*(y\text{-}y\ bar)}{(n-1) * S.D.(x) * S.D.(y)} \qquad = \frac{\sum(x\text{-}x\ bar) * (y\text{-}y\ bar)}{\sqrt{\sum(x\text{-}x\ bar)^2 * (y\text{-}y\ bar)^2}}$$

By using this formula r can be calculated as

r = 4.25/1.67 * 2.92 = 0.87.

By standardizing coefficient value has been limited between -1 and +1. Thus, positive sign indicates movement of both variables in the same direction and high magnitude of r indicates high correlation between the variables.

10.2 Self- check question

1) wo variables have a correlation coefficient of -0.5. What does this mean?

a) The variables are perfectly negatively correlated.

b) The variables are not correlated.

c) The variables are positively correlated but with a weak relationship.

d) The variables are negatively correlated with a strong relationship.

10.3 Types of correlation

Statistically the correlation value ranges between -1 indicating a perfect negative correlation, and +1 indicating a perfect positive correlation. A value of zero indicates no correlation at all. For example, distance travelled by a car is negatively related with petrol remaining in the tank. Productivity is positively related with experience and amount of time spent on work. Whereas, as described in the example above age of mother and daughter has no correlation with foot size.

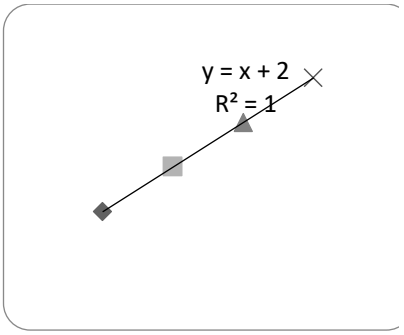
Figure 1 depicts these three different types of correlation by using scatter diagrams. A scatter diagram is simply a graph that plot scores of one variable with the scores of another variable. A scatter diagram tells several things about the data such as whether there exists a relationship between the variables, what kind of relationship it is and whether any cases are distinctly different from the others.

(i) Perfectly positive correlation: A correlation of +1 between two variables indicate a perfect association which implies that if one variable shows increasing or decreasing score then other variable also moves in the same direction.

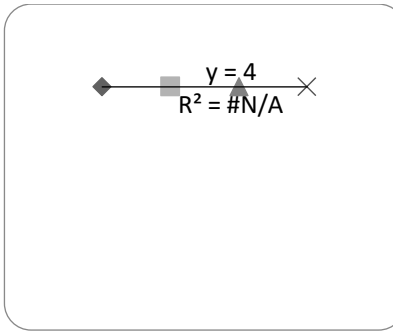
(ii) No correlation: A correlation of 0 between two variables indicate no association.

(iii) Perfectly negative correlation: A correlation of -1 between two variables indicate a perfect association which implies that if one variable shows increasing or decreasing score then other variable moves in the opposite direction.

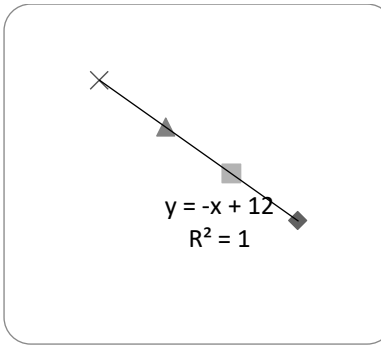
Figure 1



Perfect positive correlation r=+1



No correlation r=0



Perfect negative correlation r=-1

10.4 Pearson coefficient of correlation

The coefficient in equation

r = covariance (x,y) / {S.D.(x) * S.D.(y)}

$$= \frac{\sum(x-\text{x bar})*(y-\text{y bar})}{(n-1) * S.D.(x) * S.D.(y)} \qquad = \frac{\sum(x-\text{x bar}) * (y-\text{y bar})}{\sqrt{\sum(x-\text{x bar})^2 * (y-\text{y bar})^2}}$$

is known as Pearson coefficient of correlation. Pearson's correlation relies on a number of assumptions.

- The relationship between the variables is linear.
- The points are evenly distributed along the straight line. This is the assumption of homoscedasticity. If the data has the points unevenly spread along the proposed straight line then the Pearson correlation is not an accurate measure of the association.
- The data are drawn from normally distributed populations.
- The data collected must be interval or ratio.

Example 1: It has been proposed that students who spent maximum time on studying statistics would achieve highest marks. Data for ten students was collected regarding time spent on studying and marks obtained as shown in table 2.

Table 2

Time spent (hrs. per week)	Marks obtained (out of 100)	(x-x bar)	(x-x bar)²	(y-y bar)	(y-y bar)²	(x-x bar) * (y-y bar)
40	58	11	121	2	4	22
43	73	14	196	17	289	238
18	56	-11	121	0	0	0
10	47	-19	361	-9	81	171
25	58	-4	16	2	4	-8
33	54	4	16	-2	4	-8
27	45	-2	4	-11	121	22
17	32	-12	144	-24	576	288
30	68	1	1	12	144	12
47	69	18	324	13	169	234
x bar=29	**y bar=56**		**1304**		**1392**	**971**

r = 971 / √1304*1392

= 0.72

These results indicate that as study time increases, statistics exam performance also increases, which is a positive correlation.

10.4 Self-check question

1) What does a Pearson's correlation coefficient of 0.8 indicate?

a) A strong positive correlation.

b) A weak positive correlation.

c) A strong negative correlation.

d) A weak negative correlation.

2) What is Pearson's correlation coefficient?

a) A measure of the strength and direction of the linear relationship between two variables.

b) A measure of the average distance between data points and the mean.

c) A measure of the probability that two events will occur together.

d) A measure of the spread of a dataset.

10.5 Spearman Rank correlation

In the previous section Pearson correlation was discussed as a technique to measure degree of association between two variables. The primary condition for applying Pearson correlation was the presence of interval data. But when only ordinal data is available then Spearman rank correlation should be applied to achieve the same objective.

The formula for Spearman rank correlation is derived from Pearson correlation coefficient and is as follows:

$r = 1 - \{6\sum d^2/n(n^2 - 1)\}$

where n = number of pairs being correlated

d = difference in rank of each pair

The process of calculating Spearman rank correlation begins by assigning ranks to the ordinal data of two variables under consideration. For one variable the highest value is given rank 1, the next highest is given rank 2 and so on. Similar process is adopted for ranking of second variable. Importantly, the difference in ranks 'd' is found by subtracting the rank of a member of one group from the rank of its corresponding member of the other group. The differences are then squared and summed.

Spearman rank correlation is interpreted similarly as Pearson correlation. Positive correlations indicate that high value of one variable is associated with high value of other variable and negative correlations indicate that high value of one variable is associated with low value of other variable.

Example 2: Is there a correlation between distance travelled by a salesperson and sales achieved. In table 3 sales achieved by nine salesperson in rupees and distance travelled in kms to cover a particular territory is given. Assuming data to be ordinal determine correlation coefficient.

Table 3

Sales (Rs. '000)	Distance (kms)	Rank of sales	Rank of distance	d	d^2
150	1500	9	9	0	0
210	2100	8	8	0	0
285	3200	7	3	4	16
301	2400	6	6	0	0
335	2200	5	7	-2	4
390	2500	4	5	-1	1
400	3300	3	2	1	1
425	3100	2	4	-2	4
440	3600	1	1	0	0
					$\sum d^2 = 26$

Spearman rank coefficient:

$$r = 1 - \{6\sum d^2/n(n^2 - 1)\}$$

$$r = 1 - \{6 * 26/9(81-1)\}$$

$$r = 0.783$$

Positive r indicates sales achieved increases if a salesperson covers more distance per territory. It is not a perfect correlation which implies that other factors than distance travelled plays role in determining sales.

10.6 SUMMARY

Correlation is a statistical device which helps us in analysing the covariation of 2 or more variables. Though it helps us in determining the degree of relationship between 2 or more variables, it does not tell us anything about cause and effect relationship. Even a high degree of correlation does not necessarily mean causational or functional relationship, though the existence of causation always implies correlation. The correlation can be due to purely by chance and both the correlated variables may be influenced by one or more other variables. Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect.

10.7 Glossary

- **Correlation coefficient:** is a statistic given by Karl Pearson to measure the linear relationship between two variables. its value ranges from -1 to +1 where -1 indicates perfect negative correlation, +1 indicates perfect positive correlation and value of zero implies no relationship between two variables.
- **Spearman's rank correlation:** is a measure of correlation between two ordinal variables.

10.8 SHORT ANSWER QUESTIONS

1. *Explain assumptions of Pearson correlation.*
2. *Determine value of r for the following data*

X	*4*	*6*	*7*	*11*	*14*	*17*	*21*
Y	*18*	*12*	*13*	*8*	*7*	*7*	*4*

3. *What is the difference between interval and ordinal data? Explain by giving example.*

10.9 LONG ANSWER QUESTIONS

1. Explain the difference between Pearson's correlation coefficient and Spearman's rank correlation coefficient.
2. Discuss the advantages and disadvantages of using correlation
3. *Calculate Spearman rank correlation for following data*

X	*4*	*5*	*8*	*11*	*10*	*7*	*3*	*1*
Y	*6*	*8*	*7*	*10*	*9*	*5*	*2*	*3*

10.10 ANSWERS OF SELF CHECK QUESTIONS

10.1 a) Answer - A simple correlation implies the study of a relationship between only two variables.

10.2 d) The variables are negatively correlated with a strong relationship.

10.4 1) Answer: a) A strong positive correlation.

2) Answer: a) A measure of the strength and direction of the linear relationship between two variables.

10.11 SUGGESTED READINGS

1. S.P.Gupta : Statistical Methods.
2. S.C.Gupta & V.K.Kapoor : Fundamentals of Mathematical Statistics.

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS &RESEARCH METHODOLGY: BSRM 201

Lesson No. 11 **AUTHOR: KAJAL**

REGRESSION ANALYSIS

Structure

11.0 OBJECTIVES

After reading this lesson, the student should be able to

- Understand the concept and use of Regression.
- Formulate Regressional Equations.

11.1 INTRODUCTION

The statistical technique of estimating the unknown value of dependent variable from the known values of an independent variable is called regression analysis. Galton introduced the concept of regression in 1877 where he studied the case of one thousand fathers and sons and concluded that the tall fathers tend to have tall sons and short fathers have short sons, but the average height of the sons of a group of tall fathers is less than that of the fathers and the average height of the sons of a group of short fathers is greater than that of the fathers.

The straight line or the line of best fit with the help of an equation

$$Y = a + bx$$

Here x is an independent variable whereas Y is dependent variable.

11.2 Regression: *Does money lead to happiness?*

Correlation analysis as discussed in previous section indicated the strength of relationship between two variables. But the technique did not tell the direction of relationship. For instance,

- Sales volume is associated with advertisement expenditure. But does increase in advertisement resulted in increase in sales or increase in sales motivated the company to increase expenditure on advertisement.
- Hotel occupancy is related with tourist flow. But does increase in hotels lead to high tourist inflow or more tourists motivates building of more hotels.
- Performance in a job is related with salary. But does increase in salary results in better performance or vice versa?

Whether two variables are associated or not can be evaluated by using correlation analysis but it does not indicate which variable is the cause and which is the effect? The causal relationship can be evaluated by using regression analysis.

Regression analysis is the process of building a model involving two variables that can be used to predict one variable by another variable. The most elementary regression model is called simple regression in which the variable to be predicted is called the dependent variable and variable which predicts is called independent variable. The dependent variable is denoted by 'y' and independent variable is denoted by 'x'. The model is denoted by following equation.

Dependent or predicted variable (y) = intercept (a) + slope (b) * independent or predictor variable (x) .....(1)

11.3 Regression Equation and Coefficients

The first step in regression line is to find whether two variables are associated or not. This can be done by drawing scatter diagram as discussed in previous chapter of correlation. Scatter diagram is the graphical representation of data of two variables. All the data points might not fall in a straight line. For example table 4 shows data regarding advertising expenditure and corresponding sales. The questions are:

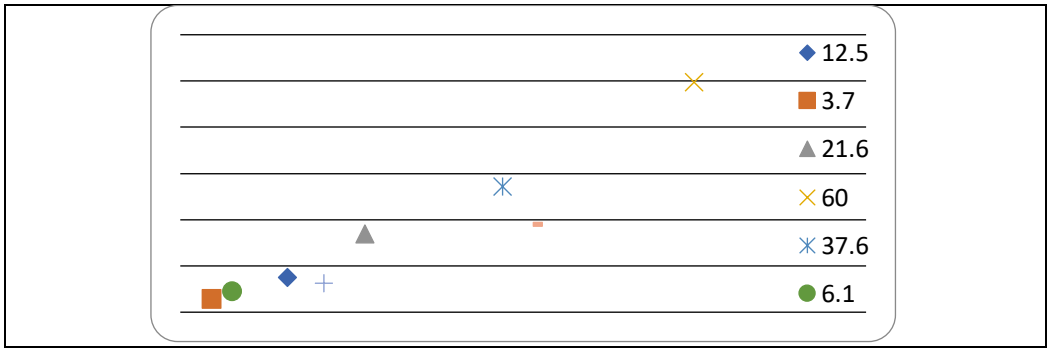
(i) Is there a relationship between two variables?

(ii) Can one variable be used to predict another variable?

Table 4

Advertising expenditure (millions)	12.5	3.7	21.6	60	37.6	6.1	16.8	41.2
Sales	148	55	338	994	541	89	126	379

Fig.2



The relationship between two variables can be deduced by drawing the scatter diagram (Fig.1). It is not a perfect relationship as data points are scattered. But diagram shows a positive relationship between sales and advertisement expenditure indicating that with increase in one variable other variable also increases. But it is important to note here that scatter diagram does not tell causality i.e. whether advertisement expenditure results in sales or is it vice-versa. To understand causality following literature which includes model building and its accuracy has to be appreciated.

For accurate prediction a straight line covering maximum data points should be drawn. Now as data points are scattered so many straight lines can be drawn. This line which touches maximum points and minimizes error is called as regression line. This line as discussed above can be denoted by a model depicted by following equation:

$$y = a + bx$$

where y = dependent variable

a = intercept

b = slope intercept and

x = independent variable.

For instance in the above example if sales is considered as dependent and advertisement expenditure as independent variable then the model would become as:

$$\text{Sales} = a + b * (\text{advertisement expenditure})$$

This type of model is called as deterministic model that produces exact output for a given input. But sales are not a function of advertisement expenditure only. There can be other factors also like distribution, sales people motivation etc. which have an impact on sales. So a more proper model can be as:

$$\text{Sales} = a + b * (\text{advertisement expenditure}) + \text{error}$$

where error denotes other factors than advertisement expenditure which impact sales. This kind of model is called as probabilistic model.

To determine the regression equation values of ‘a’ and ‘b’ has to be determined. This process is termed as least square analysis. This method is used to develop a regression model by producing the minimum sum of squared error values. The error will be minimized if regression line passes through maximum data points. Thus least squares method is used to find best fit line. A best fit line is one which minimizes the amount of difference between observed data and the line. The data points which does not fall on the line lie either above or below

the line. The distance between these points and line is called residual. For some points this residual is positive and for some negative. To avoid cancelling out each other these differences are squared and added. This sum would be least for the best fit line. That's why this method is called as least squares a method of drawing the line. Thus, in this process entire aim would be calculate values of 'a' and 'b'.

Slope intercept 'b' can be calculated by using the following equation:

$b = \sum(x - \text{x bar})(y - \text{y bar})/\sum(x - x^2)$

y intercept of the regression line can be calculated as:

a = y bar – b * x bar

The regression line depending on which variable is dependent and which variable is independent can give two different results. The following example indicates the importance of establishing causality between two variables.

Example 3:

(a) If sales is dependent variable (y) and advertisement expenditure is independent variable (x) then compute regression equation and also predict sales if advertisement expenditure is 50 millions:

Table 5

Advertisement expenditure (X)	Sales (Y)	x – x bar	y - y bar	(x –x bar)*(y – y bar)	(x – x bar)²
12.5	148	-12.4375	-185.375	2305.602	154.6914
3.7	55	-21.2375	-278.375	5911.989	451.0314
21.6	338	-3.3375	4.625	-15.4359	11.13891
60	994	35.0625	660.625	23163.16	1229.379
37.6	541	12.6625	207.625	2629.052	160.3389
6.1	89	-18.8375	-244.375	4603.414	354.8514
16.8	126	-8.1375	-210.375	1711.927	66.21891
41.2	379	16.2625	45.625	741.9766	264.4689
X bar = 24.93	**Y bar = 333.37**			**Sum = 41051.69**	**Sum = 2692.11**

b = 41.051/2692.11 = 15.24

a = 333.37 – 15.24*24.93 = -46.89

then regression equation becomes y = -46.89 + 15.24 * x

The slope 'b' of the regression equation implies that with every one unit increase in x i.e. advertisement expenditure sales would increase by 15.24.

If expenditure on advertisement is increased to 50 million then sales would be:

y = -46.89 + 15.24 * 50

= 715.61

i.e. sales would be 715 units with increase in advertisement expenditure to 50 million.

(b) If sales is independent variable (x) and advertisement expenditure is dependent variable (y) then compute regression equation and predict expenditure on advertisement if sales were 350 units.

Table 6

Advertisement expenditure (y)	Sales (x)	y – y bar	x - x bar	(x –x bar)*(y – y bar)	(x – x bar)^2
12.5	148	-12.4375	-185.375	2305.602	34363.89
3.7	55	-21.2375	-278.375	5911.989	77492.64
21.6	338	-3.3375	4.625	-15.4359	21.39063
60	994	35.0625	660.625	23163.16	436425.4
37.6	541	12.6625	207.625	2629.052	43108.14
6.1	89	-18.8375	-244.375	4603.414	59719.14
16.8	126	-8.1375	-210.375	1711.927	44257.64
41.2	379	16.2625	45.625	741.9766	2081.641
y bar = 24.93	**x bar = 333.37**			**Sum = 41051.69**	**Sum = 697469.9**

b = 41.051/697469.9 = 0.058

a = 24.93 – 0.058*333.37 = 5.315

then regression equation becomes y = 5.315 + 0.058 * x

The slope 'b' of the regression equation implies that with every one unit increase in x i.e. sales, advertisement expenditure would increase by 0.058.

If sales were 350 units then advertisement expenditure would be:

y = 5.315 + 0.058 * 350

= 25.61

So, by changing dependent and independent variable different regression line can be constructed which could be used for predictive purposes.

11.3 Self-check question

a) For the following data determine the equation of regression line to predict y from x

X	53	47	41	50	58	62	45	60
y	5	5	7	4	10	12	3	11

b) Slope of regression line represents …….

c) Assessment of regression line needs estimation of constant and slope intercept by using method of …….

11.4 Assessing the accuracy of model: How can it be inferred that proposed model is good?

In the previous section we discussed that an effective model between two variables can be constructed for predictive purposes by using best fit line method. But as shown in equation (4) in every model there would be an error term which should be studied before finalizing the accuracy of the model. This error denotes the difference between observed and expected values. As the analysis is always done on sampled data and decision is interpreted for entire population so a difference between sampled i.e. observed data and expected data for population is expected. In checking the accuracy of model the aim is to assess the impact and amount of error. High and significant error would ask for rectification in the model. This difference between observed 'y' values and expected 'y cap' values given by 'y – y cap' is termed as residual.

For example performance in exam of a student is dependent on the amount of time he/she spends on revising the subject. Presumably, higher the time spent on revising higher he/she will score. To assess this hypothesis a sample of 50 students is taken and data regarding their marks and amount of revision time is recorded. Regression analysis is applied and it was found that this hypothesis was good for not all the students as some students even if spending higher time were not able to improve their performance. Thus, there must be some other factors which cause exam performance. These other factors form part of unexplained variance indicated by residuals, whereas impact of revision time on exam performance is known as explained variance indicated by inclusion of independent variable. Therefore, it becomes necessary to evaluate residuals or unexplained variance to assess the accuracy of model.

The significance of explained and unexplained variance is understood by comprehending the concepts of standard error of estimate and coefficient of determination.

11.4.1 Standard Error of Estimate (s_e)

The standard error of estimate provides a single measure of error which can be used to understand the magnitude of error in the model. Thus, it is used to examine the regression error. So, the first step in estimating accuracy of model is to interpret the error indicated by residuals. As discussed in the previous section residuals denote the difference between observed (y) and expected (y cap) values. The concept is explained by furthering the sales and advertisement expenditure example. This analysis is also indicative of the fact that which of the variable is dependent and which is independent i.e. which of the two regression equations is more representative of the model. The regression equation having less standard error should provide the relationship between dependent and independent variable.

From regression equation analysis by considering sales as dependent and advertisement expenditure as independent variable standard error calculation methodology is depicted in following table.

