Research Methodology

DCOM408/DMGT404

Edited by:
Dr. Anand Thakur

LOVELY PROFESSIONAL UNIVERSITY

RESEARCH METHODOLOGY

Edited By

Dr. Anand Thakur

Printed by
EXCEL BOOKS PRIVATE LIMITED
A-45, Naraina, Phase-I,
New Delhi-110028
for
Lovely Professional University
Phagwara

SYLLABUS

Research Methodology

Objectives: The general objective of this course is to introduce students to methods of research. The specific objectives are: To develop understanding of the fundamental theoretical ideas and logic of research; To develop understanding of the issues involved in planning, designing, executing, evaluating and reporting research; To introduce students to many of the technical aspects of how to do empirical research using some of the main data collection and analysis techniques.

DCOM408 RESEARCH METHODOLOGY

Sr. No.	Description
1	**An Introduction to Research:** Meaning, Process, Defining, Research Problem: Selection, Understanding, Necessity of defined problem, Research Design , need and types of Research Design
2	**Sampling Design:** Steps of Sampling design, Characteristics of good Sampling Design, Different types of Sampling Design.
3	**Measurement and Scaling Technique:** Tools of Sound Measurement, Techniques Of Developing Measurement Tools, Scaling meaning and Important Scaling Techniques
4	**Data Collection:** Primary (Interview, Observation and Questionnaire and Collection of Secondary Data
5	**Data Analysis-1:** Measure for Central Tendency, Dispersion
6	**Data Analysis-2:** Correlation And Regression Analysis , Time series and index number
7	**Hypothesis Testing:** Hypothesis Definition and Formulation, t test , z test , ANOVA
8	**Multivariate Analysis-1:** Classification, Important Methods of Factor analysis, factor analysis , rotation in Factor Analysis, overview of cluster analysis
9	**Multivariate Analysis-2:** Discriminant analysis, multi-dimensional scaling, conjoint analysis
10	**Report Writing:** Technique and Precaution of Interpretation, Significance of Report Writing, Layout and Types of Report

DMGT404 RESEARCH METHODOLOGY

Sr. No.	Description
1.	**An Introduction to Research:** Meaning, Process, Defining, Research Problem: Selection, Understanding, Necessity of defined problem, Research Design , need and types of Research Design.
2.	**Sampling Design:** Steps of Sampling design, Characteristics of good Sampling Design, Different types of Sampling Design.
3.	**Measurement and Scaling Technique:** Tools of Sound Measurement, Techniques Of Developing Measurement Tools, Scaling meaning and Important Scaling Techniques
4.	**Data Collection:** Primary (Interview, Observation and Questionnaire and Collection of Secondary Data.
5.	**Data Analysis:** Measure for Central Tendency, Dispersion, Correlation And Regression Analysis
6.	**Hypothesis Testing:** Hypothesis Definition and Formulation, t test , z test , ANOVA
7.	**Multivariate Analysis:** Classification, Important Methods of Factor analysis, factor analysis , rotation in Factor Analysis, overview of cluster analysis, discriminant analysis, multi dimensional scaling , conjoint analysis.
8.	**Report Writing:** Technique and Precaution of Interpretation, Significance of Report Writing, Layout and Types of Report.
9.	**Time series**
10	**Index number**

CONTENT

Unit 1: Introduction to Research

Notes

CONTENTS

Notes

Objectives

After studying this unit, you will be able to:

- Recognize the meaning and objectives of research
- Define research in the expression of different authors
- Generalize the Process of research
- Differentiate between different types of research

Introduction

Research means technical and organized search for relevant information on a particular topic. It is defined as an academic activity that involves identifying the research problem, formulating a hypothesis, collecting and analyzing data and reaching specific conclusions in the form of solutions or general theories. The primary objective of research is to find solutions for problems in a methodical and systematic way. A research depends on the field in which the research work is performed. Various types of researches can be done for different fields, like fundamental research for identifying the important principles of the research field and applied research for solving an immediate problem. However, all these researches primarily follow two approaches, quantitative and qualitative. The quantitative approach focuses on the quantity of the data obtained from the research, while the qualitative approach is concerned with the quality of the obtained data.

1.1 Meaning of Business Research

Business research is a systematic and objective process of gathering, recording and analyzing data for aid in making business decisions. Business research comes within the purview of social science research. Social science research refers to research conducted by social scientists (primarily within sociology and social psychology), but also within other disciplines such as social policy, human geography, political science, social anthropology and education. Sociologists and other social scientists study diverse things: from census data on hundreds of thousands of human beings, through the in-depth analysis of the life of a single important person to monitoring what is happening on a street today-or what was happening a few hundred years ago.