Example 4:

Table 7

Sales (y)	Advertisement expenditure (x)	y cap	y – y cap	(y – y cap)^2
148	12.5	143.61	4.39	19.2721
55	3.7	9.498	45.502	2070.432
338	21.6	282.294	55.706	3103.158
994	60	867.51	126.49	15999.72
541	37.6	526.134	14.866	220.998
89	6.1	46.074	42.926	1842.641
126	16.8	209.142	-83.142	6912.592
379	41.2	580.998	-201.998	40803.19
y bar = 333.37	**x bar = 24.93**			**Sum= 70972.01**

The equation of the model is:

Sales (y) = a + b*(advertisement expenditure, x)

From calculations in example 1 value of regression coefficients was found to be

a = -46.89 and b = 15.24

by putting these values the equation becomes

y = -46.89 + 15.24*(x)

for each value of x calculate expected value of y i.e. y-cap. For example,

y cap = -46.89 + 15.24*12.5 = 143.61

Similarly y cap was calculated for all the observations and difference between observed (y) and expected (y cap) was calculated. These values indicate the error or residuals. As the sum of residuals would be approximately equal to zero as positive values would negate the negative error so to avoid this square of residuals is taken and sum calculated.

This total of the residuals squared column is called sum of squares of error (SSE)

Thus, SSE = $\sum(y - y\ cap)^2$ = 70972.01

Now, the question arises how to interpret SSE. A better way of interpreting SSE is through standard error of estimate denoted as s_e. The standard error of estimate is the standard deviation of error. This is calculated to put meaning to the magnitude of error as it is difficult to comprehend whether calculated SSE is high or low. To do this standard error of estimate is calculated by using following formula:

$$s_e = \text{sqrt}(SSE/n-2)$$

where n is number of observations.

For the above example s_e = sqrt(70972.01/8-2)

= 108.75

This standard error of estimate is a good measure to compare the models. The model with a lower s_e would be considered to better fit the data and a better predictor than a model with higher s_e.

11.4.2 Coefficient of determination (R square)

Coefficient of determination depicted as R square is another widely used and effective measure of accuracy of suggested model. It indicates the proportion of variability in the dependent variable (y) explained by the independent variable (x). Higher the value of R square higher is the variability explained. The range of R square is always between 0 and 1. As discussed in previous sections some of variability in the dependent variable could be explained by the introduction of independent variable and remaining goes unexplained. R square is the indicator of explained variation. Thus, higher value of R square is more acceptable. For instance analysis of coefficient of determination can indicate amount of variation in sales because of change in advertisement expenditure or variation in exam performance because of change in revision time.

Meaning and Analysis of R square:

The calculation of R square is explained by using sales and advertisement expenditure example as follows:

Step 1: Initially data was calculated regarding sales of a certain item and it was found there was variability in sales of the item.

Step 2: the researcher wanted to find the reason of sales variability and to create a model where in future sales can be predicted by studying those reasons.

Step 3: Advertisement expenditure was considered to be one of major reasons which causes change in sales volume. The data for advertisement expenditure was collected corresponding to sales and a scatter diagram was plotted. It was found from scatter diagram in addition to information given by correlation analysis that two variables were positively related.

Step 4: Now, for prediction purposes a model was formulated wherein sales were considered to be dependent on advertisement expenditure. By using this model a regression equation was devised with the help of least squares method. This method was selected to find the best fit regression equation with minimum of error.

Step 5: But entire variation in sales data was not explained by advertisement expenditure. The difference between observed and expected values of sales was termed as residuals or unexplained variation. The aim was to have more understanding of variation in dependent variable for more accurate prediction.

Step 6: The variation of dependent variable, in this case, sales data from its mean value is considered to be total variation. This variation is called as Total Sum of Squares denoted as SST and given by following equation:

SST = y – y bar

To explain this variation an independent variable, in this case, advertisement expenditure was introduced. The amount of variation explained as discussed in previous sections is because of regression line. This explained variation because of regression line is denoted by *y cap.* The difference between y cap and mean value is called as Sum of Squares of Regression denoted as SSR and given by following equation:

SSR = y cap – y bar

The remaining variation which goes unexplained is the residual or called as Sum of Squares of Error denoted as SSE and given by:

SSE = y – y cap

Thus complete equation becomes

SST = SSR + SSE

y – y bar = (y cap – y bar) + (y – y cap)

As understood by the definition, coefficient of determination is an indicator of how much of *y* is explanatory and how much of it goes unexplained. So R square represents the amount of variance in the outcome explained by the regression equation relative to how much variation was there to explain in the first place. Therefore, as a percentage it represents the percentage of the variation in the outcome that can be explained by the model:

R square = SSR/SST

or R square = (SST-SSE)/SST

= 1 – (SSE/SST)

From the above equation it is clear that value of R square can never be greater than 1.

Example 5:

Table 8

Sales (y)	Advertisement expenditure (x)	y cap	y – y cap	SSE=(y – y cap)^2	y – y bar	SST = (y – y bar)^2
148	12.5	143.61	4.39	19.2721	-185.75	34503.06
55	3.7	9.498	45.502	2070.432	-278.75	77701.56
338	21.6	282.294	55.706	3103.158	4.25	18.0625

994	60	867.51	126.49	15999.72	660.25	435930.1
541	37.6	526.134	14.866	220.998	207.25	42952.56
89	6.1	46.074	42.926	1842.641	-244.75	59902.56
126	16.8	209.142	-83.142	6912.592	-207.75	43160.06
379	41.2	580.998	-201.998	40803.19	45.25	2047.563
y bar = 333.37	**x bar = 24.93**			**Sum= 70972.01**		**Sum = 696215.5**

R square = 1 – (SSE/SST)

= 1 – (70972.01/696215.5) = 0.898

Interpretation: R square value of 0.898 indicates that 89.8% of variation in sales is explained by advertisement expenditure. This implies accuracy of suggested model. Remaining variation goes unexplained and is depicted as residuals by sum of square of errors.

11.4.2 Self- check question *(True/False)*

a) Standard error of estimate represents the standard deviation of error of regression models.

b) Standard error estimate of a model can be used to interpret the accuracy of a model in isolation.

c) Value of Coefficient of determination can be greater than one.

11.5 Correlation vs. Regression

Understanding of the two concepts facilitates us to infer the similarities and differences between correlation and regression. The major similarity between two statistical tools is that both are measures of association. Higher value of correlation coefficient (r) and higher value of coefficient of determination (R square) indicates strong relationship between two variables. But correlation analysis suffers with two shortcomings. One, it does not tell which variable is the cause and which variable is the effect. Does time spend on revising a subject leads to better performance in exam or better exam performance motivates the candidate to study and revise more? Second, correlation does not have a predictive power. It can be used to link performance with level of stress but it cannot be used to infer that with a specific amount of stress how an employee will perform?

These limitations are removed by regression analysis. The most important procedure in regression analysis is formulation of model having accurate or near accurate predictive power. The steps involved and methodology has been discussed in detail in previous sections.

Coefficient of determination R square is related to coefficient of correlation r. in case of simple linear regression i.e. if there is only one predictor and it is linearly associated with the dependent variable then coefficient of correlation r is square root of coefficient of determination.

coefficient of correlation (r) = square root {coefficient of determination (R square)}

11.6 Summary

There are several different measures to evaluate relationship between two variables. This chapter has only discussed Pearson and Spearman rank correlation. It is necessary to understand that relevant assumptions regarding type of data and its distribution should be checked before selecting a particular technique to find

association of two variables. Pearson correlation should be used only for interval data following a normal distribution whereas Spearman rank correlation is a non-parametric test to find association where data is ordinal rather than interval. Both techniques are interpreted in similar fashion as their correlation coefficients lie between +1 and -1. Positive correlation means that as value of one variable increases that of other also increases. Negative correlation means as value of one variable decreases that of other tends to increase. For r values near zero little or no correlation is present.

Regression analysis is a statistical tool that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other. This chapter discusses bivariate (two variables) regression where there is one independent variable *x* used to predict another dependent variable *y.* The regression model involves building of a regression equation which consists of slope of line as a coefficient of x and a y intercept value as a constant. The accuracy of regression model can be evaluated by using certain statistics. Standard error of estimate and coefficient of determination are two such statistics which are discussed in this chapter. These statistics involves evaluating observed and expected values of dependent variable. The standard error of estimate is the standard deviation of the error of model. This value of standard error of estimate can be used to analyze magnitude of error instead of studying each residual value. The coefficient of determination is the proportion of total variance of the y variable predicted by x. this value ranges from 0 to 1. High value of coefficient of determination can be interpreted as a significant indicator of accurate model. Lastly, relationship between correlation and regression analysis has been discussed. In the case of bivariate regression correlation coefficient is square root of coefficient of determination.

11.7 Glossary

- **Regression analysis:** is the process of constructing a model involving two variables which can be used to predict dependent variable from independent variable.
- **Least squares method:** is used to create best fit line called as regression line with least error. This method estimates the value of y intercept and slope intercept to create most accurate regression line which is used for prediction of dependent variable.
- **Standard error of estimate:** is the standard deviation of error of regression models which can be used to compare the efficacy of regression models.
- **Coefficient of determination:** represents the proportion of dependent variable explained by independent variable
- **Sum of squares of error:** is the sum of residuals squared in a regression model
- **Total sum of squares (SST):** is the sum of squared deviations about the mean of a set of values

11.8 SHORT ANSWER QUESTIONS

1. What is regression?
2. What is the difference between linear regression and multiple linear regression?
3. What are the assumptions of linear regression?

11.9 LONG ANSWER QUESTIONS

1. Explain the concept of multiple regression and its advantages over simple linear regression.
2. Describe the different types of regression models and their respective applications.

11.10 ANSWERS OF SELF CHECK QUESTIONS

11.3 a) $y = -11.335 + 0.355x$

b) with increase in a unit of independent variable dependent variable increases by amount of slope of line.

c) Least squares method

11.4.2 a) True

b) False

c) False

MBA-Distance Education (First year)

Semester-2

Lesson No. 12 _____ **AUTHOR : A.S. BHATIA**

RESEARCH METHODOLOGY

STRUCTURE

13.1 INTRODUCTION

The theory of probability owes its origin to the study of games of chance or gambling. In the games of chance, under the given conditions, more than one result is possible and which one of these results will actually appear cannot be predicted. For example, if a coin is tossed we cannot predict whether it will land head or tail, however we do not know that it will land on one or the other, so that the set of all possible outcomes is known. Probability theory is designed to deal with uncertainties regarding the happening of given phenomena. The word "probable" itself indicates such a situation, its dictionary meaning is "likely though not certain to occur."

The aim of probability theory is to provide a mathematical solution to all such situations arising in the games of chance. In order to develop theory of such random processes, we introduce the following definitions.

13.1.1 Random Experiment

An experiment conducted repeatedly under essentially homogenous conditions is known as random experiment. The tossing of a fair die, drawing a card from a pack of cards are the operations which are statistically as experiments.

13.1.2 Trial and Event

The performing of a random experiment is known as trial and outcome or combination of outcomes are termed as event.

For example, if a coin is tossed repeatedly, the result is not unique. We may get any of the two faces head or tail. So tossing a coin is trial and getting a head or tail is an event.

13.1.3 Favourable Cases of Events

It is termed as number of outcomes of a random experiment which result in the happening of an event.

For example, in tossing a die the number of favourable cases to an even number is 3 viz., 2, 4 or 6.

13.1.4 Independent Events

Two events are said to be independent when the happening of an event does not affect the happening of the other event.

For example, in tossing a die repeatedly, the event of getting ’4' in the first throw is indepdent of getting '4' in second, third or subsequent throws.

13.1.5 Mutually Exclusive Events

Events are said to be mutually exclusive, if the happening of any one of them rules out the happening of all others. In other words no two or more of them can happen simultaneously in the same trial.

For example, in tossing a coin the events head and tail are mutually exclusive. **Exhaustive Events**

The total number of all possible outcomes in any trial is known as exhaustive events. For example, in tossing a coin the possible outcomes are head and tail, are exhaustive events. In throwing a die, the possible outcomes, 1,2,3,4,5 and 6 are exhaustive events.

13.1.6 Equally Likely Events

Events are said to be equally likely if there is no reason to expect any one in preference to other.

For example, in tossing an unbiased coin, head or tail are equally likely events. In throwing an unbiased die, all the six faces are equally likely to come.

13.1.7 Simple Event, Compound Event

A single event is called a simple event and when two or more than two events occur in connection with each other then their simultaneous occurrence is called a compound event.

For example, when two coins are tossed, the occurrence of a head on the first coin and a head on the second coin is a compound event.

13.1.8 Complementary Events

If A and B are two events which are mutually exclusive and exhaustive, then A is called complementary of event B and vice versa. For instance in tossing a coin the event of appearing head and tail are complementary events.

13.1.9 Sample Space

A sample space is defined as the set of all possible outcomes of an experiment and is defined by capital alphabets. For example,

(i) In tossing a coin sample space is given by S=(H, T).

(ii) In tossing two coins simultaneously, sample space is given by S=(HH, HT, TH, TT).

(iii) In tossing there coins simultaneously sample space is S=(HHH, HHT, HTH, THH, HTT, THT, TTH, TTT).

(iv) In throwing a die, sample, space is S=(l, 2, 3, 4, 5, 6).

13.1.10 Algebra of Events (Event "A or B" and Event "A and B")

Let 'S' be the sample space of an experiment and two events defined by the condition either "A or B both occur" and "both A and B occur". These events will be called the events "A or B" and "A and B" and will be represented by the subsets AuB and AnB respectively.

13.2 DEFINITION OF PROBABILITY

The following are three approaches of the probability :

1. Classical or a priori approach
2. Relative frequency approach (Statistical or Empirical)
3. Axiomatic approach.

12.2.1 Classical or Priori Approach

In classical approach, "the probability is the ratio of the number of favourable cases to the total number of equally likely cases." If the probability of an event A is denoted by p, then

$$p \text{ or } P(A) = \frac{\text{Number of favourable cases}}{\text{Total number of equally likely cases}} = \frac{m}{n}$$

where m denotes number of favourable cases for an event A and n denotes total number of equally likely and mutually exclusive events.

3 . 1

For example, in tossing a die, the probability of getting an odd number is ^ie'2>

because any of the six equally likely event (1, 2, 3, 4, 5, 6) three are odd (1, 3, 5).

In symbols of any event can happen in m ways out of total equally likely and mutually exclusive n' ways, then the probability of happening of an event (called its success) is

m
/
—\ p or P\A) = — and the probability of non-happening of an event (called its failure) is q or P [Aj.

p(A) = l- —= 1-P(A)-^^ n n

Thus P (A) + P **(a)** = 1 i.e. p + q = 1.

12.2.2 Relative Frequency Approach (Statistical or Empirical)

According to Von Mises, "If an experiment is repeated after essential homogeneous anc identical conditions, then the limiting value of the ratio of the number of times the event occur is the number of trails. As the number of trails become infinitely large; it is called the probability of happening of the event, it being assumed that this limit is finite and unique.

If an event 'A' occurs in m times out of n, its relative frequency is . The probability of

'A' is defined as limit of ratio, ™ when n become sufficiently large.

Symbolically P(A) = $\lim_{n\to\infty} \frac{m}{n}$

12.2.3 Axiomatic Approach of Probability

In axiomatic approach, some concepts are laid down the following properties known as axioms are defined.

1. The probability of an event A i.e. P(A) varies between two limits of 0 and 1. It can be neither less than zero nor more than unity.
2. If there is a certainty for the event A to happen then the probability of A is unity.
   Symbolically : P (A) = 1
3. If there is certainty of the event A does not happen then the probability of A is zero.
   Symbolically : P (A) = 1
4. The probability of the happening or not happening of an event A is :

   P(A) + P (a) = 1 or

   P(A) = 1 - P **(a)** or

   p(a) = 1 - P(A)

12.2.4 Example 1. Two coins are tossed. Find the probability of 2 heads. What is the probability of exactly 1 head ?

Sol. The possible outcomes are (HH, HT, TH, TT)

Total number of outcomes = 4

/. P (2 heads) = ^

Exactly one head means 1 head and 1 tail. Therefore there are 2 favourable cases.

/. P (exactly one head) = $\frac{1}{4} = \frac{1}{2}$

12.2.5 Example 2. Three coins are tossed once. Find the probability of (i) 3 heads, (ii) exactly 1 heads (iii) at least 2 heads, (iv) at most 2 heads (v) no tails. Sol. The possible outcomes are HHH, HHT, HTH, THH, HTT, THT, TTH, TTT .'.

total number of outcomes = 8

(i) Number of favourable cases = 1

Required probability - $\frac{1}{O}$

Exactly 2 heads means 2 heads and one tail

Number of favourable cases = 3

Required probability = $\frac{3}{O}$

(i) At least 2 heads means, 2 head 1 tail or 3 heads no tail.
.•. Number of favourable cases = 4

/. Required probability = $\frac{2}{o} = \frac{1}{2}$

(ii) At most 2 heads, are in the cases
HHH, HHT, HTH, THH, HTT, THT, TTH, TTT
Number of favourable cases = 7
Required probability - $\frac{7}{o}$

(iii) No tail means all heads .\ Favourable cases HHH
Number 1 of favourable cases = 1

Required probability = $\frac{1}{O}$

12.2.6 Example 3. A card is drawn from a well suffeled pack of 52 cards. Find the probability of drawing (i) king (ii) heart (iii) a seven of heart.

Sol. Total number of possible drawing is = 52

(i) Number of favourable cases of drawing a king = 4
Probability of drawing a king = $\frac{4}{OZ} = \frac{1}{1\ o}$

(ii) Number of favourable cases of drawing a heart =13
Probability of drawing a heart = $\frac{13}{—} = \frac{1}{1}$

(iii) Number of favourable cases of drawing seven of heart = 1

.. Probability of drawing a seven of heart = $\frac{1}{oz}$

12.2.7 Example 4. A die is thrown once. Find the probability of getting (i) even number (ii) number greater than equal to 3 (iii) number between 2 and 5.

Sol. A die has six faces marked 1, 2, 3, 4, 5, 6

(i) Even number on the face are 2, 4, 6 .•. Number of favourable cases is = 3
P (even number) = $\frac{3}{6} = \frac{1}{2}$

(ii) Number >3 on the face of die are 3, 4, 5, 6 .'. Number of favourable cases is = 4

P (a number > 3) - g = g $^{4\ 2}$

(iii) Numbers between 2 and 5 are 3, 4 Number of favourable cases is = 2

/. P (a number between 2 and 5) = $\frac{2}{}=\frac{1}{}$

12.2.8 Example 5. From an urn containing 6 red and 3 black balls, a ball is drawn at random. What is the probability of drawing a red ball ?

Sol. Random drawing of balls ensured equally likely outcomes. There are 6 red balls out of total of 6+3=9 balls.

P (a red ball) = |=|

13.3 SELF CHECK EXERCISE - I

(a) Two dice are thrown. Describe the sample space of this experiment.

(b) If a coin is tossed two times, describe the sample space associated to this experiment.

13.4 USE OF COMBINATIONS

We know number of combinations of n things taken r at a time is $^{n}C_{j}$. n!
where

13.4.1 Principle of Association

If one operation can be performed in m ways and the second operation can be performed in n ways, then the two operations simultaneously can be performed in mxn ways.

13.4.2 Example 6. Four balls are drawn from a box containing 7 red and 5 white balls. Find the chance that the balls drawn are all red.

Sol. Total number of balls 7 + 5=12, 4 balls can be drawn out of 12 balls in $^{12}C_4$

12! 12.11.10.9 ways -
4 j s |– 4.3.2.1

/. Total number of all equally likely cases = 495 = 4131" 3.21 $\frac{7!\_ 7.6.5}{}$

Now 4 red balls can be drawn out of 7 in $^{7}C_4$ ways.

.■. Total number of favourable cases = 35

Hence the required chance of drawing 4 red balls = ^T=gg $\frac{35\quad 7}{}$

13.4.3 Example 7. A bag contains 4 white and 3 black balls. If 4 balls are drawn at random, what is the probability that 2 are white and 2 black ?

Sol. Total number of balls = 4+3=7

4 balls can be drawn out of 7 in $^7C_4$ ways = ~ ^

Total number of all possible cases for the event = 35

Again, 2 white balls can be drawn out of 4 in $^4C_2$ways and 2 black balls can be drawn out of 3 in $^3C_2$ ways.

^{2} white and 2 black balls can be drawn in ways $^4C_2 . ^3C_2$

.. Total number of favourable cases = ^{4} Co.^{3} Co = $\frac{43\ 33}{z\ 21\ 21}$ = 77r$^{x'}$XT

Hence the required probability of drawing 2 white and 2 black balls = $\frac{18}{oo}$

13.4.4 Example 8. Two unbiased dice are thrown. Find the probability that the sum of the faces is not less than 10.

Sol. Total number of outcomes with the throwing of two dices = [(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2) .... (6, 6) = 36

Let A be the event that the sum of the faces not less than 10, i.e.

A = [(4, 6), (5, 5), (6, 4), 5, 6), (6, 5), (6, 6)] = 6

Hence the required probability = = — $\frac{6}{36}$

13.5 SELF CHECK EXERCISE - II

a) A bag contains 4 white, 5 red and 6 blue balls. Three balls are drawn at random from the bag. The probability that all of them are red, is –.

13.6 PROBABILITY RULES OF ADDITION

'Here we shall discuss the rule to find the probability of the sum or union of two or more events. For this purpose, we first prove two theorems called the theorems of total probability.

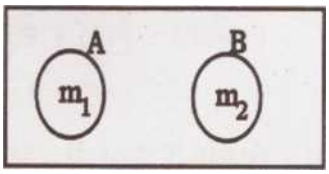
13.6.1 Theorem 1. If A and B be two mutually exclusive events then

P (AuB) = P(A+B) = P(A) + P(B)

Proof. : Let n be the total number or exhaustive, equally likely cases of the experiment.

Let m, and $m_2$be the number of cases favourable to the happening of the events A and B respectively.

• **P(A) = HL,p(B)** = E2
n n



Since the events A and B are mutually exclusive.

There cannot be any sample point common to both events A and B.

The event A or B can happen in exactly m,+m_2 ways.

P iAuBI = Probability that either A or B will occur

m **i** — m 2 m i m.2

n n n

= P(A)+P(B)

Hence P(AuB) or P(A+B) =P(A) + P(B)

Note : The result of this theorem can be extended to any number of mutually exclusive events.

13.6.2 Example 9. If a die is rolled, what is the probability that the roll yield a 3 or 4.

Sol. A die can be thrown (or rolled) in six ways and the possible outcomes are 1, 2, 3, 4, 5, 6.

Now P(A) = Probability of getting 3 = ^

P(B) = Probability of getting 4 = ^

As the events are mutually exclusive.
Probability of getting 3 or 4 = P(AuB) = P(A) + P(B)

1 1 = 1
6 + **6** " 3

13.6.3 Example 10. One card is drawn from a standard pack of 52. What is the probability that it is either a king or a queen ?

Sol. There are 4 king and 4 queens in a pack of 52 cards.

/. The probability that the card drawn is a king = P(A) = — 4

and the probability that the card drawn is a queen = P(B) = — 4

Since the events are mutually exclusive, the probability that the card drawn is either a king or a queen i.e. P(AuB) = P(A)+P(B)

4 8 **2**
" 52 + 52 ~ 52 " 13

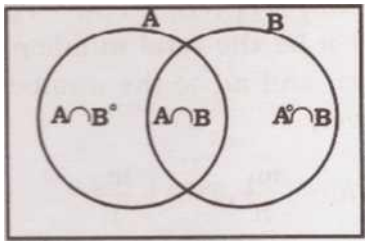
13.6.4 Theorem II. If the events A and B be any two not mutually exclusive events, then

P(AuB) = P(A) + P(B) - P (AnB).

Sol. Two events A and B are not mutually exclusive.

Then from the figure, event AuB can be decomposed into two mutually exclusive events A and A°nB.
By theorem I

P(AuB) = P(A) + P (A°nB) - (1)



Again event B can be decomposed into two mutually exclusive events AnB and A°uB from the figure.

.\ By theorem I, P(B) = P(AnB)+(A°uB)

or P(A°nB) = P(B) - P(AnB) (2)

Hence from 1 and 2, we get

P(AuB) = P(A) + P(B) - P (AnB) (3)

13.6.5 Example 11. One card is drawn from a standard pack of 52. What is the probability that it is either a king or a heart.

Sol. There are 4 king and 13 heart in a pack and 52 cards.

The probability that the card drawing king P(A) = $\frac{4}{}$

The probability that the card drawing heart P(B) = $\frac{13}{}$

The probability that the king is of heart.

P(AriB) = ^

The probability that the card drawn is either a king or a heart is P(AuB)

= P(A) + P(B) - P(AnB)

$$\frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$$

13.7 SELF CHECK EXERCISE - III

(a) What is the probability of picking a card that was red or black.

(b) Two events A and B are mutually exclusive : P(A) = 1/5 and P(B) = 1/3, find the probability that

(i) Either A or B will occur, (ii) both A and B will occur,

(ii) Neither A nor B will occur.

13.8 MULTIPLICATION THEOREM

13.8.1 Theorem 3. If two events A and B are independent, the probability that they both will occur is equal to the product of their individual probability. Symbolically, if A and B are independent events, then

P(A and B) = P (AnB) = P(A). P(B)

Proof. If an event A can happen in n_t ways of which a, are successful and the event B can happen in n_2 ways of which a_2 are successful. We can combine each successful event in the first with each successful event in the second case. Thus, the total number of successful happenings in both cases is * a^

Similarly, the total number of possible cases is n,xn_2. Then by definition the probability of the occurrence of both events A and B is

$$P(AnB) = \frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} \times \frac{a_2}{n_2}$$

But ^- = P(A)and^- = P(B) n_1 n_2

. P (A^B) = P(A) x P(B)

The theorem can be expanded to three or more independent events.

Thus P(A, B and C) = P(A) * P(B) x p(C)

13.8.2 Example 12. A die is thrown twice. Find the probability of a number of greater than 4 on each throw.

Sol. Let A be the event of getting 'a number greater than 4 on first throw" and B be the event 'a number greater than 4 on the second throw." Event A and B are independent.

P(A and B) = P(AoB) = P(A). P(B)

= P (5, 6). P (5, 6)
= 1 2 _i J_

~ 66 " 33 ~ 9

13.8.3Example 13. A problem in statistics is given to three students whose chance of solving $\frac{1}{1}$, 1 it are and ^ respectively. What is the probability that the problem will be solved.

Sol. Let the three students be denoted by A, B and C.
Then probability and A solves the problem

i.e. P(A) = ^

P(B| - ±

P(C| -i

.'. $P(A^C)$ = P(A fails to solve) = l-P(A) = 1-^ = ^

P(B°) = P(B fails to solve) = l-P(B) = 1-= — = —
3 3

$P(C^C)$ = P(C fails to solve) = l-P(C) = 1-^- = ^

P (A, B, C fails to solve) = $P(A^c \cap B^c \cap C^c)$ = ~~~ = —
'2 3 4 4

(.'. events are independent)

Hence the problem will be solved = 1-P (Problem is not solved)

$= 1-\frac{1}{4}=\frac{3}{4}$

Thus the required probability = — 3

13.9 SELF CHECK EXERCISE - IV

Two cards are randomly chosen one after the other from a standard deck of 52 playing cards without replacement. What is the probability of choosing two kings?

Glossary

- **Trial and Event**: The performing of a random experiment is known as trial and outcome or combination of outcomes are termed as event.
- **Independent Events**: Two events are said to be independent when the happening of an event does not affect the happening of the other event.
- **compound event:** when two or more than two events occur in connection with each other then their simultaneous occurrence is called a compound event

13.10 Short Answer Questions

Q.1. Explain various approaches to probability.

Q.2. Explain the terms mutually exclusively events and independent events. Give one example for each. State and prove the addition rule of probability.

that (i) either A or B will occur, (ii) both A and B will occur, (iii) neither A nor B will occur.

Q.4.In a single throw of two dice, determine the probability of getting a total of 8 or 12.

LONG ANSWER QUESTIONS

1. Discuss the Bayes' theorem and its applications in various fields.
2. Explain the difference between probability distributions and cumulative distribution functions.
3. If two coins are tossed simultaneously, what is the probability of getting exactly two heads?

ANSWERS OF SELF CHECK QUESTIONS

12.3 a) 36

b) 4

12.5 a) 2/91

12.7 a) 1

b) (i) 8/15 (ii)1/15 (iii)7/15

19.9 a) 1/221

13.11 SUGGESTED READINGS

1. Statistical Methods : S.P. Gupta
2. Theory and Problems of Statistics : Spiegal R. Murray
3. Statistical Methods : R.C. Joshi

Chapter 13: Probability Distributions

13.0 Objectives

The students should be able to capture the following concepts:

- Meaning and significance of probability distributions
- Type of probability distributions
- Characteristics and computation of Binomial distribution
- Characteristics and computation of Poisson distribution
- Characteristics and computation of Normal distribution

13.1 Introduction

Distributions play a significant role for a statistician in deciding which analysis should be applied to the data for decision making. Distributions indicate the behavior of data showing certain characteristics. For instance, in a class of 100, majority of students score average marks, few score very less and few score very high. So most likely such data show a symmetrical behavior. In an organization very few people earn very high salaries and a large chunk of employees earn average or small wages. Such data most likely shows asymmetrical characteristics. Thus, a manager or a decision maker should understand the distributions of data.

Probability distributions are depiction, either numerically or graphically, of probabilities of all the possible outcomes in an experiment. For example, in an experiment of throwing a dice there can be six possible outcomes and for each outcome a probability of its occurrence can be computed. The selection of a particular kind of probability distribution depends on the type of variable. Firstly, the variable should have a random outcome i.e. its outcome should be dictated by chance rather than by inference. In a toss of coin the occurrence of head or tail is governed by chance rather than by previous outcomes. Secondly, are the outcomes of the variable being counted or measured. An automobile repair workshop encounters number of cars daily. Assigning probability to a randomly selected car being a foreign made car requires *counting* of cars whereas assigning probability to a potato chip pack to be underweight requires *measuring* the weight of potato chip pack.

Depending on the type of data various probability distribution can be categorized into discrete and continuous random distributions. ***Discrete random probability distributions*** involve distributions of those variables the outcome of which can be counted and probability is determined for such outcomes like number of one rupee coins in a bag, number of defective parts in a sample of produced parts, etc. The random variable which can be counted called discrete variable can occur over a sample like finding defective parts in a sample of 20 parts or getting heads when experiment of tossing a coin is conducted for five times. In these cases number of defective parts or number of heads are counted. In the next step probability of obtaining defective parts in the given sample or probability of getting heads is obtained. The random variable can also occur over a given interval. For instance, average number of foreign made cars coming to car repair store in a ten minute interval or average number of chocolate chips in a chocolate cookie. In these cases, time and a cookie is an interval. In the next step probability of getting a certain number of foreign made of cars in the given interval or probability of getting certain number of chocolate chips in the given interval of cookie is obtained.

Continuous random distributions involve distributions of variables which can be measured like time, weight, height, volume etc. It is difficult and infeasible to determine exact amount of sugar in one kilogram of sugar. The weight can be slightly higher or lower than one kilogram. So, probability of the interval in which weight of one kilogram of sugar is computed rather than probability of exact outcome. Probability is not measured in terms of success or failure but whether the outcome is within the given interval or not. For example, determining probability that a student would score in a given interval of percentage would result in assigning him first division or otherwise. In another example, in an army recruitment drive if average height for selection is six feet, then it would be difficult and infeasible to find a candidate with exact height of six feet and ultimately assigning probability to such an event. In this case the experiment is picking of a candidate and random outcome be the height of the selected outcome. Now, height is a measurable data and assigning probability to an exact height outcome would be infeasible. So, it would be more appropriate and feasible to determine probability that a candidate has a height within an interval of five feet eight inches and six feet two inches. Now, if probability is to be found for more than one candidate then it would result in formulation of continuous probability distribution.

This chapter discusses two kind of discrete probability distributions: Binomial and Poisson, and one continuous probability distribution: normal distribution.

13.1 Self Check Questions 1

a) Which one of these variables is a continuous random variable?

i. The time it takes a randomly selected student to complete an exam.
ii. The number of tattoos a randomly selected person has.
iii. The number of women taller than 68 inches in a random sample of 5 women.
iv. The number of correct guesses on a multiple choice test.

b) Continuous random variables are obtained from data that can be measured rather than counted.

(i) True (ii) False

c) Discrete variables have values that can be measured.

(i) True (ii) False

d) A(n) _____ is one in which values are determined by chance.

e) A(n) _____ probability distribution consists of the values in which a random variable can assume the corresponding probabilities of the values.

13.2 Binomial distribution

Binomial distribution is a kind of discrete random distribution which involves assigning probabilities to the experiments which can be counted and have only two outcomes. For instance, if the experiment is appearing in an exam then the candidate can have only two outcomes i.e. pass or fail. In an experiment of tossing of a coin the outcome can only be head or tail. Both outcomes cannot occur simultaneously.

13.2.1 Characteristics of Binomial Distribution:

- *The experiment involves n identical trials*

  This means that if throwing a dice is an experiment and a sample of five trials is taken then the experiment should involve only throwing of the dice. The trials should not be different from each other.

- *Each trial has only two possible outcomes as success or failure*

  As the experiment is conducted on the basis of its outcome being random, so an experiment can have only outcomes in terms of either a success or a failure.

- *Each trial is independent of the previous trials*

  For instance, in tossing of coin experiment if an outcome is head in the first trial, then it does not have any impact on the occurrence of head in succeeding trials. Thus, in binomial distribution the trials are mutually exclusive and independent where two trials cannot occur simultaneously and their outcomes are independent of each other. This constraint is possible in cases where the experiment by nature produces

independent trials such as tossing of coin, rolling of dice, result of an exam etc. or the experiment is conducted with replacement. For instance if of all items in a bin 5% are known to be defective then probability of finding a defective item when one is withdrawn from the bin is p = 0.05. If this item is not replaced in the bin and second item is withdrawn from the bin then probability of getting a defective item would be different from the first trial as sample size or the number of limited items in the bin has changed. The binomial distribution would not be applicable in such cases as it does not allow probability of an outcome to change from trial to trial. This condition can be relaxed if population is very high. For instance in a potato chip manufacturing facility number of chips produced daily is enormous. So to check its quality a small sample is taken. In such cases even without replacement the independence assumption is generally met.

- *The probability of success denoted by p and that of failure denoted by q =1-p remains constant for every trial.*

This implies that probability of occurrence of one outcome in a trial is independent of occurrence of same outcome in another trial. For instance, probability of getting head in one trial is 0.5, which would be the same incase of other trials.

Some of the examples of binomial distribution are:

- Suppose a vending machine has an error rate of 5%. If 50 individual use this machine, what is the probability that less than 2 individuals encounter error?
- A company promises to deliver pizzas within half an hour. The probability of failure is 10%. To check the success of the policy the manager decides to take a sample of 20 deliveries. What is the probability that less than one delivery failed?
- After logging on to amazon.com the probability of buying is 20%. What is the probability that out of 100 people logging at a particular time less than four people would not buy?

13.2.2 Computation of Binomial probability problem

This would be illustrated by conducting an experiment of tossing a coin three times i.e. number of trials denoted by n = 3. There can be only two possible outcomes, head (H) or tail (T). The outcome of random variable is denoted by X. The probability of getting a head is termed as success and is denoted by p whereas getting tail is failure and is denoted as q = 1-p. For illustration, if trial is conducted for the first time i.e. n=1, the outcome can be head or tail. If the outcome is head with probability p, then second trial is conducted (n=2) and outcome can be head or tail. If outcome is head with probability p, then third trial is conducted (n=3) and outcome can be head or tail. If outcome is head then for three trials the probability would be p*p*p. Similarly, probability for all the cases can be obtained which is shown in table 1.

Table 1

Events	Probability	No. of successes
HHH	p*p*p	3
HHT	p*p*(1-p)	2

HTH	p*(1-p)*p	2
HTT	p*(1-p)*(1-p)	1
THH	(1-p)*p*p	2
THT	(1-p)*p*(1-p)	1
TTH	(1-p)*(1-p)*p	1
TTT	(1-p)*(1-p) *(1-p)	0

Probability of occurring of any event from above illustration can be generalized as

$P(X) = p^x(1-p)^{n-x}$

For instance, probability of 2 successes i.e. probability of getting head twice in three trials implies that when, p=0.5

$P(X=2) = 0.5^3 * (1-0.5)^{3-2}$

But according to table 2 successes can be obtained three times. Thus, final formula in case of binomial distribution becomes

$$\mathbf{P(X) = {}^nC_x * p^x(1-p)^{n-x}}$$

Where C is combination calculated by using formula n! / (x!)*(n-x)!

Example 1: According to government figures 6% of all workers in a city are unemployed. While conducting a small survey for the city what is the probability of getting two or fewer unemployed workers in a sample of 20?

Solution:

Two or fewer unemployed employees indicate 0, 1 or 2. Thus solution becomes

$P(X <=2) = P(X=0) + P(X=1) + P(X=2)$

$= {}^{20}C_0 * 0.06^0(1-0.06)^{20-0} + {}^{20}C_1 * 0.06^0(1-0.06)^{20-1} + {}^{20}C_2 * 0.06^2(1-0.06)^{20-2}$

$= 0.8850$

13.3 Poisson Distribution

Poisson distribution is another discrete random distribution involving assigning probabilities to countable outcomes over an interval. It is different from binomial distribution because here number of trials is not given whereas in binomial distribution trials happen over a given sample space. For instance, number of erroneous calls over a given number of calls is a binomial distribution, whereas number of erroneous calls over a given interval of say five minutes is a Poisson distribution. Another difference between two distributions is that in Poisson distribution instead of sample space events happen over an interval. For instance, number of defects per carpet, number of chocolate chips per cookie, number of customers coming in a bank during lunch hour etc.

13.3.1 Characteristics of Poisson Distribution

- *Occurrence of success in any one interval is independent of that in any other interval.*

For example, number of chips in a cookie is independent of those in another cookie; arrival of customer during an interval is not dependent on arrival of another customer in other interval; a customer passing through a security gate from another customer passing through the same gate at different point in time.

- *Probability of observing more than one success in any one interval is zero.*

This implies that if happening of an event over a continuum ranges from zero to infinity then that continuum should be broken down to smaller intervals in such a way that only one event should happen in that interval. For example, number of customers coming in a bank during lunch hour is a discrete event. Now, suppose on an average 180 customers arrive during one hour. In this case the interval is big i.e. of one hour so probability of observing more than one success i.e. arrival of a customer is different from zero. So, interval is broken down into shorter duration say in seconds. In one second chance of arrival of more than one customer is very rare. Thus, the occurrence of more than one success in that interval is zero.

- *Probability of success in an interval is same for other intervals.*

This implies intervals to be mutually exclusive i.e. events cannot occur simultaneously. For instance, by taking same example when interval of one hour is broken down into 3600 seconds then probability of arrival of one customer out of 180 would be 180/3600 = 0.05. As only one customer can arrive per interval of one second so probability of arrival of another customer in second interval of one second would be same i.e. 0.05. A corollary of this characteristic is that probability of occurrence of an event is interval dependent. If interval is of one second then probability is 0.05 but if interval is of two seconds then probability of occurrence would also be doubled to 0.1 as now two customers might arrive during increased interval.

Some of examples of Poisson distribution are:

- Number of accidents per hour.
- Number of wrong delivery of newspapers per morning.
- Number of customers coming in a restaurant every five minutes on a busy Saturday night.
- Number of defective pens per carton.
- Number of golf players in a small city.
- Number of people having a rare disease in population of ten lakh.

13.3.2 Computation of Poisson Distribution problem

If a Poisson distribution phenomenon is studied over a long period of time, then a long run average of that event can be estimated denoted by **lambda (λ).** A binomial distribution required n and p to describe occurrence of discrete variables, whereas a Poisson distribution can be described by lambda. The formula to calculate probability of occurrences of a discrete variable over a given interval is

$$P(X) = e^{-\lambda} * \lambda^{x} / x!$$

where λ = long run average

e = exponential constant = 2.718

x = number of occurrences per interval for which the probability is being computed.

Example 2: Number of calls received by an operator between 9 and 10 am has a Poisson distribution with a mean of 12. What is the probability that an operator received at least 5 calls during:

(i) 9 and 10 am

(ii) 9 and 9:30 am

(iii) 9 and 9:15 am

Solution:

(i) $P(X>=5) = 1 – [P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)]$

Solving for $P(X=0) = 2.178^{-12} * 12^0 / 0!$

Similarly probability for x = 1, 2, 3 and 4 can be found.

(ii) As discussed earlier lambda is dependent on interval duration. In this case interval duration is halved so value of long run average i.e. average number of events happening in that interval will also be halved. Thus, $\lambda = 6$

(iii) In this case $\lambda = 4$

Calculations are left as exercise for student.

13.3 Self Check Questions 2

a) A medical treatment has a success rate of .8. Two patients will be treated with this treatment. Assuming the results are independent for the two patients, what is the probability that neither one of them will be successfully cured?

(i) 0.5 (ii) 0.36 (iii) 0.2 (iv) *.04 (this is (1 – .8)(1 – .8) = (.2)(.2) = .04)*

b) The probability of a success must remain the same for each trial in a binomial experiment.

(i) True (ii) False

c) In binomial experiments, the outcomes are usually classified as successes or failures.

(i) True (ii) False

d) In a binomial experiment, the outcomes of each trial must be dependent on each other.

(i) True (ii) False

e) When sampling is done without replacement, the binomial distribution does not give exact probabilities because the trials are not independent.

(i) True (ii) False

f) A coin is tossed five times. Find the probability of getting exactly three heads.

(i) 0.3750 (ii) 0.1563 (iii) 0.2500 (iv) 0.3125

g) If a student randomly guesses at 20 multiple-choice questions, find the probability that the student gets exactly four correct. Each question has four possible choices.

(i) 0.19 (ii) 0.17 (iii) 0.08 (iv) 0.23

h) One of the requirements for a binomial experiment is that there must be a _____ number of trials.

i) The Poisson distribution is used when n is small and p is large.

(i) True (ii) False

j) Which one of these variables is a binomial random variable?

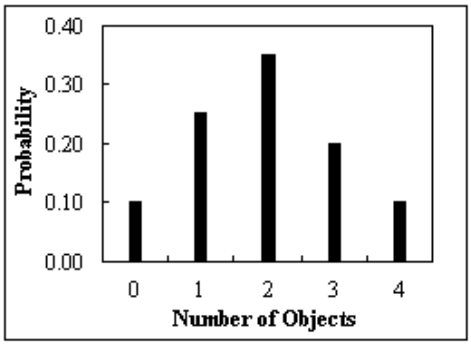
(i) time it takes a randomly selected student to complete a multiple choice exam

(ii) number of textbooks a randomly selected student bought this term

(iii) *number of women taller than 68 inches in a random sample of 5 women*

(iv) number of CDs a randomly selected person owns

k) The figure below represents the probability distribution for selecting a number of objects out of a container. Construct a probability distribution from this graph.