Social scientists use many different methods in order to describe, explore and understand social life. Social methods can generally be subdivided into two broad categories. Quantitative methods are concerned with attempts to quantify social phenomena and collect and analyze numerical data, and focus on the links among a smaller number of attributes across many cases. Qualitative methods, on the other hand, emphasize personal experiences and interpretation over quantification, are more concerned with understanding the meaning of social phenomena and focus on links among a larger number of attributes across relatively few cases. While very different in many aspects, both qualitative and quantitative approaches involve a systematic interaction between theories and data.

1.1.1 Research Objectives

Research in common man's language refers to "search for Knowledge".

Research is an art of scientific investigation. It is also a systematic design, collection, analysis and reporting the findings & solutions for the marketing problem of a company. Research is required because of the following reasons:

Notes

1. To identify and find solutions to the problems
2. To help making decisions
3. To develop new concepts
4. To find alternate strategies

1. ***To Identify and Find Solutions to the Problem:*** To understand the problem in depth

Example: "Why is that demand for a product is falling"? "Why is there a business fluctuation once in three years"? By identifying the problem as above, it is easy to collect the relevant data to solve the problem.

2. ***To Help making Decisions:***

Example: Should we maintain the advertising budget same as last year? Research will answer this question.

3. ***To Find Alternative Strategies:*** Should we follow pull strategy or push strategy to promote the product.
4. ***To Develop New Concepts:***

Example: CRM, Horizontal Marketing, MLM, etc.

1.1.2 Marketing Research

Marketing research is an important part of overall business research. Systematic collection and analysis of data relating to sale and distribution of financial products and services is called marketing research. Market research is an early step in the marketing process, and includes an analysis of market demand for a new product, or for existing products, as well as appropriate methods of distributing those products. Techniques in market research include telephone polling and focus group interviews to determine customer attitudes, pricing sensitivity, and willingness to use delivery alternatives. Marketing research, or market research, is a form of business research and is generally divided into two categories: consumer market research and business-to-business (B2B) market research, which was previously known as industrial marketing research. Consumer marketing research studies the buying habits of individual people while business-to-business marketing research investigates the markets for products sold by one business to another.

Did u know? Most large banks have their own market research departments that evaluate not only products, but their Brick and Mortar branch banking networks through which most banking products are sold.

Self Assessment

Fill in the blanks:

1. Business research comes within the purview of ..... research.
2. ..... Market research, which was previously known as industrial marketing research.

Notes

3. ..... methods are concerned with attempts to quantify social phenomena and collect and analyse numerical data.

1.2 Defining Research

Various authors and management gurus have defined research in different ways. Usually a research is said to begin with a question or a problem. The purpose of research is to find solutions through the application of systematic and scientific methods. Thus, research is a systematic approach to purposeful investigation. Some of the proposed definitions of research are:

According to Redman and Mory, research is a systematised effort to gain new knowledge.

According to Clifford Woody, research comprises defining and redefining problems, formulating hypotheses or suggesting solutions; collecting, organising and evaluating data; making deductions and reaching conclusions; and at last carefully testing the conclusions to determine whether they agree with the formulated hypothesis or not.

D. Slesinger and M. Stephenson in the Encyclopedia of Social Sciences define research as: 'the manipulation of things, concepts or symbols for the purpose of generalising to extend, correct or verify knowledge, whether that knowledge aids in construction of theory or in the practice of an art.'

Self Assessment

Fill in the blanks:

4. The purpose of research is to find solutions through the application of ..... and ..... methods.
5. Research is a systematised effort to gain .....
6. Research is a systematic approach to ..... investigation.

1.3 Research Process

Until the sixteenth century, human inquiry was primarily based on introspection. The way to know things was to turn inward and use logic to seek the truth. This paradigm had endured for a millennium and was a well-established conceptual framework for understanding the world. The seeker of knowledge was an integral part of the inquiry process. A profound change occurred during the sixteenth and seventeenth centuries. The Scientific Revolution was born. Objectivity became a critical component of the new scientific method. The investigator was an observer, rather than a participant in the inquiry process. A mechanistic view of the universe evolved. We believed that we could understand the whole by performing an examination of the individual parts. Experimentation and deduction became the tools of the scholar. For two hundred years, the new paradigm slowly evolved to become part of the reality framework of society.

The research process is a step-by-step process of developing a research paper. As you progress from one step to the next, it is commonly necessary to backup, revise, add additional material or even change your topic completely. This will depend on what you discover during your research. There are many reasons for adjusting your plan. For example, you may find that your topic is too broad and needs to be narrowed, sufficient information resources may not be available, what you learn may not support your thesis, or the size of the project does not fit the requirements.

Notes

Notes The research process itself involves identifying, locating, assessing, analyzing, and then developing and expressing your ideas. These are the same skills you will need outside the academic world when you write a report or proposal for your boss.