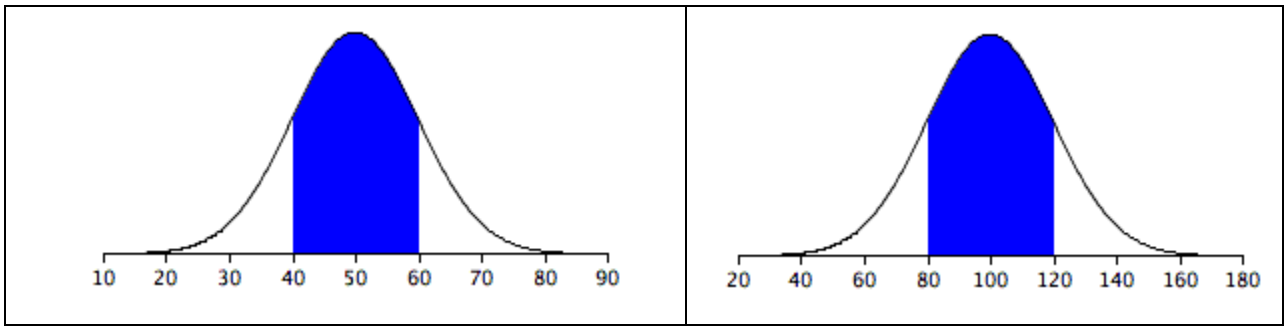


13.4 Continuous probability distribution: Normal Distribution

Normal distribution is the most applied distribution because of its use in various decision making processes. The distribution is appropriate for assigning probability for the occurrence of continuous data. The data which is not counted but measured like, height, weight, length, speed etc. are put under normal distribution. Many researchers use scaling techniques like Likert scale to measure customers' satisfaction or any other behavioral characteristic. These variables are continuous in nature as it measures an individuals' behavior. The data regarding these variables follow normal distribution making it easy for decision maker to reach an inference.

Graphically Normal distributions are represented by bell shaped curve as shown in Fig. 1. Figure 1 shows a normal distribution with a mean of 50 and a standard deviation of 10. The shaded area between 40 and 60 contains 68% of the distribution.

Fig. 1	
Normal distribution with a mean of 50 and standard deviation of 10.	A normal distribution with a mean of 100 and a standard deviation of 20.



The normal distribution shown in Figure 1 is specific example of the general rule that 68% of the area of any normal distribution is within one standard deviation of the mean.

13.4.1 Characteristics of normal distribution:

- It is a continuous distribution.
- It is a symmetrical distribution about its mean implying that each half of the distribution is the mirror image of the other half.
- It is unimodal. The graph has only one peak indicating that in the continuous data there is only one value which has highest frequency.
- Area under the curve is one.

The normal distribution bell shaped curve is described by two parameters: mean and standard deviation. For each value of mean and standard deviation the graph would carry a different shape. If mean of one graph is more and standard deviation is less than other then it would have higher peak and narrow width. Whereas, high standard deviation spreads the graph wide. For a perfect normally distributed graph mean would be equal to zero with one standard deviation.

Because of different normal distribution with different value of mean and standard deviation the generalized or standardized formula for normal distribution is given by z distribution.

$$z = (x - \mu) / \sigma$$

where x depicts the value of data for which probability in comparison to mean value has to be found. If the value of x is less than mean the z score would be negative and if x value is more than mean then z score would be positive. If x value is equal to mean value then z score would be equal to zero implying that the selected data (x) has zero deviations from the mean.

Example 3: In case of GMAT test with a mean of 494 and a standard deviation of 100, what is the probability that a candidate would score

(i) greater than 700

(ii) equal to or less than 550

(iii) between 300 and 600

Solution:

(i) the z score for this problem is:

$z = (700 - 494) / 100$

$= 2.06$

From z tables this value of z gives the probability as 0.4803. But this probability is for the area between mean value and 700. The question asks for probability greater than 700. As we know that in case of normal distribution the area of curve is 1 and mean divides the curve into two equal halves. The probability for one half is 0.5 So, the required probability is

= 0.5 (probability of x greater than mean) – 0.4803 (probability of x between mean and 700)

= 0.0197 (probability of greater than 700)

This means, that probability that a candidate would score greater than 700 would be 1.97%.

(ii) the z score for this problem is:

$z = (550 - 494) / 100$

$= 0.56$

The associated probability of x between mean 494 and 550 from the z tables is 0.2123. but the question asks for finding the probability that a candidate will score less than 550. So the required probability is:

= 0.5 (probability of x less than mean) + 0.2123 (probability of x between 550 and mean)

= 0.7123

This means, that probability that a candidate would score less than 550 would be 71.23%

(iii) the z score for this problem is calculated in two parts:

between 300 and mean:

$z = (300-494) / 100$

$= -1.94$

The associated probability from z tables for x between 300 and 494 is 0.4738.

Between mean and 600

$z = (600-494) / 100$

$= 1.06$

The associated probability for x between 600 and 494 is 0.3554.

So, the required probability is

= 0.4738 (probability of x between 300 and 494) + 0.3554 (probability of x between 494 and 600)

= 0.8292

This means, that probability of a candidate scoring between 300 and 600 would be 82.92%.

Example 4: An agency publishes data on solid waste generation. One year the average number of waste generated per person per day was 3.58 kgs. Suppose the daily amount of waste generated per person is normally distributed,

with a standard deviation of 1.04 kgs. Of the daily amounts of waste generated per person, 67.72% would be greater than what amount?

Solution: the mean and standard deviation are given but x and z are unknown. The problem is to determine specific x value when 0.6772 of the x values are greater than that value. If 0.6772 values are greater than x then 0.6772-0.5000 = 0.1772 are between x and the mean value. According to z table the value of z for 0.1772 is 0.46. because x is less than the mean, the z value actually is -0.46.

Solving the z equation gives:

$z = (x - \mu) / \sigma$

$-0.46 = (x - 3.58) / 1.04$

Thus, $x = 3.10$

So, 67.72% of the daily average amount of solid waste per person weighs more than 3.10 kgs.

13.4 Self Check Questions 3

a) Heights of college women have a distribution that can be approximated by a normal curve with a mean of 65 inches and a standard deviation equal to 3 inches. About what proportion of college women are between 65 and 67 inches tall?

(i) 0.75 (ii) 0.50 (iii) 0.25 (iv) 0.17

b) Suppose that vehicle speeds at an interstate location have a normal distribution with a mean equal to 70 mph and standard deviation equal to 8 mph. What is the z-score for a speed of 64 mph?

(i) −0.75(ii) +0.75(iii) −6 (iv) +6

c) Pulse rates of adult men are approximately normal with a mean of 70 and a standard deviation of 8. Which choice correctly describes how to find the proportion of men that have a pulse rate greater than 78?

(i) Find the area to the left of z = 1 under a standard normal curve.

(ii) Find the area between z = −1 and z = 1 under a standard normal curve.

(iii) Find the area to the right of z =1 under a standard normal curve.

(iv) Find the area to the right of z = −1 under a standard normal curve.

13.5 Summary

Probability distributions are important to understand the characteristic of data which would help a researcher to apply different statistics tools. This chapter discusses three probability distributions: binomial, Poisson and normal distributions. The application of these distributions depends on type of data under consideration. Binomial and Poisson distributions are applied where data is countable over a sample space or under given interval. Normal distribution is a continuous distribution where data is measurable like volume, height, consumer behavior etc. The chapter understands these distributions by illustrating certain numerical problems.

13.6 Glossary

- **Discrete distributions:** the probability distributions which involve countable data like, number of 10 Rs. notes etc.
- **Continuous Distributions:** the probability distributions which involve measurable data like height, weight, volume etc.
- **Binomial distribution:** is a kind of discrete distribution in which there are only two possible outcomes on the occurrence of an experiment. It involves specific number of experiments. For instance, determining probability of certain number of defective products from a specific number of manufactured products.
- **Poisson distribution:** is also a discrete distribution where events happen over an interval which could be time or space.
- **Normal Distribution:** it is a kind of continuous distribution which involves measurable data such as many human behaviors and timing or capacity of machines. This is most widely used distribution as many statistical applications are based on normally distributed data.

13.7 Short Questions

1. How binomial and Poisson distributions similar and different?
2. Explain discrete and continuous distributions.

13.8 Long Questions

1. According to a survey by a consumer magazine 60% of all consumers have dialed an '100' or '200' telephone number for information about some product. Suppose a random sample of 25 consumers is contacted about their buying habits
   (i) What is the probability that 15 or more of these consumers have called '800' or '900' telephone number for information about some product?
   (ii) What is the probability that fewer than 10 of these consumers have called '800' or '900' telephone number for information about some product?

2. The average number of annual trips per family to amusement parks is Poisson distributed with a mean of 0.6 trips per year. What is the probability of randomly selecting a family and finding:
   (i) The family did not make a trip to an amusement park last year?
   (ii) The family took exactly one trip to an amusement park last year?
   (iii) The family took two or more trips to an amusement park last year?
3. According to a report the average monthly household cellular phone bill is $60. Suppose local monthly household cell phone bills are normally distributed with a deviation of $11.35. What is the probability that a randomly selected monthly phone bill is

(i) More than $85.

(ii) Between $45 and $70

(iii) Between $65 and $75

(iv) No more than $40.

13.9 Answers to Self check questions/ Self assessment exercise

Self check 1

A.i

B.i

C.ii

D.random variable

E.discrete

Self check 2

A.iv

B.i

C.i

D.ii

E.i

F.d

G.a

H.fixed

I.b

J.c

K.

No. of objects, x	0	1	2	3	4
P(x)	0.10	0.25	0.35	0.20	0.10

Self check 3

A.iii

B.i

C.iii

13.10 References/ Suggested Readings

- Black, K., *Business Statistics For Contemporary Decision Making,* Fifth Edition, Wiley India,
- Keller, G., *Statistics for Management,* First India Reprint 2009, Cengage Learning India Private Limited.
- Stevenson, W.J., *Operations Management,* Ninth Edition, Tata McGraw Hill, New Delhi, 2009.
- Donald R. Cooper & Pamela S. Schindler, *Business Research Methods,* Tata McGraw-Hill Publishing Co. Ltd., New Delhi, 9th Edition.
- S.P. Gupta, *Business Statistics*, Sultan Chand, New Delhi.

MBA-Distance Education
(first year)

BSRM 201
BUSINESS STATISTICS &
RESEARCH METHODOLOGY

Lesson No.14

AUTHOR : DR. PARMOD K. AGGARWAL

TESTING OF HYPOTHESIS

STRUCTURE

14.1 INTRODUCTION

The main object of the sampling theory is to study the Tests of Hypothesis or Tests of Significance. In many circumstances, we are to make decisions about the population on the basis of only sample information. For example, on the basis of sample data, (i) a quality control manager is to determine whether a process is working properly, (ii) a drug chemist is to decide whether a new drug is really effective in curing a disease, (iii) a statistician has to decide whether a given coin is unbiased, etc. Such decisions are called statistical decisions (a simply decisions). The theory of testing of hypothesis employs various statistical techniques to arrive at such decisions on the basis of the sample study.

14.2 BASIC CONCEPTS OF HYPOTHESIS TESTING

Hypothesis (or Statistical Hypothesis) : In attempting to arrive at decisions about the population on the basis of sample information, it is necessary to make assumptions about the population parameters involved. Such an assumption (or statement) is called a statistical hypothesis which mayor may not be true.

The following basic concepts are used in the study of tests of hypothesis :

14.2.1 Null Hypothesis

In testing of hypothesis we always begin with an assumption or hypothesis (i.e. assumed value of a population parameter). This is called Null Hypothesis. The null hypothesis asserts that there is no (significant) difference between the sample statistic and the population

parameter and whatever the observed difference is there, is merely due to fluctuations in sampling from the same population. Null hypothesis is usually denoted by the symbol H_o. R.A. Fisher defined null hypothesis as the hypothesis which is tested for possible rejection under the assumption that it is true. In other words, the hypothesis (regarding some characteristic of population) which is to be verified with the help of a random sample or the hypothesis which is under test is called null hypothesis. For example, it we want to test the hypothesis that the mean of the population to be taken as po, then the null hypothesis (H_o) is $n = n_o$.

14.2.2 Alternative Hypothesis

Any hypothesis different from the null hypothesis (H_o) is called an alternative hypothesis and is denoted by the symbol H,. The two hypothesis H_o and H_1 are such that if one is accepted, the other is rejected and vice versa.

For example, if we want to test whether the population mean p has a specified value no then (I) Null Hypothesis is H_o : n = po and (ii) Alternative Hypothesis may be (a) H, : p = n_o (i.e, p > M_o or M < M_0) or (b) H, : p > n_o or (c) H, : M < M_0- Thus, there can be more than one alternative hypothesis.

14.2.3 Type I and Type II Errors

In the process of hypothesis testing we usually come across same sort of errors, called errors in hypothesis testing which are grouped in two types as : (i) Type I Errors, and (ii) Type
II Errors.

(i) Type I Errors: Type I errors are made when we reject the null hypothesis H_o though it is true, In other words, when H_o is rejected despite its being true, then it is callcd Type I errors. The probability of making a type I error is denoted by p (E,) = a and the probability of making a correct decision is then 1 - a i.e, p * 1 - a.

(ii) Type II Errors: Type II errors are made when we accept the null hypothesis though it is false, In other words, when H_o is accepted despite its being false, then it is called type II errors. The probability of making a type II error is denoted by p, Thus p (E_2) = p.

The following table illustrates Type I and Type II errors.

	To Accept H_o	To Reject H_o
H is True 0	Correct Decision p = 1 - a	Type I Error, p = a
H is False o	Type II Error P= *P*	Correct Decision *p* = 1 - p

While testing hypothesis, attempts are made to minimize both the types of errors, although it is not at all possible to reduce them both at the same time.

14.2.4 Level of Significance

The refers to the degree of significance with which we accept or reject a particular hypothesis. Since 100% accuracy is not possible in taking a decision over the acceptance or rejection of a hypothesis, we have to take the decision at a particular level of confidence which would speak of the probability of one being correct or wrong in accepting or rejecting a hypothesis. In most of the cases of hypothesis testing, such a confidence is fixed at 5% level, which implies that our decision would be correct to the extent of 95%. For a greater precise, however, such a confidence may be fixed at 1% level which would imply that the decision would be correct to the extent of 99%. This level is usually denoted by the symbol,

a (alpha) which represents the probability of committing the Type I error (i.e. of rejecting a null hypothesis which is true). The level of confidence for significane), is always fixed in advance before applying the test procedures, It is important to note that if no level of significance isgiven, then we always take a = 0.05.

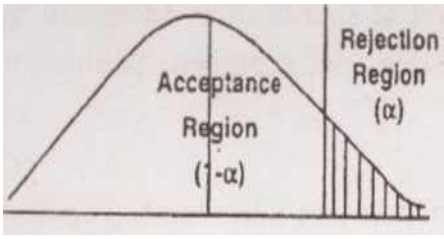
14.2.5 Critical Region or Rejection Region

The critical region or rejection region is the region of the standard normal curve corresponding to a pre-determined level of significance. The region under the normal curve which is not covered by the rejection region is known as Acceptance Region. Thus, the statistic which leads to the rejection of null hypothesis H_o gives us a region known as Rejection Region or Critical Region, While those which lead to acceptance of H_o give us a region called as Acceptance Region.

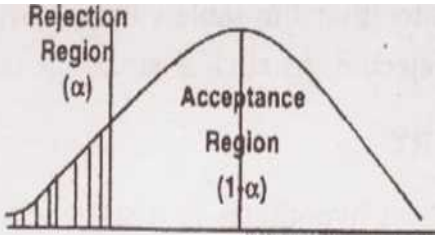
14.2.6 Two Tailed Test and One Tailed Test

A test of any statistical hypothesis where the alternative hypothesis is expressed by the symbol (<) or the symbol (>) is called a one tailed test since the entire critical region lies in one tail of the distribution of the test statistic. The critical region for all alternative hypothesis containing the symbol (>) lies entirely on the right tail of the distribution while the critical region for an alternative hypothesis containing a less than (<) symbol lies entirely in the left tail. The symbol indicates the direction where the critical region lies. A test of any statistical hypothesis where the alternative is written with a symbol *=' is called a two-tailed test, since the critical region is split in to two equal parts, one in each tail of the distribution of the test statistic, The following figures illustrate one'tailed and two tailed tests :

Right-Tailed Test

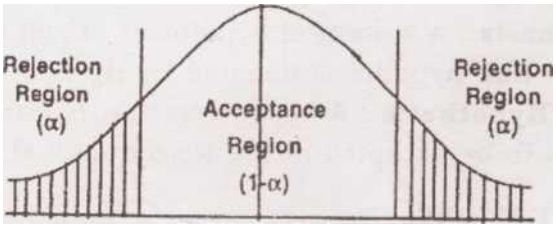


Left-Tailed Test



Two Tailed Test

(L.O.S.ct)



14.2.7 Critical Value

The critical values of the standard normal variate (Z) for both the two-tailed and one

tailed tests at different level of significance are very often required in hypothesis testing. The following table gives critical values for both one failed and two failed tests at various level : significance

Level of Significance (a)	a = 0.10	a = 0.05	a = .01	a = 005
Critical values of Z (for one tailed test)	- 1.26 or + 1.26	- 1.645 or + 1.645	- 2,33 or + 2.33	- 2.56 or + 2.56
Critical values of Z (for two tailed test)	- 1.645 or + 1.645	- 1.96 or + 1,96	- 2.58 or + 2.56	- 2.61 or + 2.81

14.2 Self Check Questions 1

a.What is Statistical Hypothesis?
1. A proven fact about population
2. A guess about the sample information
3. An assumption or statement about population parameters based on sample information
4. A statement about the accuracy of sample data

b. What may be the nature of Statistical Hypothesis?
1. Always True
2. Always False
3. May or may not be true
4. Constantly Changing

c. What is the Null Hypothesis(Ho)?
1. A proven fact about the population
2. An assumption about the sample
3. The hypothesis under test, assumed to be true
4. An alternative hypothesis

d. What is Type I Error in Hypothesis testing?
1. Rejecting the Null Hypothesis when it is true
2. Accepting the Null Hypothesis when it is true
3. Rejecting the Alternative Hypothesis when it is true
5. The confidence level for accepting or rejecting a hypothesis

e. The Null Hypothesis is denoted by the symbol_____.

f. The level of significance os often fixed at_____% representing the probability of committing Type I Error.

14.3PROCEDURE OF TESTING A HYPOTHESIS:

Testing of a hypothesis passes through the following steps :

(1) Set up a null hypothesis: It is denoted H_0. Null hypothesis assumes that difference between any values to be compared is not significant.

(2) Set up a suitable level of significance: A suitable level of significance is determined to test the null hypothesis. In practice, 5% significance level is used.

(3) Set up a suitable test of statistic : A number of test statistics like Z, t, y^2, F etc. may be applied to test the null hypothesis. It is decided only on the basis of available information.

(4) Doing necessary calculation: After selecting appropriate statistic, computations relating to the test statistic are made and values are worked out.

(5) Making Decisions: In the process of hypothesis testing, results are interpreted at the final stage. For this purpose, we compare the computed value of a test statistic with the table value at a predetermined level of significance. If computed value is greater that the table value of 5% or 1% level of significance, then null hypothesis is rejected. In such a situation the sample does not represent population.

14.3 Self Check questions 2

a. Which step involves selecting an appropriate statisitc sich as Z, t, y2, F, etc.?

1. **Step 1. Set up a null hypothesis**
2. **Step 2. Set up a suitable level of significance**
3. **Step 3. Set up a suitable test statistic**
4. **Step 4. Doing necessary calculations**

b. The null hypothesis assumes that there is a significant difference between values being compared.

1. **True**
2. **False**

14.4 SUMMARY

A statistical hypothesis is a statement about a population parameter. The statement is tentative in the sense that it carries some belief or involves an assumption that may or may not be found valid on verification. When the act of verification is done on basis of sample evidence, testing is called statistical testing and the hyothesis tested is known as statistical hypothesis.

14.5 GLOSSARY

- **Null Hpothesis** : A statistical hypothesis stated with a view to testing its validity is called as a null hypothesis denoted by H_0.
- **Alternate Hypothesis** : A hypothesis that is framed in opposite to null hypothesis, and comes to be accepted at the Rejection of H_0 is called Alternate Hypothesis.

14.6 SHORT ANSWER QUESTIONS

Q.1 What do you understand by Null and Alternate Hypothesis?

Q.2 State the procedure in detail for testing of a hypothesis, citing example.

14.7 LONG ANSWER QUESTIONS

Q.1 List some test statisitcs that may be applied to test the null hypothesis

Q.2 Explain the steps involved in testing a hypothesis

Q.3 Discuss the importance of the level of significance in hypothesis testing

14.8 ANSWERS TO SELF CHECK QUESTION

14.2 a. 3

b. 3

c. 3

d) 1

e) Ho

f) 5%

14.3 a) 3

b) False

14.9 SUGGESTED READINGS

1. *Statistics* by Croxton and Cowden
2. Statistical Methods by S.P. Gupta
3. Fundamental of Mathematical Statistics by S.C. Gupta and V.K. Kapoor
4. *Statistical Analysis* by T.L. Kaushal.

LESSON 15 **HYPOTHESIS TESTING : CHI-SQUARE TEST AND T-TEST**

STRUCTURE

15.0 OBJECTIVES

After reading this chapter, the reader should be able to :

- Understand the importance and significance of tests of association and goodness of fit.
- Understand the importance and significance of students' t-distribution or t-tests.
- Practically apply & evaluate the concepts of chi square and t-tests.

15.1 INTRODUCTION

A marketing manager wants to know the market potential for his company's new product in Western and Southern India, so that he can decide in which region to launch the product first. So he has test marketed the product, in 25 supermarkets each in both the regions for a period of three months. The results of the test marketing are that the Western region has registered mean monthly sales of 25,000 units and the Southern region has registered mean monthly sales of 26,000 units. The marketing manager is unsure as to whether the sales in the Western region are significantly higher than those in the Southern region- enough to decide on launching the product first in the Western region.

In a pre-poll survey (with 2 samples consisting of 900 urban and 800 rural youth respectively) undertaken by a media company, Party A was preferred by 43% of the urban youth and also by 47% of the rural youth. The analyst at the company doesn't know whether difference in the proportions is statistically significant so as to state with confidence that the rural youth are more inclined towards party A than the urban youth.

The above examples give some idea of the kinds of dilemmas commonly faced by statistical researchers. In these situations, researchers use hypothesis tests to analyze the assc ::ation or differences between means and proportions. The hypothesis testing :edures can be classified into two key types - tests of association, and tests of differences. Tests of association are used in situations where researchers evaluate the statistical relationship between the variables. Tests of differences are concerr.e: -.t - making judgments regarding the differences between populations.

15.2 CHI-SQUARE TEST

Tests of association are used in situations where the researcher has to evaluate whether there is any association between the variables under study. For example, researchers face situations where they want to know whether there is any association between brand preference and income levels, or whether the inflation in an economy and the stock market index are related. Prominent tests of association are the Chi-square test, correlation analysis and regression analysis. In order to evaluate the statistical significance of association among the variables involved in the cross-tabulation, researchers use statistical technique called the Chi-Square test. The Chi-square test is usually used by the researchers in two ways - test of independence and test of goodness of fit. The test of independence is used to evaluate whether there is any association between two variables. The goodness of fit test is used to identify whether there is any significant difference between the observed frequencies and the expected frequencies.

15.2.1 CHI-SQUARE TEST-GOODNESS OF FIT

In some situations researchers would like to see how well the observed frequency pattern will fit into the "expected frequency'" pattern. That is the researcher wants to know whether there is any significant difference between the observed frequencies of a particular variable and the expected frequencies of that variable. In such cases the Chi-square test is used to test whether the fit between the observed distribution and the expected distribution is good. If the value of Chi-square, denoted by / is less than the tabular value corresponding to level of significance then we deduce that the observed data corresponds to the expected data. It implies that there is no difference between the observed data and the expected data. However if the value / is greater than the tabular value corresponding to level of significance then we deduce that the fit between the observed data and expected data is not good. The procedure for this test is as follows:

Step 1 : Formulate the null and alternate hypothesis

Step 2 : Calculate the expected values

Step 3 : Calculate the / using the formula $f=Z(O_i-E.)^2/E_i$

Step 4 : Decide upon the level of significance and degrees of freedom

Step 5 : Determine the critical value and compare with calculated %² value.

Step 6 : Deduce the business research conclusion

The following example will provide a clearer understanding of the procedure.

Consider an e-commerce site which plans to test the effectiveness of three different advertisements in the print media. Each advertisement will be run for a month in various national newspapers. Now the marketing communications manager wants to know the impact of these advertisements on the level of hits received by the web site in three different months. The number of hits received in each month and the corresponding advertising program is given in Table 15.1.

Table 15.1

Advertising Program	**Month**	**Hits Received**
L	January	8700
II.	February	9100
III.	March	8780

The manager wants to know whether there is any difference between the hits received during each month the advertising was run.

Step 1 : Formulate the hypotheses

Null hypothesis can be formulated as - there is no significant difference between the number of hits received each month when different advertising programs were run. In other words tile numbers of hits received are same in all three months. The alternate hypothesis can be formulated as - there is a significant difference between the number of hits received each month when different advertising programs were run.

Thus, H_0: Number of hits received is the same in each month.

Hj : There is a significant difference between number of hits received during each month when different advertising programs are run.

Step 2 : Determine expected values

Expected frequencies are those values (hits) that are expected when the null hypothesis is true. As stated above null hypothesis in this case is that the number of hits for each month is not significantly affected by advertising. Hence the expected value for each month will be same i.e. 8860 as calculated below.

The expected frequencies ($E_{()}$) = Total value * number of categories

= Total hits received + number of months = 26,580 * 3 = 8860

Step 3 : Calculate the y^2 using the formula

$X^2 = 1(0; - E.)^2 / Ej$

$= ((8700 - 8860)^2 + 8860) + ((9100 - 8860)^2 + 8860) + [(8780 - 8860)^2 + 8860]$

= 2.88 + 6.50 + 0.72 = 10.1 Step 4 : Decide upon the level of significance and degrees of freedom

Then company has to decide upon the level of significance (a) at which it wants to perform the x^2 test- Let us set the level of significance at 10% The degrees of freedom (v) is calculated using the following formula: v = k-1

where, v = degrees of freedom, k = number of categories

In this case v = 3 - 1 = 2 Step 5 : Determining the critical value and comparing it with the calculated y^2 value

The critical value against which the Calculated x^2 is to be compared can be obtained by looking at tile corresponding tabular x^2 value for 10% level of significance with 2 degrees of freedom. The critical value for 10% level of significance with 2 degrees of freedom is 4.60. The calculated x^2 value 10.1 from step 3 is greater than the critical value.

Step 6 : Deduce the business research conclusion

As the calculated x^2 value 10.1 > 4.60, the null hypothesis is rejected. Thus, there is a significant difference between the number hits received in each month which are covered by different advertising programs.

This test can be used to find out whether there is a considerable change in the dependent variable on the whole, but it fails to ascertain tile changes in each variable.

15.2.2 CHI-SQUARE TEST - TEST OF INDEPENDENCE

Apart from testing the "goodness of fit" for a problem that involves one variable, the Chi-square test can also be used to evaluate the relationship between two or more variables This test is also known as the Chi-square test of independence. This is useful when analyzing cross-tabulations. These tests help determine whether there is any significant association between the variables involved in tile research problem. Consider the data of Table 15.2 where channel viewership is segregated according to age groups.

Table 15.2

Age group / Channel	**Channel A**	**Channel B**	**Channel C**	**Total**
15-25	20	30	30	80
25-45	80	70	50	200
45 years and above	60	40	20	120
Total	**160**	**140**	**100**	**400**

From Table 15.2 we observe that 25-45 age group respondents prefer Channel A compared to other channels. But it is not certain whether this observation is representative of the entire population or whether it is due to the sampling error. This dilemma can be resolved by subjecting the data to a Chi-Square test. The procedure followed in this test is similar to the "goodness of fit" test.

Step 1 : Formulate the null and alternate hypotheses.

Step 2 : Calculate the expected values.

Step 3 : Calculate the chi-square using the following formula x^2 = I Z (O, - E/ / E_y Where, (X represents the observed frequency in i* row and j^{lh} column

E. . represents the expected frequency in i^{1}* row and j'^{h} column *Step 4 :* Decide upon the level of significance and degrees of freedom.

Step 5 : Determine the critical value and compare it with the calculated chi-square value.

Step 6 : Deduce the business research conclusion.

Step 1 : Formulate the null and alternate hypotheses

The null hypothesis Ho can be stated as 'there is no association between the two variables. The alternate hypothesis H. can be stated as there is an association between the two variables.

The null and alternative hypotheses for the above example would be as given below. H_0 : There is no association between the age group and channel viewership Hj : There is significant association between the age group and channel viewership *Step 2 :*

Calculate the expected values

Calculation of expected values enables the researcher to decide whether null hypothesis is true or false.

Expected frequencies are the values that are expected when the null hypothesis is true. In the Table 15.2 we can see that out of total sample of 400 respondents, the viewership figures for channel A, Channel B; and Channel C irrespective of age group are 160, 140, and 100. In other words viewership figures for Channel A, Channel B and Channel C are in proportion of 160: 140: 100. If there is no influence of age group on the television channel viewership, (i.e., if the null hypothesis is true), then for each age group the viewership figures should be in a similar proportion. From Table 15.2, we find the number of respondents in the 15-25 age group is 80. If the age group variable does not have any relationship with the channel viewership, then the proportion of viewership figures for all the three channels in the 15-25 age group will be in line with the overall proportion of the sample i.e., 160: 140: 100. On the other hand, if the null hypothesis is not true, then the proportion of viewership figures for 15-25 group may not be in line with the overall proportion of the sample.

The expected frequency value for each category (age group) can be calculated using the formula

$E_{ij} = (n * n) / n$

Where, E represents expected frequency of a cell, corresponding to a particular age group and a particular channel, n represents the row total n represents the column total n represents the total sample size.

So the expected frequencies for the 15-25 age group are :

Channel A : (80 * 160) + 400 = 32 Channel B : (80 * 40) + 400 = 28

Channel C : (80 * 100) + 400 = 20

Similarly we can derive the expected values for other age categories. Table 15.3 provides the observed frequencies and corresponding expected frequencies in brackets for all the three age groups.

Table 15.3

Age group / Channel	Channel A	Channel B	Channel C	Total
15-25	20 (32)	30 (28)	30 (20)	80
25-45	80 (80)	70 (70)	50 (50)	200
45 years and above	60 (48)	40 (42)	20 (30)	120
Total	160	140	100	400

Step 3 : Calculate the chi-square value

The next step in the test of independence is the calculation of the test statistic. That is calculation of Chi-square. It is calculated using

X^2 = II $(O_{ij} - E)^2 / E_{ij}$..

The Chi-square test statistic is calculated as follows :

X^2 = ((20 - 32)² + 32] + [(30 - 28)² + 28] + [(30 - 20)² + 20] + [(80 - 80)² + 80] + [(70 - 70)² * 70] + ((50 - 50)² + 50] + [(60 - 48)² + 48] + [(40 - 42)² + 40] + [(20 - 30)² + 30]

= 16.08

Step 4 : Decide upon the level of significance and degrees of freedom :

Then company has to decide upon the level of significance (a) at which it wants to perform the Chi-square test. Let us set the level of significance at 1%. The degree of freedom is calculated using the following formula : v = (n - 1) * (k - 1)

Where v = degrees of freedom, n = number of rows, k = number of columns. Therefore v = (3 - 1) * (3 - 1) = 2 * 2 = 4

Step 5 : Determine the critical value and compare it with the calculated chi-square value.

The next step is to compare the test statistic with the critical value. The critical value can be obtained by looking at Chi-Square distribution table for 4 degrees of freedom at 1 level of significance. The value so obtained is 13.28.

Step 6 : Deduce the business research conclusion

The calculated value is greater than tabular value, i.e., 16.08 > 13.28. Thus the null hypothesis (there is no association between the age group and the channel viewership) is rejected. This implies that the channel viewership is significantly (1 % level) dependent on age group.

15.2 Self Check Questions

a. What is the Chi-Square test used for in the context of association between variables?

1. Measure of central tendency

2. Test of independence and goodness of fit

3. Regression analysis

4. Correlation analysis

b. In which situations would a researcher use the goodness of fit test with the Chi-square test?

1. Assessing correlation between variables

2. Identifying central tendency in data

3. Evaluating differences between observed and expected frequencies

4. Testing association between two variables

c. Reseachers use the Chi- square test in two ways: the test of independence and the test of goodness of_____.

d. The Chi-square test is a measure of cental tendency.

1. True

2. False

e. The Chi-square test is not suitable for evaluating the significance of association between brand preference and income levels.

1. True

2. False

f. The Chi-square test for goodness of fit assesses the fit between the observed distribution and the _____ distribution.

g. If the Chi-sqaure value(/) is greater than the tabular value at a given level of significance, it implies there is a significant difference between observed and expected data.

1. True

2. False

15.3 HYPOTHESIS TESTING WHEN THE POPULATION STANDARD DEVIATION IS NOT KNOWN

Another situation that researchers face in hypothesis testing relates to a single mean when the population standard deviation is not known and the sample standard deviation is known. When the population standard deviation is not known, the size of the sample is considered while selecting the statistical test to be used. Thus for research problems involving large samples (>30), and for a known population standard deviation, the z-test is used. If the sample size is less than 30 and the population standard deviation is not known and we need to test the hypothesis based on the sample standard deviation, we should use the t - distribution test. The procedure followed in the t-test is similar to z-test. Let us discuss the various steps involved in hypothesis testing about single mean when the standard deviation is not known.

Example : A Company, assume that it has conducted the market survey on sample of 28 customers and the mean sample price is determined as Rs 20,500. The population standard deviation is unknown and the sample standard deviation was identified as Rs 3,100. At 1% level of significance, can the company conclude that the average price is less than the hypothesized Rs. 21,000?

The problem can be described as follows :

Sample size n = 28

Sample mean X = Rs 20,500

Sample standard deviation S_x = Rs 3,100

Level of significance a = 0.01

First we need to formulate the hypotheses. As company wants to evaluate that quotation with unit price at Rs. 21,000 is competitively priced, the hypotheses can be formulated as follows :

H_0 : μ = Rs. 21, 000

H, : μ < Rs. 21,000

The next step is to decide on the statistical test to be used and determine the critical values. As the sample size is less than 30 and the population standard deviation is not known we use t-test to test the hypothesis. The level of significance is set at 1%. Thus the acceptance region will be 99% and the rejection region will be 1%. Though the t-distribution table is similar to the z-distribution table, mere is difference in the way distribution table has been derived. While the z-distribution table considers the confidence level as the basis, t- distribution table considers the level of significance as the basis. Another aspect in the t- distribution table is tile degrees of freedom. The number of degrees of freedom in t -test is measured as

$$v = n - 1$$

where, v = degrees of freedom, n = sample size.

Thus in the present case the degrees of freedom is 27 or (28 - 1). The level of significance is 0.01%.
So the critical value can be obtained by looking at the corresponding t-value in table
at 27 degrees of freedom under .01 column of one-tailed test. The value is 2.473.

The next step is to calculate the standard error of the mean. As the population standard deviation is not known, we use the sample standard deviation as an estimate to the population standard deviation. Thus,

$$o = S_x$$

As we are calculating me standard error of the mean using the sample standard deviation, the standard error of me mean will also be an estimate. Thus the standard error of the mean is calculated as follows :

$$\text{Standard error} = S_x / (n)^{1/2}$$
$$= 3100 / (28)\ 1/2 = 585.84$$

Then we need to calculate the standardized sample mean or observed t-value. The standardized sample mean can be obtained using the following formula

$$t = [(X) - M_0] / \text{standard error}$$

where, X = sample mean

H_0 = hypothesized mean so, t = (20500 - 21000) / 585.84 = -0.85

As it is a right tailed test, we reject the null hypothesis when the standardized sample mean or observed t value is greater than the critical value, and fail to reject the null hypothesis when tile standardized sample mean or observed t value is lesser than the critical value. As the observed t value -0.85 is less than 2.473, we fail to reject the null hypothesis that the price quoted by the company is competitive.

15.3.1 TESTS FOR DIFFERENCES BETWEEN MEANS: SMALL SAMPLES

When sample sizes of two samples are less than 30, procedure to test the hypothesis will differ on two aspects. One aspect will be that the t-test is used instead of the z-test. Another aspect is that determining standard error of the difference between means of two samples will be different from the procedure followed for large sample tests. Let us understand the process using an example.

A marketing manager wants to evaluate the monthly sales revenue generated by one of its newly launched product in two different regions - region A, and region B. He studied the sales revenue pattern for a month in select super markets in both the regions. The findings of the study are shown in Table 15.4.

Table 15.4

	Region A	**Region B**
Sample size	14	17
Mean weekly sales (in Rs.)	6065	5990
Sample standard deviation	805	370

Based on this survey can the company conclude that the sales will be more in region A compared to region B?

A manager should do hypothesis testing before arriving at any conclusion regarding the above problem. The procedure is similar to tile procedure we followed in large samples. *Step 1 :* **Formulate the hypothesis**
We can state hypothesis in the following way :

Null hypothesis Ho: |a, = n_2 i.e., there is no difference between sales in two regions Alternate hypothesis H, : (i_1 > \x_2 i.e. average sales in region A axe higher compared to average sales in region B.

Step 2 : Select the appropriate statistical test

As the sample size is less than 30 and population standard deviation is not known we can use the t-test to lest tile hypothesis.

Step 3 : Calculate the sample error and standardize the sample statistic

As stated earlier the procedure for calculating standard error and standardizing the sample statistic differs from the earlier method. We can recall that when population standard deviation is not known, we calculate the standard error.

However, that formula is not appropriate for small samples tests. Thus we use the following procedure to calculate the standard error. We need to assume that the unknown population variances are equal (i.e., $a,^2 = cr_2^2$).

Instead of using Oj^2 and cT_2^2 in the calculation of standard error, we use weighted average of both these values; here the weights represent the degrees of freedom of each sample. The estimate so obtained is called the pooled estimate. It is given by :

$$S_P^2 = K n_i - 1)s_i^2 + (n_2 - 1)s_2^2) / (n_i + n_2 - 2)$$

Using the pooled estimate value, we can calculate the estimated error of the difference between two sample means (when sample size is small and population standard deviation is not known). It can be calculated using the formula given below :

Standard error = sp * [(1 /n,) + (l/n_2)]$^{1/2}$ After calculating tile standard error we need to standardize tile difference between the sample means. This can be done using the following formula : t = {[(X,) - (X_2)] - (M, - n_2)| / standard error So, substituting the values provided in the table 15.4, we get S_p^2 = [13 * (805)^2 + 16 * (370)^2] / (14 + 17-2)

= 366025

We now calculate the estimated error of the difference between two sample means (when sample size is small and population standard deviation is not known)

$$\text{Standard error} = (366025)^{1'2}\ [(1/14) + (1/17)1*$$

= 217.8

Using the standard error, wc standardize the difference between the sample means by substituting above calculated values in t = [(6065 - 5990) - 0] / 217.8 = 0.34

Step 4 : Determining the critical values

We may recall that in the single mean tests involving a single sample we used the formula

v = n - 1 for determining the degrees of freedom.

However in this case two samples are involved, thus degrees of freedom can be determined using the following formula : v = n, + n_2 - 2

Substituting the values of n, and n_2 in the above equation we get v = 14 + 17 - 2 = 29

The level of significance is fixed at 5%. The test is right-tailed. Hence the accepted region consists of 0.95 of the area under the distribution curve and the rejection region is 5 : -.he nght tail of the distribution curve. The value of degrees of freedom is 29. So the ahue can be obtained by looking the corresponding t-value in table at 29 degrees of -- : — _r.der 0.05 column of one tailed test. The value is 1.699.

Szep 5 : Compare the standardized sample statistic with the critical value

The standardized difference of the two means value 0.34 is less than the critical value 1.699.

Step 6 : Deduce business research conclusion

We reject the null hypothesis if the standardized difference of two means value is greater than the critical value. As the standardized difference of two means value is less than :he critical value (0.34 < 1.699), the marketing manager fails to rej ect the null hypothesis that there is no difference between the sales of the product between the two regions.

15.3 Self check Questions

A. What statistical test is used when the population standard is not known, and the sample size is less than 30?

1. Z-test
2. Correlation test
3. T-test
4. Chi-square test

B. If the standardized difference of two means is greater than the critical value,the null hypothesis is rejected.

1. True
2. False

15.4 SUMMARY

- The Chi-square test is usually used by the researchers in two ways - test of independence and test of goodness of fit. The test of independence is used to evaluate whether there is any association between two variables. The goodness of fit test is used to identify whether there is any significant difference between the observed frequencies and the expected frequencies. Hypothesis testing about a single mean will evaluate whether there is any difference between the sample statistic and the hypothesized population mean. While comparing the sample with a known population mean, the researcher faces two situations: hypothesis testing when the population standard deviation is known and hypothesis testing when the population standard deviation is not known. In the former case z-test is used to test the hypothesis and in the latter case t-test is used when the sample size is less than 30.

15.5 GLOSSARY

- **Chi-Square Test** : The Chi-square test is used to researchers in two ways: as a test of independence and as a test of goodness of fit. While the test of independence is used to evaluate whether there is any association between two variables, the goodness of fit is used to identify whether there is any significant difference between the observed frequencies and the expected frequencies.
- **t-test** : Hypothesis testing about a single mean will evaluate whether there is any difference between the sample statistic and the hypothesized population mean. While comparing the sample with a known population mean, the researcher faces two situations: hypothesis testing when the population standard deviation is known and hypothesis testing when the population standard deviation is not known. In the former case z-test is used to test the hypothesis and in the latter case t-test is used when the sample size is less than 30.
- **Tests of differences about two means** : There are three cases in the problems that involve testing differences between two means - testing of differences between two means for large samples, testing of differences between two means for small samples and testing differences between two means for dependent samples. In the first case z-test is used whereas where small samples are used t-test is used to solve the problem.
- **Degrees of Freedom**: The number of values or quantities that can be assigned independently in a statistical distribution.

15.6 Short Question Answers

1. What is the purpose of the Chi-square test in hypothesis testing?

2. Explain the concept of degrees of freedom in the context of t-tests.

3. Describe the steps involved in hypothesis testng for a single mean when the standard deviation is not known.

4.Define degrees of freedom and explain its significance in t-test.

15.7 Long Question Answers

1. Elaborate the Chi-square test and its two primary application.

2.Examine difference between hypothesis testing for large and small sample sizes.

3. Explain the importance of confidence levels in hypothesis testing, outling how they influence the determination of critical regions and the acceptance or rejection of null hypotheses.

4.From the data given below about the treatment of 250 patients suffering from a disease, state whether the new treatment is superior to the conventional treatment.

(Given for degree of freedom = 1, chi-square 5% = 3.84)

Treatment / No. of patients	Favorable	Not Favorable	Total
New	140	30	170
Conventional	60	20	80
Total	200	50	250

5.Two types of drugs were used on 5 and 7 patients for reducing their weight. Drug A was imported and drug B was indigenous. The decrease in the weight after using the drugs for six months was as follows :

Drug A	10	12	13	1 1	14		
Drug B	8	9	12	14	15	10	9

Is there a significant difference in the efficacy of the two drugs? (Given, for degrees of freedom=10, t at 5% = 2.228)

15.8 Answers to Self Check Questions

15.2 a. 2

b. 3

c.Fit

d. False

e. False

f. expected

g. True

15.3 a. 3

b. True

15.9 SUGGESTED READINGS

- G. C. Beri, *Business Statistics,* Tata McGraw-Hill Publishing Co. Ltd., New Delhi, 2nd Edition.
- J. K. Sharma, *Business Statistics,* Pearson Education, New Delhi, 3rd Reprint, 2005.
- Murray R. Spiegel, *Statistics - Schaum's Outline Series,* McGraw-Hill, 3rd Edition.
- Richard I. Levin & David S. Rubin, *Statistics for Management,* Prentice-Hall of India, New Delhi, 7th Edition

BSRM 201

BUSINESS STATISTICS & RESEARCH METHODOLOGY

MBA-Distance Education (First Year) **AUTHOR : SANDEEP S. VIRDI**

Sesnester- 2nd

Lesson No. 16

ANALYSIS OF VARIANCE (ANOVA)

STRUCTURE

16.0 OBJECTIVES

After reading this chapter, the reader should be able to :

- Understand the importance and assumptions for Analysis of Variance.
- Perform the procedure for testing of hypothesis using One-way and Two-way ANOVA techniques.
- Describe the components of variations.

16.1 INTRODUCTION

In previous lessons we introduced hypothesis testing procedure to test the significance of differences between two samples to understand whether the means of 2 populations are equal based upon 2 independent random samples. In all these cases, the null hypothesis states that there is no significant difference among population mean that is H_0 : * |i_2. However, there may be situations were more than 2 populations are involved and we need to test the significance of differences between 3 or more sample means. We also need to test the null hypothesis that three or more populations for which independent samples are drawn have equal :r homogenous) means against the alternative hypothesis that population means are not equal.

The following are a few examples involving more than 2 populations where it is necessary to conduct a comparative study to arrive at a statistical inference :

- Effectiveness of different promotional devices in terms of sales.
- Quality of a product produced by different manufacturers in terms of an attribute.
- Production volumes in different shifts in a factory.

- : :eld for plots of land due to varieties of seeds, fertilizers, and cultivations methods.

Lr.der certain assumptions, a method known as **analysis of variance** or **ANOVA** developed by **R.** A. Fisher is used to test the significance of the difference between several populations means. It measures the overall variation within a sample drawn from populations; finds the variation between the sample means; combines these to calculate a single test static; and then uses this to carry out a hypothesis test in same way as discussed before. Basicallv, it consists of classifying and cross classifying statistical results and testing whether means of a specified classification differ significantly. In this way it is determined whether the given classification is important in affecting the results.

16.2 ASSUMPTIONS FOR ANALYSIS OF VARIANCE

The following assumptions are required for analysis of variance :

1. The data is quantitative in nature and each population has a normal distribution
2. The population from which the samples are drawn have equal variances, that is $c_T,^2 = c_{T2}^2 = a_3^2 = .. a_k^2$ for k population, i.e. homogeneity of variances.
3. Each sample is drawn randomly and is independent of other samples.

16.2 Self Check Questions

a. What is the primary nature of the data required for analysis of variance?

1.Qualitative
2.Quantitative
3.Categorical
4.Ordinal

16.3 ANALYSIS OF VARIANCE APPROACH

The first step in the analysis of variance is to partition the total variation in the sample data into the following two component variations in such a way that it is possible to estimate the contribution of factors that may cause variation :

1. The amount of variation **among the sample means** or the variation attributable to the differences among the sample means. This variation is due to assignable causes.
2. The amount of variation **within the sample observations.** This difference is considered due to chance causes or experimental (random) zero.

The observations in the sample data may be classified according to **one factor (criterion)** or **two factors (criteria).** The classification according to one factor and two factors are called **one-way classification** and **two-way classification** respectively.

16.3 Self Check Question

a. In the analysis of variance,what is the variation among the among the sample means attributed to?

1. Assignable causes
2. Random causes
3. Experimental errors

4. Population differences

b. The variation within the sample observation is considered to be due to chance causes in the analysis of variance

1. True
2. False

16.4 TECHNIQUES OF ANALYSIS OF VARIANCE

16.4.1 ONE WAY CLASSIFICATION

In one way classification, data is classified according to only one criterion. The null hypothesis is :

$$H_0 : n, = n_2 = n_3 = \ldots\ldots\ldots = n_k$$

$^{H}i : Hi * ^2 * ^3 * — \ldots\ldots\ldots\ldots\ldots\ldots *$

The following are the steps in finding out one-way analysis of variance :

1. **Calculate the variance between the samples** : The variance between the samples (groups) measures the differences between the samples mean of each group and the overall mean weighted by the number of observations in each group. The variance between the samples taken into account the random variations from observation to observation. The sum of squares between the samples is denoted by SSC. For calculating variance between the samples we take the total of the square of the deviations of the means of various samples from the grand average and divide this total by the degrees of freedom. Thus the steps in calculating variance between the samples will be :
   (a) Calculate the mean of each sample, i.e. X,, X_2 etc
   (b) Calculate the grand average X, "pronounced as X double bar". Its value is obtained as follows :

$$= \_ \frac{x, + x, + x, + \ldots.}{n, + n, + n_3 + \ldots.}$$

   (c) Take the difference between the means of the various samples and the grand average.
   (d) Square these deviations and obtain the total which will give the sum of the squares between the samples, and
   (e) Divide the total obtained in (d) by the degrees of freedom. The degrees of freedom will be one less than the number of samples i.e. if there are 4 samples then the degrees of freedom will be 4-1 = 3 or v = k - 1, where k = number of samples.
2. **Calculate the variance within the samples :** The variance (or sum of squares) within the samples measures those inter-sample differences due to chance only. It is denoted by SSE. The variance within the samples measures variability around the mean of each group. Since the variability is not affected by group differences it can be considered a measure of the random variation of values within a group. For calculating the variance within the samples we take the total of the sum of the squares of the deviation of various items from the mean values of the respective samples and divide this total by the degrees of freedom. Thus the steps in calculating variance within the samples will be:
   (a) Calculate the mean value of each sample i.e. X,, X_2, X, etc

(b) Take the deviations of the various items in a sample from the mean values of the respective samples;

(c) Square these deviations and obtain the total which gives the sum of square within the samples and

(d) Divide the total obtained in step (c) by the degrees of freedom. The degree of freedom is obtained by the deduction from the total number of items the number of samples, i.e. v = N - K, where K refers to the number of samples and N refers to the total number of al the observations.

3. Calculate the ratio as follows :

$$F = \frac{\text{Between Column Variance}}{\text{Within Column variance}}$$

The variance between the samples means is the numerator and the variance within the samples is the denominator.

4. Compare the calculated value of F with the table value of F : The table value of F is at a certain degrees of freedom at a certain critical level (generally taken as 5%). If the calculated value of the F is greater than the table value, it is concluded that the difference in sample means is significant i.e. it could not have arisen due to fluctuations of simple sampling, or in other words, the samples do not come frbm the sample population. On the other hand, if the calculated value of F is less than the table value, the difference is not significant and has arisen due to fluctuations of simple sampling.

Rationale of the test : the variation within the samples i.e. the variation of the individual observations within the samples from their own individual sample means, measures the influence of the chance forces which cause the individual observations to vaiy from one another.

Illustration 1. XYZ Traders wishes to test whether its three salesmen A, B and C tend to make sales of the same size or whether they differ in their selling ability as measured by the average size of their sales. During the last week there have been 14 sales calls - A made 5 calls, B made 4 calls and C made 5 calls. Following are the weekly sales record of the three salesmen :

Salesman A (in Rs)	Salesman B (in Rs)	Salesman C (in Rs)
300	600	700
400	300	300
300	300	400
500	400	600
0	—	500

Perform the analysis of variance and draw your conclusions.

Solution :

Let us take the hypothesis that the sales of the three salesman are of the same size or they do not differ. In order to simplify calculations let us divide each value by 100, so the coded data becomes as follows :

Salesman A (in Rs)	Salesman B (in Rs)	Salesman C (in Rs)
3	6	7
4	3	3

3	3	4
5	4	6
0	- -	5

Mean of A = —, Mean of B = —, Mean of C = —

O T j

Grand Mean, $X = \frac{3+4+5}{3} = 4$

Variance within the Samples

x,	**(X.-X,)²**	**X₂**	**(x₂-x₂p**	*3	**(x₃-x₃r**
3	0	6	4	7	4
4	1	3	1	3	4
3	0	3	1	4	1
5	4	4	0	6	1
0	9	—	—	5	0
IX, = **15**	**I(X,-X,P** = **14**	VO II	**W₂-x₂P** = 6	**IX₃** = **25**	**I(X₃-X₃?** = **10**

Total sum of squares with samples = 14 + 6 + 10 = 30 v = 14 - 3 = 11

Variance between Samples

Mean of A = —, Mean of B = —, Mean of C = —

{x,-x)'	**(x₂-x)²**	**(x₃-x)²**
1	0	1
1	0	1
1	0	1
1	0	1
1	—	1
l(x,-x) =5	/— =\² X 1 X II 0	/— =\² **l(x₃-x) =5**

Variance between samples = 5 + 0 + 5= 10 v = 3 - 1 = 2 **Analysis of Variance Table**

Sources of variation	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**Variance Ratio**
Between	2	10	5	
Within	11	30	2.73	**1.83**
Total	**13**	**40**		

For v, = 2 and v_2 = 11 $F_{0\ 05}$ = 3.98

The calculated value of F is less than the table value.

Hence, the hypothesis holds true.

We, therefore, conclude that the three salesmen do not differ in their selling activity as measured by the average of their sales.

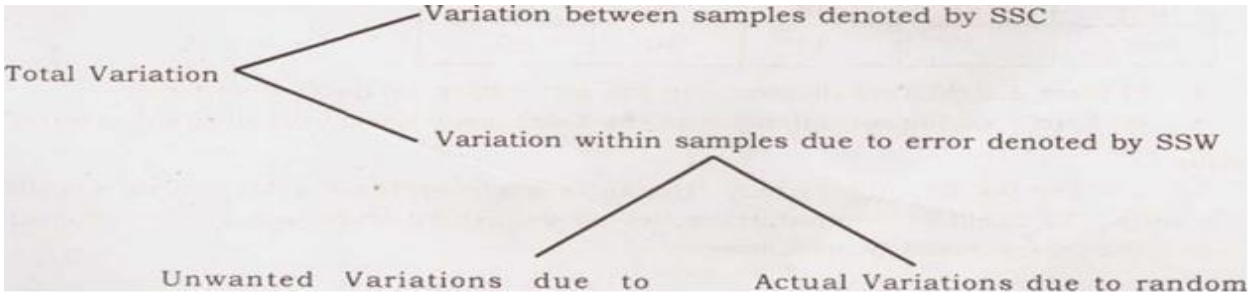
16.4.2 TWO WAY CLASSIFICATION

In one-way ANOVA classification, we illustrate the partitioning of the total variation in the sample data into 2 components: variation among the samples due to different groups or treatments and variations within the samples due to random error. However there might be a possibility, that some of the variations left in the random error from one way analysis of variation was not due to the random error or chance but due to some other measurable factor.

The two-way analysis of variance can be used to :

- Explore one criterion (or factor) of interest to partition the sample data so as to remove the unaccountable variation and arriving at a true conclusion,
- Investigate two criteria (factors) of interest for testing the difference between sample means,
- Consider any interaction between two variables

To ensure a right conclusion to be reached, each sample data (group) should be measured under the same conditions by removing variations due to these conditions by the use of a blocking factor :



difference between block means - errors - sum of squares of errors

block sum of squares (SSR) (SSE)

General ANOVA table for two way classification

Sources of variation	Sum of Square	Degrees of Freedom	Mean Square	Test Statistic
Between Columns	SSC	c - 1	MSC = SSC / (c - 1)	F, = MSC / MSE
Between Rows	SSR	r - 1	MSR = SSR / (r - 1)	F_2 = MSR / MSE
Residual Error	SSE	(o-l) * (r—1)	MSE = SSE / (c- l)x(r - 1)	
Total	SST	(n - 1)		

As stated above, total variation consists of three parts :

(i) Variation between columns, SSC;

(ii) Variation between rows, SSR;

(iii) Actual variation due to residual error, SSE.

That is

SST = SSC + SSR + SSE or SSE = SST - (SSC + SSR)

The degrees of freedom associated with SST are (cr - 1), where c and r are the number of columns and rows respectively.

Degrees of freedom between columns = c - 1 Degrees of freedom between rows = r - 1

Degrees of freedom for residual errors = (cr - 1) - (c - 1) - (r - 1) = (c - 1) (r - 1)

The test-statistic, F, for analysis of variance is given by

F, = MSC / MSE; MSC > MSE or MSE / MSC; MSE > MSC F_2 = MSR / MSE; MSR > MSE or MSE / MSR; MSE > MSR

DECISION RULE

If $F_{cal} < F_{tabl(.}$, then accept Null Hypothesis, H_0, Otherwise reject H_0.

Here it might be pertinent to mention that Residual is the measuring rod for testing significance. It represents the magnitude of variation due to forces called chance.

Illustration 2. The following table gives the number of refrigerators sold by 4 salesmen in 3 months May, June and July.

Month	**Salesmen**			
	A	**B**	**C**	**D**
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

- Is there a significant difference in the sales made by the four salesmen?
- Is there a significant difference in the sales made during the different months?

Solution :

Let us take the hypothesis that "The sales made by the 4 salesmen do not differ significantly". To facilitate computations, let us deduct 40 from each given value, after deduction the values would be as follows :

Is there a significant difference in the sales made by the four salesmen?

- Is there a significant difference in the sales made during the different months?

Solution :

Let us take the hypothesis that "The sales made by the 4 salesmen do not differ significantly". To facilitate computations, let us deduct 40 from each given value, after deduction the values would be as follows :

Month	**Salesmen**				
	A	**B**	**C**	**D**	**Total**
May	10	0	8	-1	17
June	6	8	10	5	29
July	-1	4	0	-1	2
Total	15	12	18	3	48

Correction Factor i.e. C.F. = — = ---------- = 192
N 12

Sum of squares between salesmen :

$$\frac{(ix_1)^2}{(IX_3)^2} + \frac{(ix_2)^2}{(IX_4)^2} \quad \frac{}{N} \quad \frac{}{N} \quad \frac{}{N} \quad \frac{}{N}$$

$$(15)^2 + (12)^2 + (18)^2 + (3)^2$$

2 3 3 3 =

75 + 48 + 108 + 3

- 192 = 42

Sum of squares between months :

$$= \frac{(17)^2}{3} + \frac{(29)^2}{4} + \frac{(2)^2}{4}$$

= 72.25 - 210.25

+ 1 - 192 = 91.5 Total

Sum of squares :

= $(10)^2 + (6)^2 + (-1)^2 + (0)^2 + (8)^2 + (4)^2 + (8)^2 + (10)^2 + (0)^2 + (-1)^2 + (5)^2 + (-1)^2$-

192 = 408 - 192 = 216

ANALYSIS OF VARIANCE TABLE

Sources of Variation	Sum of Squares	v	Mean Sum of Squares
Between salesmen	42.0	3	42 / 3 = 14.00
Between months	91.5	2	91.5 / 2 = 45.75
Residual	82.5	6	82.5 / 6 = 13.75
Total	216	11	

14

F (Salesman) = $\frac{3}{10.7}$ = 1.018

The table value of F at 5% for v, = 3 and v_2 = 6 is 4.76. Since the calculated value less than the table value, the hypothesis holds true.

Hence the sales ma'de by the four salesmen do not differ significantly.

F (Months) = $\frac{45.75}{13.75}$ = 3.33

The table value of F at 5% for v, = 2 and v_2 = 6 is 5.14. Since the calculated value is less than the table value, the hypothesis holds true.

Hence the sales made during the different months do not differ significantly.

16.4 Self Check Questions

a. In one-way classification, data is classified according to:

1. Two criteria
2. Only one criterion
3. Three criteria
4. Random criteria

b. What does SSC stand for in the context of one-way analysis of variance?

c. The degree of freedom for variance between the samples in one-way ANOVA are calculated as_____.

16.5 SUMMARY

In one way ANOVA we illustrate the partitioning of the total variation in the sample data into two components, variation among the samples due to different groups or treatments and variations within the samples due to random error. There is also a probability that some of the variation left in the random error from one way analysis of variance was not due to random error or chance but due to some other measurable factor. Consequently, F value would then be small and responsible for the rejection of null hypothesis. In two way analysis of variance, another term called blocking factor' is introduced to remove the undesirable accountable variation. The two-way analysis of variance can be used to explore one criterion (or factor) of interest to partition the sample data so as to remove the unaccountable variation and arriving at a true conclusion, or investigate two criteria (factors) of interest for testing the difference between sample means and consider any interaction between two variables.

16.6 GLOSSARY

- **Analysis of Variance** : A statistical procedure for determining whether the means of several different populations are equal.
- **Blocking** : The removal of a source of variation from the error term in the analysis of variance.
- **One Way Analysis of Variance** : Analysis of variance in which only one criterion (variable) is used to analyze the difference between more than 2 population means.
- **Treatment** : Different levels of a factor.
- **Two Way Analysis of Variance** : Analysis of variance in which two criteria (variables) are used to analyze the difference between more than 2 population means.

16.7 SHORT QUESTIONS

1. State one assumption necessary for analysis of variance regarding the distribution of data.

2. What are the two component variations that the total variation in sample data is partitioned into during the analysis of variance?

3. Define the term “Degrees of Freedom” in the context of one-way ANOVA.

16.8 LONG QUESTIONS

1. Explain the step-by-step process of calculating the variance between the samples in one-way classification.

2. Explain various techniques of Analysis of Variance in detail.

16.9 ANSWERS TO SELF CHECK QUESTION.

16.2 a. 2

16.3 a. 1

b. TRUE

16.4 a. 2

b. "SUM OF SQUARE B/W THE SAMPLE"

c. $v=k-1$, where k is the number of samples

16.10 SUGGESTED READINGS

- Sharma, J. K.; **Business Statistics,** Pearson Education, New Delhi, 2005, 3rd Reprint.
- Gupta, S. P.; Statistical Methods, Sultan Chand & Sons, New Delhi, 2007.

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS & RESEARCH METHODOLGY: BSRM 201

Lesson No.17 **AUTHOR : SHILPI GOYAL**

SPSS AND REPORT PRESENTATION

STRUCTURE

17.0 OBJECTIVES

After reading this chapter, the student should be able to :

- Understand the steps used in manipulating data with SPSS.
- Identify how to use SPSS in certain quantitative statistical calculations and help in better presentation of research report.

17.1 INTRODUCTION

Researchers differ in the way they prepare a research report. Research report is considered a major component of the research study for the research task remains incomplete till the report has been presented and /or written. The purpose of research is not well served unless the findings are made known to others. Research results must invariably enter the general store of knowledge. All this explains the significance of writing research report. Writing of report is the last step in a research study and requires a set of skills somewhat from those called for in respect of the earlier stages of research. This task should be accomplished by the researcher with utmost care; he may seek the assistance and guidance of experts for this purpose. Now a days, there are many statistical packages

available to help the researcher to carry out a detailed analysis and present the research report in a better way. SPSS is one of the most commonly used packages.

17.2 SPSS - STATISTICAL PACKAGE FOR SOCIAL SCIENCES

SPSS stands for Statistical Package for the Social Sciences. SPSS is a statistical and data management package for analysts and researchers. It provides us with a broad range of capabilities for the entire analytical process. SPSS is a full-range package that provides all levels of statistical analysis, data manipulation, graphical representation and report writing, among other features. With SPSS, we can generate decision-making information quickly using powerful statistics, understand and effectively present our results with high-quality tabular and graphical output, and share our results with others using a variety of reporting methods, including secure web publishing. Results from our data analysis enable us to make smarter decisions more quickly by uncovering key facts, patterns, and trends. SPSS is available for Windows only. Programs and data created and used in SPSS can be transported and used in SPSS running on other platforms, though some modifications may need to be made in programs written under one version and used under a different version.

SPSS is the statistical package most widely used by political scientists. The several reasons why :

1. Force of habit: SPSS has been around since the late 1960s.
2. Of the major packages, it seems to be the easiest to use for the most widely used statistical techniques;
3. One can use it with either a Windows point-and-click approach or through syntax (i.e., writing out of SPSS commands
4. Many of the widely used social science data sets come with an easy method to translate them into SPSS; this significantly reduces the preliminary work needed to explore new data.

There are also two important limitations that deserve mention at the outset :

1. SPSS users have less control over statistical output than, for example, Stata or Gauss users. For novice users, this hardly causes a problem. But, once a researcher wants greater control over the equations or the output, she or he will need to either choose another package or learn techniques for working around SPSS limitations;
2. SPSS has problems with certain types of data manipulations, and it has some built in quirks that seem to reflect its early creation. The best known limitation is its weak lag functions, that is, how it transforms data across cases. For new users working off of standard data sets, this is rarely a problem. But, once a researcher begins wanting to significantly alter data sets, he or she will have to either learn a new package or develop greater skills at manipulating SPSS.

Overall, SPSS is a good first statistical package for people wanting to perform quantitative research in social science because it is easy to use and because it can be a good starting point to learn more advanced statistical packages.

17.2 Self Check Questions

a. What does SPSS stand for?

b.SPSS has been around since the late_____.

c.SPSS users may face limitations in controlling statistical output compared to packages like_____or_____.

d. Which of the following is limitation of SPSS in terms of statistical output control?

1. SPSS offers more contraol than other statistical packages.

2. Users have less control compared tp other packages like Stata or Gauss.

3. SPSS has the same level of control as other packages.

4. Control over output.

17.3 STARTING UP SPSS

SPSS is usually part of the general network available in the computer labs and residence halls of most college campuses. To activate SPSS, sign-on to the network with your username and password. Then click the **Start** icon in the lower left-hand corner of the screen followed by **Network Programs > Academic Applications > SPSS for Windows.** If SPSS is not found on the "Network Programs" group, it may be installed as a "local program" in which case the proper sequence is **Start > Programs > SPSS for Windows** > [If you encounter an "Empty" button at this point look down the list for another "SPSS for Windows."] **SPSS 10.0 for Windows.** Another possibility is that a shortcut already exists on the desktop, in which case double-clicking it will open the SPSS program.

17.4 THE SPSS PROGRAM

After clicking the SPSS icon, there is a short wait and the SPSS program appears. SPSS 9.0 for Windows begins with two windows. The top window offers several options which may be useful eventually, but the easiest thing to do is close the top window which then gives access to the main program. At this point one of the most sophisticated and popular data analysis programs is available. Thanks to a user-friendly interface, it is possible to do almost anything from the simplest descriptive statistics to complex multivariate analyses with just a few clicks of the mouse. The program is also quite "smart" in that it will not execute a procedure unless the necessary information has been provided. Although it can be frustrating when working with complex procedures, it saves a lot of time in the long run because the feedback is immediate and corrections can be made on the spot.

17.5 DATA IIf PUT FOR SPSS

SPSS appears on the screen looking like most other Windows programs. Two windows are initially available: the data input window and the output widow. When SPSS first comes up, it is ready to accept new data. To begin entering data, look at the menu options across the top of the screen:

File Edit View Data Transform Analyze Graphs Utilities Window Help

Clicking on one of these options opens a menu of related options, many of which will not be available until enough information has been provided to allow the procedure to run. To begin the process of a computing a correlation in SPSS Release 9.0, click on the **Data** option, then click on **Define Variable.** This will open an input window that allows you to define the first variable by giving it a name and other information that will make it easier to use the variable

in statistical analyses and interpret the output. When this window is opened the default name for the variable is displayed and highlighted. Just type a name for the first variable that uses less than eight characters. For example, the first variable in the above example could be called **colgpa,** a name that is less than eight characters and gives a good indication of the nature of the variable (college GPA). It is also useful to have more information about the variable and this can be done by clicking on the **Labels...** button which appears at the bottom of the window. This button opens another window which allows you to add more information about the variable, including an extended label, such as **College GPA for 1999.** You can also add what are called **value labels** using this same window. Value labels allow you to give names to particular values of a nominal or categorical variable. For example, most studies have a variable called Sex that can take on two values, 1 = Female or 2 = Male. The value labels option allows you to have these labels attached to all the output from statistical analyses which simplifies interpretation and reporting. Entering value labels also means you don't need to remember how the variable was coded (i.e., whether males were coded 1 or 2) when you view the output. After entering a variable name and value labels for the first variable, close the Labels... input window by clicking the Continue button.-Then click the OK button on the Define Variable window. The next step is to use the mouse and left mouse button or the arrow keys to reposition the cross to the first cell in the second column of the data input spreadsheet. Then define the second variable using the same process. Continue defining variables until all the variables have been defined. Release 10.0 has a different approach to entering infomation about the variables, ^t the bottom of the SPSS window are two tabs, Data View and Variable View. If you are using Release 10.0, click on the Variable View tab and another spreadsheet will appear. This spreadsheet contains detailed information about each variable. To enter new information double-click on the first box in the "Name" column and type in the name you are assigning to the variable (limited to 8 characters). Other information about the variables can be entered by adding or changing the information in the other columns. An extended, more informative, variable label can be added to the "Label" column and value labels can be added using the "Values" column. To enter value labels, click on the box corresponding to the variable and then click the grey button. A window for adding variable labels will then appear.

Once the variables have been defined, the data can be entered into the spreadsheet. (These tasks can be done in the opposite order, as well.) This requires working with the New data or spread sheet window. Release 10.0 requires clicking the **Data View** tab at the bottom of the screen to bring up this view. To begin, make sure the cursor is flashing at the top of the spreadsheet window and that the upper left cell of the spreadsheet is highlighted. To highlight a cell use the mouse to move the cross to the desired cell of the spreadsheet and click the left mouse button. The arrow keys also work well to navigate around the spreadsheet. Now begin entering data by typing the first piece of data. In the above example the first entry would be **1.8.** This number will appear at the top of the spreadsheet. Hit **<ENTER>** to move the data into the correct cell. Notice that after hitting **<ENTER>** the second cell in the first column is now highlighted. The next piece of data **(3.9)** can be entered using the same procedure. Thus, data is automatically entered vertically. Continue until all the data for the first column have been entered. After entering all the data for the first column, use the mouse or arrow keys to highlight the first cell in the second column and begin entering the second column of data using the same technique. If a piece of data is missing (e.g., the participant did not answer one or more of the questions on a survey), simply hit **<ENTER>** when the input cell at the top of the spreadsheet is empty. This will cause a dot to appear in the spreadsheet cell which is interpreted by SPSS as missing data. SPSS has very flexible options for handling missing data. Usually, the default or standard option is the best one to use.

In larger studies with a lot of variables, it may be more convenient to go across or horizontally, entering all the data for the first participant followed by all the data for the second participant, continuing until all the data have been entered. In order to do this it will be necessary to make more frequent use of the mouse and left button or the arrow keys to highlight the next cell going across. When data for a large study is being entered, it is best to work with a partner. One person can read the data and the other can type. This greatly increases speed and accuracy.

17.6 ADVANCED DATA ENTRY AND FILE HANDLING

Sometimes a researcher has data stored in a text file that was created manually or by optical scanning. A text file may be in either fixed width format (each variable aligned in fixed- width columns) or delimited format (each variable delimited or separated by a specific character such as a comma or tab). Data in a fixed width format consists of a constant number of lines for each participant or case with each case beginning on a new line. Each variable should be in the same location for each participant (i.e., biological sex could be coded in the fifth column of the second line for each participant). Missing data should be designated by blanks so the remaining data remains aligned in the correct location, and for convenience in setting up the SPSS file. Data stored in the delimited format has a specific character (comma, semicolon, tab, etc.) separating each variable for each participant.

To convert text data into an SPSS file, click **File > Read Text Data,** select the data file to be read, click the Open button, and then follow the six steps of the SPSS Test Import Wizard.

Step 1 asks whether the format is one that has been used previously to read data into an SPSS file. If this is the first time you have used the Text Import Wizard, click the No button and then Next.

Step 2 requires identifying the type of file format, fixed or delimited, and whether variable names are at the top of the file. After clicking the appropriate buttons, click Next.

Step 3 allows you to indicate which line of the text file is the first line of the data allowing exclusion of any file labels or other extraneous identifiers at the beginning of the text file. If the data are in fixed column format, indicate the number of lines per case and the number of cases to import.

In Step 4, follow the directions to indicate where each variable is located by inserting break lines at the appropriate positions. Note that this needs to be done once for each line of data that makes up a case.

In Step 5 variable names and formats are assigned.

In Step 6, the file and syntax may be saved. The procedure is similar when the delimited format has been used.

Data may also be captured or imported from database programs by following the steps under **File> Database Capture.**

17.6 Self Check Questions

a. How can a researcher convert text data into an SPSS file?

1. Click File> Save
2. Click File> Read Text Data
3. Click Edit> Convert
4. Click Tools> Import Data

b. In a delimited format, each variable is separated by a specific character.

1. True

2. False

c. What are the specific character that separated each variable in a delimited format?

17.7 COMPUTING THE PEARSON CORRELATION

After entering the data, the next step is to order the program to actually compute the correlation coefficient for you. Use the mouse to go to the top of the screen and click on the following sequence : **Analyze > Correlate > Bivariate.** This will open another input window. You will see two boxes with the one on the left containing the complete list of variables for the study. [Note: The variables will appear in alphabetical order which is the default variable display. However, it can often be more convenient to display the variables in the same order as they appear on the spreadsheet or input window. The display order can be changed by clicking **Edit > Options.** Then change "Alphabetical" to "File" by clicking the empty circle next to File' under "Variable Lists." Unfortunately, SPSS will need to be exited and then reloaded before this option will take effect.) The box on the right will be empty. In between the boxes is a right-pointing arrow. The sequence for computing a correlation is to highlight variables from the list on the left and then use the mouse to click the right-pointing arrow. This will cause each highlighted variable to jump to the box on the right. Each variable in the box on the right will be included in the correlation matrix computed by SPSS. Thus, in order to compute the correlation between COLGPA and STUDYHRS, move both variables over to the box on the right. A variable can be removed from the box orf the right by highlighting it and clicking the arrow in the middle which will now face in the opposite direction. Once the variables you want to correlate are in the right-hand box, the OK button could be clicked which would cause the correlations to be computed and appear in an Output window. However, there are a couple of additional points worth considering.

F:rs: :*. :ar. re extreme!;, helpful t: click the Options button which appears at the bottom of the input window. This will cause another input window to appear. Generally, all options can be left on their default settings. However, one option allows you to print means and standard deviations for each variable in the analysis by just clicking the box. This is worth doing. The other options should be left alone unless you have a specific reason for changing one. At this point you must click the Continue button in order to close this box and move on with your task. The next step is simply to click the OK button. After a short delay, an Output window will appear with the results of your analysis. The information in the output file can be viewed or saved to a disk using standard Windows conventions. Additional analyses can be performed and their results will be appended to the end of the current output window so the results of a complex series of analyses can be contained in one output window. Be sure to give this file a name that will remind you of its contents. The results for the correlation example are shown below :

Descriptive Statistics

	Mean	Std. Deviation	N

COLGPA	2.9455	.7285	11
STUDYHRS	23.8182	13.4001 . * .	11

CORRELATIONS

		COLGPA	STUDYHRS
COLGPA	Pearson Correlation	1.0000	.868**
	Sig. (2-tailed)	·	.001
	N	1 1	1 1
STUDYHRS	Pearson Correlation	.868**	1.0000
	Sig. (2-tailed)		.001
	N	11	1 1

**. Correlation is significant at the 0.01 level (2-tailed).

To interpret the output, look at the table labeled Correlations. This is a correlation matrix with three numbers for each correlation. The top number is the actual Pearson correlation coefficient which will range from -1.00 to +1.00. The further away the correlation is from zero, the stronger the relationship. The correlation between study hours and college GPA in this fictional study was .868 which represents an extremely strong relationship. The next number is the probability. Remember, you are looking for probabilities less than .05 in order to reject the null hypothesis and conclude that the correlation differs significantly from a correlation of zero. The third number is the sample size, in this case 11. Correlation coefficients that can not be computed will be represented as a dot.

Another nicc thing to do when computing a correlation is to look at the scatter diagram. To produce a scatter-plot, click **Graphs > Scatter > Define** >. Use the same technique as before to transfer variables to the x-axis and y-axis boxes. Then click **OK** and the graph will appear in the Output window. To insert the plot in another document, click on **File > Copy,** open your word processing document, and Paste it into the document.

17.7 Self Check Questions

a. To compute a correlation in SPSS, go to the top of the screen and click_____> Correlate>Bivariate.

b. The Pearson correlation coefficient ranges from -1.00 to +1.00, and the further away from zero, the weaker the relationship.

1. True
2. False

c. In the Correlation matrix table, what does the top number represent?

17.8 SAVING OUTPUT AND DATA FILES

If you attempt to close either the data input or data output windows of SPSS, the program will respond w'ith another window prompting you to save the file with either a user-supplied name or a generic name. Output files are given the extension, .spo, and data files are given the extension, .sav. The usual Windows conventions with respect to saving and reopening files apply using commands under the File menu.

17.9 OTHER HELPFUL FEATURES OF SPSS

There are a number of additional features available in SPSS that can be extreme:;. helpful for the beginning researcher. These features will be described briefly.

17.9.1 Transformations

Two particularly valuable features are available from the Transform menu: Recode, and Compute. The purpose of a recode is very simple. Imagine a variable that is coded from 1 to 5. Sometimes extreme values are not selected by very many individuals. Thus, it may be desirable to combine individuals who responded with either a 4 or a 5 into a single category such as 4. The recode feature is the way to do this. Another situation that often calls for a recode is when a variable is part of a scale but the scoring needs to be reversed before it can be added to other items to make a total score. This is also accomplished with the recode command. To do a recode, click Transform > Recode > Into Same Variables... or Into Different Variables... and enter the required information. The choice of recoding into the Same or Different Variables is a question of whether it is desirable to preserve the old data. By doing the recode into a Different variable, the old data can be preserved in case a mistake is made or another recoding procedure is tried.

The Compute... command is also under the Transformations menu. This command allows the researcher to construct an equation for changing the scale of a variable. The main usefulness of the procedure is for remedying the situation where the raw data do not meet statistical assumptions. A transformation using the Compute... command can often bring the data back into conformity with statistical assumptions. Most statistics books have a discussion of the various common types of transformations and their potential benefits. The Compute... option can also be used for computing new variables. For example, a new variable can be computed that represents the sum of scores on several other variables. This feature is useful when adding the scores on individual items of a test or personality measure to obtain a total score.

17.9.2 Exploratory Data Analysis

Exploratory data analysis is a process of carefully examining data prior to performing inferential statistical tests. Access to exploratory data analysis techniques can be obtained by clicking Analyze > Descriptive Statistics > Explore... which leads to plots (boxplots; stem and leaf) and descriptive statistics that can help greatly in the early stages of data analysis. Distributions can also be tested for normality.

17.9.3 Help Features

The Help menu provides access to information about specific Topics, a Tutorial, a Statistics Coach, and other useful features. It is also possible to click the right mouse button while pointing to a term of interest which will result in a display of the definition of that term. The dialog or input boxes also have context-specific Help buttons.

17.9.4 Reliability Analysis

Reliability is one of the most important characteristics of good psychological measures. To compute the standard measure of internal consistency, coefficient alpha, click Analyze > Scale > Reliability Analysis.... The variables that make up the scale to be analyzed should be transferred to the Items box. Then click the Statistics... button and request all the descriptive statistics plus the inter-item correlations.

17.9.5 Moving Output to Other Applications

Often, it will be desirable to move output to another application such as a \#ord- processing file. This operation will prove especially useful in research methods courses. To do this, copy the table, chart, or plot that you wish to move using the Edit menu. Then open up the target application (for example a word processing file into which you would like to copy the item) and select Paste or Paste Special... from the Edit menu. This is the simplest method of including SPSS output in another file. There are other methods which enable you to update the table or chart with SPSS. You can learn more about these processes by searching the Help files in SPSS. Transferring information from one program to another is often one of the most difficult tasks to accomplish with modern technology so it is wise to seek consultation when difficulties are encountered.

17.9 Self Check Questions

a. Which menu in SPSS provides access to features such as Recode and Compute?

1. Analyze

2. Transformations

3. Explore

4. Help

b. What is the purpose of exploratory data analysis in SPSS?

1. Conduct inferential statistical tests

2. Examine data prior to statistical tests

3. Perform reliability analysis

4. Move output to other applications

c. What is the purpose of the Recode feature in SPSS?

d. Exploratory data analysis in SPSS includes techniques like reliability analysis.

1. True

2. False

17.10 SUMMARY

These directions in this lesson describe some basic analyses and point the way toward more advanced procedures. Very complex analyses can also be easily performed with the help of this package. . The best way to learn the advanced features of SPSS for Windows is to explore the program using data from an original study. SPSS is much faster, convenient, and accurate than computing the analysis by hand.

17.11 GLOSSARY

- **Report** - A structured written presentation directed to interested readers in response to some specific purpose, aim or request.
- **SPSS** - A full-range package that provides all levels of statistical analysis, data manipulation, graphical representation and report writing, among other features.

17.12 SHORT ANSWER QUESTIONS

1. Explain various advantages of SPSS.

2. What are the various limitations of SPSS.

3. Explian fixed width format and delimited format in text files.

17.13 LONG ANSWER QUESTIONS

1. Explain the significance of SPSS in preparation of a research report.

2. Discuss the steps used in operating SPSS to compute various statistical tools.

17.14 ANSWERS TO SELF CHECK QUESTIONS

17.2 a. STATISTICAL PACKAGE FOR SOCIAL SCIENCES

b. 1960's

c. Stata and Gauss

d. 2

17.5 a. 2

b. True

c. Comma "," Semicolon ";" Tab etc

17.7a. Analyze

b. False

c. Pearson Correlation Coefficient

17.9 a. 2

b. 2

c. To modify

d. False

17.15 SUGGESTED READINGS

- SPSS Instruction Manual
  www.idrc.ca/en/ev-56466-201-1-DO-TOPIC.html

MBA-Distance Education (First year)

Semester-2

BUSINESS STATISTICS & RESEARCH METHODOLGY: BSRM 201

Lesson No. 18

AUTHOR : SANDEEP S. VIRDI

REPORT WRITING

STRUCTURE

18.0 OBJECTIVES

After reading this chapter, the reader should be able to :

- Understand the importance and need for Report writing.
- Types of reports and the guidelines for writing a report.
- The fundamental structure of a Business Research Project Report.

18.1 INTRODUCTION

As part of the research proposal, the sponsor and the researcher agree on what types of reporting will occur both during and at the end of the research project. Depending on the budget for the project, a formal presentation may not be part of the reporting. A research sponsor, however, is sure to require a written report and a poorly presented report can destroy a research project. This fact prompts researchers

to make special efforts to communicate clearly and effectively their findings, analysis of findings, interpretations, conclusions and recommendations.

18.1.1 SIGNIFICANCE OF REPORT WRITING

Research report is one of the vital aspects of research and is considered a major constituent of the research study, for the research task remains incomplete till the report has been presented and / or written. As a matter of fact even the most brilliant hypotheses, highly well designed and conducted research study, and the most striking generalization and findings are of little value unless they are effectively communicated to others. The purpose of research is not well served unless the findings are made known to others,

Writing of report is the last step in a research study and requires a set of skills somewhat different from those called for in respect of the earlier s.ages of research.

18.1.2 TYPES OF RESEARCH REPORTS

Depending on its intended audience, the research report may be either technical or popular in orientation. While both approaches describe the research study, its methodology, findings, conclusions and recommendation, they can differ considerable in terms of detail, writing style, use of technical terms and length. In general, the higher the executive status of the audience, the shorter the report will tend to be.

18.1.2.1 TECHNICAL REPORT

The technical report is generally intended for other researchers, or for research managers. The report should enable another researcher to critique methodology, check calculations and accuracy and to follow everything which is done on a step by step basis. A brief definition of every technical term should be given.

A general outline of a technical report can be as follows :

1 Summary of Results
2 Nature of the Study
3 Methods Employed
4 Data
5 Analysis of Data and Presentation of findings
6 Conclusions
7 Bibliography
8 Technical Appendices
9 Index

18.1.2.2 POPULAR REPORT

The popular report is intended for a more general audience, one that is not that conversant with the details of research methods and terminology. Compared to the technical report, the presentation will be a bit livelier with increased attention to

headlines, flow diagrams, charts, tables and occasional summaries for the purpose of stressing major points. A popular report gives emphasis on simplicity and attractiveness, practical aspects and policy implications.

A general outline of a popular report can be as follows:

1. Findings and their Implications
2. Recommendations for Actions
3. Objectives of the Study
4. Methods Employed
5. Results
6. Technical Appendices

As different kinds of audiences may be interested in the results of the same research study, it is sometimes necessary to write both a technical report and a popular report.

18.1 SELF CHECK QUESTIONS

a. What is the last step in a research study?

b. Who are the intended audience for technical report?

1. General Audience
2. Research managers and other researchers
3. Students
4. Policy makers

c.The purpose of research is not well served unless the findings are communicated to others.

1. True
2. False

18.2 GUIDELINES FOR WRITING A REPORT

Researchers who are effective in report writing agree that there are a series of guidelines which should be followed :

Such guidelines can be enumerated as under :

- **Consider the Audience** : Make the report clear; use only words familiar to the readers and define all technical terms. To make the comparison of figures easier, use percentages, rounded off figures, ranks or ratios; put the exact data in a table within the text or in the appendix. Use graphic aids (charts, graphs, pictures, etc.) wherever they help clarify the presentation of data.
- **Address the Information Needs** : Remember the research report is designed to communicate information to decision makers. Make sure that it clearly relates the research findings to the objectives of the management.

- **Be Concise, Yet Complete** : Most managers will not want to read about the details of a research report. Knowing what to include and what to leave out is a difficult task. It is up to you, the researcher, to take into account the information needs of the decision maker when writing your report.
- **Be Objective** : You will probably face at least one situation in which you know that the client will not easily accept the results. The findings may conflict with the decision maker s experience a judgment or they may reflect unfavorably on the wisdom of previous decisions. In these circumstances, there is a strong temptation to start the report by making the result more acceptable to the management. A professional researcher, however, will present the research findings in an objective manner :.e.. without b:as and will defend their validity if they are challenged by the client.
- Style : The style of writing a research report is important because it shows a way of presentation. Here are a few a tips to help you write a report that is easy to read.
  - Write in brisk, business like English
  - Use short words and sentences.
  - Be concise.
  - Use the active voice.
  - Consider appearance - space makes a long report easier to read.
  - Write in present tense.

18.2 SELF CHECK QUESTIONS

a. What is the role of graphic aids in a research report?

1. Complicate the presentation
2. Make the report lengthy
3. Clarify the presentation of data
4. Reduce the clarity of the report

b. In report writing, it is crucial to be objectibe and present the research findings without_____.

c. Including unnecessary details in a report is recommended for thoroughness.

1. True
2. False

d. Graphic aids, such as chart and graphs, can help clarify the presentation of data.

1. True
2. False

18.3 SUGGESTED STRUCTURE

Most writers agree with Robson (2002) on the general structure to adopt for a project report that is the end product of your research. This is :

- TITLE PAGE
- TABLE OF CONTENTS
- LIST OF TABLES / FIGURES / GRAPHS ETC
- EXECUTIVE SUMMARY / ABSTRACT
- INTRODUCTION
- LITERATURE REVIEW
- METHODOLOGY
- RESULTS
- LIMITATIONS
- CONCLUSIONS AND RECOMMENDATIONS
- REFERENCES
- APPENDICES

This suggested structure should not inhibit anyone from adopting something different. The precise structure to be adopted is important than the necessity for the reader to be absolutely clear what the report is saying and to meet the assessment criteria. The structure should have a logical flow. The readers should know the journey on which they are being taken, and should know at all times the point in the journey that has been reached. Above all, the structure adopted should enable the reader, having read the report, to identify the storyline clearly.

Following are the broad sections outlining their purpose and content.

(a) TITLE PAGE

The title page should contain a title which conveys the essence of the study, the data, name of the organisation submitting the report and the name of the recipient organisation.

(b) TABLE OF CONTENTS

The table of contents sequentially lists the topics covered in the report along with their page references, its purpose is to help readers find the particular sections of the report that are of most concern to them.

(c) LIST OF TABLES / FIGURES / GRAPHS ETC

This table lists the titles and the page numbers of all visual aids i.e. tables, figures, diagrams, graphs, etc. It can be placed either on the same page with the table of contents or on a separate page.

(d) EXECUTIVE SUMMARY / ABSTRACT

The Executive Summary or the Abstract is probably the most important part of the report because it may be the only part that some persons will read. It is a short summary of the complete content of the project report.

For those who intend to read the whole report the abstract prepares them for what is to come. It should contain four short paragraphs with the answers to the following questions:

5. What were my research questions, and wrhy were these important?
6. How did I go about answering the research questions?
7. What did I find out in response to my research questions?

8. What conclusions do I draw regarding my research questions?

Smith (1991) lists five principles for the writing of a good abstract. He argues that:

6. It should be short. Try to keep it to a maximum of two sides of A4.
7. It must be self-contained. It must summaries the complete content of your report.
8. It must satisfy your reader's needs. Your reader must be told about the problem, or central issue, that the research addressed and the method adopted to pursue the issue. It must also contain a brief statement of the main results and conclusions.
9. It must convey the same emphasis as the report, with the consequence that the reader should get an accurate impression of the report's contents from the abstract.
10. It should be objective, precise and easy to read. The project report contents page should give you the outline structure for the abstract. Summarizing each section should give an accurate resume of the content of the report. Do ensure that you stick to what you have written in the report. The abstract is not the place for elaborating any of your main themes. Be objective.

Writing a good abstract is difficult. The obvious thing to do is to write it after finishing the report. It is suggested that it should be drafted at the start of writing so that you have got your storyline abundantly clear in your mind. You can then amend the draft when you have finished the report so that it conforms to the five principles above.

(e) INTRODUCTION

The introduction should give the reader a clear idea abut the central issue of concern in your research and why you thought that this was worth studying. It should also include a full statement of your research question(s) and research objectives. This will give brief details of the content of each chapter and present an overview of how your storyline unfolds. This will usually be a fairly brief chapter, but it is vitally important. The nature of the ..-.rxd action is conditioned by the diversity of the audience and their familiarity with the - v ar. r. project. The more diverse the audience, the more extensive the introduction. The 'ocuction must clearly explain the nature of the decision problem and the research .j.ective. Background information should be provided on the product(s) or service(s) involved a_r.z tr.e circumstances surrounding the decision problem.

(f) LITERATURE REVIEW

The critical literature review forms the foundation on which the research is built. Its ~a:n purpose is to help develop an understanding and insight into the relevant previous research and the trends that have emerged. The main purposes of your literature review are to set your study within its wider context and to show the reader how your study supplements the work that has already been done on your topic. The literature review, “—’fore, may inform directly any specific hypotheses that your research was designed to -oe hypotheses will also suggest a particular research approach, strategy and data collection methods. Following are some of the other purposes that a literature review is supposed to accomplish :

1. To help refine further the research questions and objectives
2. To highlight possibilities that have been overlooked implicitly in

research to date

3. To discover explicitly recommendations for further research
4. To help avoid simply repeating work that has already been done
5. To sample current opinions in newspapers, professional and trade journals
6. To discover and provide an insight into research approaches, strategies and techniques that may be appropriate to your research

(g) METHODOLOGY

This should be a detailed chapter giving the reader sufficient information to make an estimate of the reliability and validity of your methods. Following table is a useful list of the points that you should include in the method chapter.

POINTS FOR INCLUSION IN THE METHOD CHAPTER

Setting

y What was the research setting?

/ Why did you choose that particular setting?

/ What ethical issues were raised by the study, and how were these addressed?

Participants / How many?

/ How were they selected?

/ What were their characteristics?

/ How were refusals/non-returns handled?

Materials

/ What tests/scales/interview or observation schedules / questionnaires were used?

/ How were purpose made instruments developed?

/ How were the resulting data analyzed?

Procedures

/ What were the characteristics of the interviewers and observers, and how were they trained?

/ How valid and reliable do you think the procedures were?

/ What instructions were given to participants?

/ How many interviews / observations / questionnaires were there; how long did they last; where did they take place?

/ When was the research carried out?

(h) RESULTS

It may well be that your report will contain more than one results chapter. The question you should ask yourself is: Is more than one results chapter necessary to communicate my findings clearly? The results chapter or chapters are probably the most straightforward to write. It is your opportunity to report the facts that your research discovered. This is where your will include such tables and graphs that will illustrate your findings (do not put these in the appendices). The chapter may also contain verbatim quotes from interviewees, or sections of narrative account that illustrate periods of unstructured observation.

There are two important points to bear in mind when writing your results. The first is to stress that the purpose is to present facts. It is normally not appropriate in this chapter to begin to offer opinions on the facts. This is for the following chapter. Many of us become confused about the difference between findings and conclusions. One way of overcoming the confusion is to draw up a table with two columns. The first

should be headed what I found out' and the second 'what judgments I have formed as a result of what I found out'. The first list is entirely factual (for example, 66per cent of respondents indicated they preferred to receive email messages rather than paper memos) and therefore the content of your findings chapter. The second list will be your judgments based on what you found our (for example, it appears that electronic forms of communication are preferred to traditional and therefore the content of your conclusions section.

The second point links to the first. Drawing up a table will lead you to a consideration of the way in which you present your findings. The purpose of your project report is to Communicate the answer to your research question to your audience in as clear a manner as possible. Therefore you should structure your findings in a clear, logical and easily understood manner. There are many ways of doing this. One of the simplest is to return to the research objectives and let these dictate the order in which you present your finding in a clear, logical and easily understood manner. There are many ways of doing this. One of the simplest is to return to the research objectives and let these dictate the order in which you present your findings. Alternatively, you may prefer to report your findings thematically. You could present the themes in descending order of importance.

(i) LIMITATIONS

The limitation section of the report presents a dilemma. Every research project has short-comings which need to be communicated in a clear and a concise manner. The purpose of this section is not to undermine the quality of the research project but rather to enable the reader to judge the validity of the research study results. The limitations in a marketing project generally involve sampling, non-response, inadequacies and methodology weaknesses etc.

(j) CONCLUSIONS AND RECOMMENDATIONS

Logically, for each finding there should be at least one conclusion. This suggests that:
the conclusions chapter should be at least as long as the findings chapter(s). This is certainly the case. It is your conclusions that will demonstrate whether you have answered the research question and show the degree of insight that you exhibit in reaching your conclusions. It is the second major opportunity in the research process to demonstrate real originality of thought. In the conclusions you are making judgments rather than reporting facts. The key questions to ask of each of the findings are : 'So what?' and importantly, 'To what extent have answered my research question(s) and met my research objective(s)?'

You may find that the clearest way to present your conclusion is to follow a similar structure to the one used in your findings section. If that structure reflects the research objectives then it should make certain that your conclusions would address the research question(s)

You may also have a final section in your conclusion chapter(s) called' discussion' alternatively you may make this a separate chapter with this general heading. Here you would turn to your conclusions and ask such questions as: What does this mean?' What are the implications for organizations?, What are the implications for the current state of knowledge of the topic?. How does it add to the literature? What are the implications for future research? The conclusions chapter should not include new material but the discussion may do so, as long as it is

germane to the point you are making about your conclusions.

Using a matrix in the planning of the content for the results and conclusion chapters

Research questions	Results (what factual information did I discover in relation to the specific research questions?)	Conclusions (what judgments can I make about the results in relation to the specific research questions?)
What are the operational differences between different shifts in the production plant?	Cases of indiscipline in the last six months have been twice as frequent on the night shift as on the day shift.	The night shift indiscipline problems may be due to the reluctance of operators to work on this shift

Figure 2

In a management report this would normally form the content of a chapter specifically devoted to recommendations. Even if you do not specify any practical implications of your research you may comment in the conclusions chapter on what your research implies for nay future research. This is a logical extension of a section in the conclusions chapter that should be devoted to the limitations of your research. These limitations may be the size of sample, the snapshot nature of the research, or the restriction to one geographical area of an organization. Virtually all research has its limitations. This section should not be seen as confession of your %veaknesses, but as a mature reflection on the degree to which your findings and conclusions can be said to be the truth.

(k) REFERENCES

A range of conventions are used to reference the material of other writers' material that you have cited in your text.

For books and pamphlets the order may be as under :

1. Name of author, last name first.
2. Title, underlined to indicate italics.
3. Place, publisher, and date of publication.
4. Number of volumes.

Example

- *Kothari, C. R.,* Research Methodology : Methods and Techniques, *New Age International Publishers, New Delhi, 2007* For magazines and newspapers the order may be as under :

1. The Name of the Author, Last Name First.
2. The Title of Article, in Quotation Marks.
3. The Name of Periodical in Italics,
4. The Volume / Issue and Number.

5. The Date of the Issue.
6. The Pagination.

Example

Rossa, Robert V., "Coping with Short-term International Money Flows", The Banker, London, Vol - 3, No, 2, September, 1971, pp. 28-30.

The above examples are just the samples for bibliography entries and may be used, but one should also remember that they are not the only acceptable forms. The only thing important is that, whatever method one selects, it must remain consistent.

1) APPENDICES

In general, appendices should be kept to the minimum. Your project report will stand or fall on the quality of the main text. However, your appendices should include a blank copy of your questionnaire, interview or observation schedule, where these have been conducted in a language different from that in which you write your submitted project report you will need to submit both this version and the translation.

The Management Report

C3

In the typical management report this may be the most important section. The hard pressed executive reading your report may turn to your recommendations first to see what action needs to be taken to tackle the issue.

Whether you include a recommendation section depends on the objectives of you research. If you are doing exploratory research you may well write recommendations, among which will be suggestions for the pursuit of further research. However, if your research is designed to explain or describe, recommendations are less likely,

Length of the Project Report

You will probably have guidelines on the amount of words your project report should contain. Reports that exceed the word limit are usually excessively verbose. It is more difficult to be succinct. Do not fall into the trap of writing a long report because you did not have the time to write a shorter one.

18.3 SELF CHECK QUESTIONS

A) In referencing books and pamphlets, what is the correct order of information?

1. Title, place, and date of publication
2. Name of author, title, place, publisher, and date of publication
3. Date of publication,title and author
4. Author's last name, title, and place of publication

B) For magazines and newspapers, what information is included in the refrence entry?

1. Author's last name, title, place, and date of publication
2. Title of article, name of periodical, volume/issue number, date of the issue, and pagination
3. Author's first name, volume/issue number, and pagination
4. Place of publication,title, and date of the issue

18.4 COMMON GRAMMATICAL ERRORS

OFTEN WE WRITE	THE CORRECT WAY IS
1. Each pronoun should agree with their antecedent.	Each pronoun should agree with its antecedent
2. Just between you and I, case is important	Just between you and me, case is important
3. A preposition is a poor word to end a sentence with	A preposition is a poor word with which to end a sentence
4. Verbs has to agree with their subject.	Verbs have to agree with their subject.
5. Do not use no double negatives	Do not use double negatives.
6. Remember to never split an infinitive.	Remember never to split an infinitive.
7. When dangling, do not use participles.	Do not use dangling participles
8. Avoid cliches like the plague	To avoid cliches like the plague.
9. Do not write a run-on sentence it is difficult when you got to punctuate it so it makes sense when the reader reads what you wrote.	Do not write a run-on sentence. It is difficult to punctuate it, so that it makes sense to the reader.
10. About data is included in this section,	What about sentence fragments?
11. The data is included in this section.	The data are included in this section

Figure 3

18.5 CHECKLIST FOR EVALUATING THE FIRST DRAFT

- Is there a clear structure?
- Is there a clear storyline?
- Does your abstract reflect accurately the whole content of the report?
- Does your introduction state clearly the research question(s) and objectives?
- Does your literature review inform the later content of the report?
- Are your methods clearly explained?
- Have you made a clear distinction between findings and conclusions in the two relevant chapters?

- Have you checked all your references and presented these in the required manner?
- Is there any text material that should be in the appendices or vice versa?
- Does your title reflect accurately your content?
- Have you divided up your text throughout with suitable headings?
- Does each chapter have a preview and a summary?
- Are you happy that your writing is clear , simple and direct ?
- Have you eliminated all jargon?
- Have you eliminated all unnecessary quotations?
- Have you checked spelling and grammar?
- Have you checked for assumptions about gender?
- Is your report in a format that will be acceptable to the assessing body?

18.6 ORAL PRESENTATION

At times oral presentation of the results of the study is considered effective, particularly in cases where policy recommendations are indicated by project results. The merit of this approach lies in the fact that it provides an opportunity for give and take decisions which generally lead to a better understanding of the findings and their implications But the main demerit of this sort of presentation is the lack of any permanent record concerning the research details and it any be just possible that the findings may fade away from people's memory even before an action is taken. In order to overcome this difficulty, a written report may be circulated before the oral presentation and referred to frequently during the discussion. Oral presentation is effective when supplemented by various visual devices. Use of slides, wall charts and blackboards is quite helpful in contributing to clarity and in reducing the boredom, if any. Distributing a board outline, with a few important tables and charts concerning the research results, makes the listeners attentive who have a ready outline on which to focus their thinking . This very often happens in academic institutions where the researcher discusses his research findings and policy implications with others either in a seminar or in a group discussion.

18.7 PRECAUTIONS FOR WRITING RESEARCH REPORTS

Research report is a channel of communicating the research findings to the readers of the report. A good research report is one which does this task efficiently and effectively. Following are some of the precautions which can be kept in view while writing a research report :

1. While determining the length of the report, one should keep in view the fact that it should be long enough to cover the subject but short enough to maintain interest.
2. A research report as far as possible should not be dull, and try to sustain interest of the reader throughout.
3. Abstract terminology and technical jargon should be avoided in a research report and the report should be able to convey the matter as simply as possible
4. Readers are often interested in acquiring a quick knowledge of the main findings and as such the report must provide a ready availability of the

findings. For this purpose the charts, graphs and the statistical tables may be used for the various results in the main report in addition to the summary of the findings.

5. The layout of the report should be well thought out and must be appropriate and in accordance with the objective of the research problem.
6. The report should be free from grammatical mistakes and errors, and must be prepared strictly in accordance with the techniques of composition of report writing such as the use of quotations, footnotes, documentation, proper punctuation and use of abbreviations in footnotes and the like
7. The report must present the logical analysis of the subject matter. It must reflect a structure wherein the different pieces of analysis relating to the research problem fit well
8. A research report should show originality and should necessarily be an attempt to solve some intellectual problem.

r Towards the end, the report must also state the policy implications relating to the problem under consideration.

11 Appendices should be enlisted in respect of all the technical data in the report.

1 1 Bibliography of sources consulted is a must for a good report and as such must be prepared and appended at the end.

12 Index is also considered as essential part of a good report and as such must be prepared and appended at the end.

13. Report must be attractive in appearance, neat and clean, whether typed or printed.

Calculated confidence limits must be mentioned and the various constraints experienced in conducting the research study may also be stated in the report.

15. Objective of the study, the nature of the problem, the methods employed and the analysis techniques adopted must all be clearly stated in the beginning of the report in the form of introduction.

18.9 GLOSSARY

- **Report** : An account or statement to relate, as to what has been learned by observation or investigation.
- **Executive Summary / Abstract** : A short summary of the complete content of the project report.
- **Literature Review** : Text that helps develop an understanding and insight into the relevant previous research ar.i the trrr.is that have emerged.
- **Reference** : A note in a pobbcab— referring the reader to another passage or source, footnote used to direct a rraracT eis«r*here fcr additional information.

- **Bibliography** : A list of source materials that are used or consulted in he preparation of a work or that are referred to in the text, which includes the description and identification of the editions, dates of issue, authorship, and typography of books or other written material.
- **Appendix** : Supplementary material at the end of a book, article, document.

18.8 SUMMARY

Research report is one of the vital aspects of research and is considered a major c:nstituent of the research study, for the research task remains incomplete till the report has been presented and / or written. Writing of report is the last step in a research study and requires a set of skills somewhat different from those called for in resrect :f the earlier stages of research. Research report is a channel of communicating tie 'tsearch findings to the readers of the report. A good research report is one which

this task efficiently and effectively. A research report as far as possible should not be ar.d should try to sustain interest of the reader throughout. The layout of the icpavt should be. well thought out and must be appropriate and in accordance with the ifcy nur of the research problem. The report must present the logical analysis of the sahiect matter. It must reflect a structure wherein the different pieces of analysis ■rr tc the research problem fit well.

18.10 SHORT ANSWER QUESTIONS

1. What is the significance of research report writing in research process?

2. What are the different types of research reports? Explain.

3. Discuss various tips for maintaing good writing style in research report writing.

18.11 LONG ANSWER QUESTIONS

1. What are the Precautions to be kept in mind or the Errors which can creep in, while writing report?

2. Explicate in with illustrations, the structure of a Business Research Project

18.12 ANSWERS TO SELF CHECH QUESTIONS

18.1 a. Report Writing
b. 2
c. True

18.2 a. 3
b. Bias
c. False
d. True

18.3 a. 2
b. 2

18.13 SUGGESTED READINGS

- Cooper, Donald R. and Schindler, Pamela S.; Business Research Methods, Tata McGraw Hill, New Delhi, 2007, 9th Edition.
- Bryman, Alan and Bell, Emma; Business Research Methods, Oxford University Press, New Delhi, 2006, 1st Indian Edition.
- Kothari, C. R.; Research Methodology - Methods and Techniques, New Age International Publishers, New Delhi, 2007, Revised 2nd Edition.
- Bhattacharya, Dipak Kumar; Research Methodology, Excel Books, New Delhi, 2006, 2nd Edition.
- Saunders, Mark; Lewis, Philip and Thornhill, Adrian; Research Methods for Business Students, Pearson Education, New Delhi, 2004, 3rd Edition.