There are nine steps in the research process, that can be followed while designing a research project. They are as follows:

1. Formulate the problem
2. Evaluate the cost of research
3. Prepare the list of information
4. Research design decision
5. Data collection
6. Select the sample type
7. Determine the sample size
8. Organize the field work
9. Analyze the data and report preparation

Defining the research problem and formulation of hypothesis are the hardest steps in the research process.

1.3.1 Problem Formulation

Problem formulation is the key to research process. For a researcher, problem formulation means converting the management problem to a research problem. In order to attain clarity, the MR manager and researcher must articulate clearly so that perfect understanding of each others is achieved.

While problem is being formulated, the following should be taken into account:

1. Determine the objective of the study
2. Consider various environment factors
3. Nature of the problem
4. State the alternative

1. ***Determine the objective:*** Objective may be general or specific. General - Would like to know, how effective was the advertising campaign.

   The above looks like a statement with objective. In reality, it is far from it. There are two ways of finding out the objectives precisely. (a) The researcher should clarify with the MR manager "What effective means". Does effective mean, awareness or does it refer to sales increase or does it mean, it has improved the knowledge of the audience, or the perception of audience about the product. In each of the above circumstances, the questions to be asked from audience varies (b) Another way to find objectives is to find out from the MR Manager, "What action will be taken, given the specified outcome of the study."

Notes

Example: If research finding is that, the previous advertisement by the company was indeed ineffective, what course of action the company intends to take (a) Increase the budget for the next Ad (b) Use different appeal (c) Change the media (d) Go to a new agency.

Caution: If objectives are proper, research questions will be precise. However we should remember that objectives, do undergo a change.

2. ***Consider environmental factors:*** Environmental factors influence the outcome of the research and the decision. Therefore, the researcher must help the client to identify the environmental factors that are relevant.

Example: Assume that the company wants to introduce a new product like Iced tea or frozen green peas or ready to eat chapathis.

The following are the environmental factors to be considered:

(a) Purchasing habit of consumers.

(b) Presently, who are the other competitors in the market with same or similar product.

(c) What is the perception of the people about the other products of the company, with respect to price, image of the company.

(d) Size of the market and target audience.

All the above factors could influence the decision. Therefore researcher must work very closely with his client.

3. ***Nature of the problem:*** By understanding the nature of the problem, the researcher can collect relevant data and help suggesting a suitable solution. Every problem is related to either one or more variable. Before starting the data collection, a preliminary investigation of the problem is necessary, for better understanding of the problem. Initial investigation could be, by using focus group of consumers or sales representatives.

If focus group is carried out with consumers, some of the following question will help the researcher to understand the problem better:

(a) Did the customer ever included this company's product in his mental map?

(b) If the customer is not buying the companies product, the reasons for the same.

(c) Why did the customer go to the competitor?

(d) Is the researcher contacting the right target audience?

4. ***State the alternatives:*** It is better for the researcher to generate as many alternatives as possible during problem formulation hypothesis.

Example: Whether to introduce a Sachet form of packaging with a view to increase sales. The hypothesis will state that, acceptance of the sachet by the customer will increase the sales by 20%. Thereafter, the test marketing will be conducted before deciding whether to introduce sachet or not. Therefore for every alternative, a hypothesis is to be developed.

1.3.2 Evaluate the Cost of Research

There are several methods to establish the value of research. Some of them are (1) Bayesian approach (2) Simple saving method (3) Return on investment (4) Cost benefit approach etc.

Notes

Example: Company 'X' wants to launch a product. The company's intuitive feeling is that, the product failure possibilities is 35%. However, if research is conducted and appropriate data is gathered, the chances of failure can be reduced to 30%. Company also has calculated, that the loss would be ₹ 3,00,000 if product fails. The company has received a quote from MR agency. The cost of research is ₹ 75,000. The question is "Should the company spend this money to conduct research?"

Solution:

$$\text{Loss without research} = 3,00,000 \times 0.35$$

$$= ₹\ 1,05,000$$

$$\text{Loss with research} = 3,00,000 \times 0.30$$

$$= ₹\ 90,000$$

$$\text{Value of research information} = 1,05,000 - 90,000$$

$$= ₹\ 15,000$$

Since the value of information namely ₹ 15000 is lower than the cost of research ₹ 75,000, conducting research is not recommended.

1.3.3 Preparing a List of Needed Information

Assume that company 'X' wants to introduce a new product (Tea powder). Before introducing it, the product has to be test marketed. The company needs to know the extent of competition, price and quality acceptance from the market. In this context, following are the list of information required.

1. ***Total demand and company sales:***

Example: What is the overall industry demand? What is the share of the competitor? The above information will help the management to estimate overall share and its own shares, in the market.

2. ***Distribution coverage:***

Example:

(a) Availability of products at different outlets.

(b) Effect of shelf display on sales.

3. ***Market awareness, attitude and usage:***

Example: "What percentage of target population are aware of firm's product"? "Do customers know about the product"? "What is the customers' attitude towards the product"? "What percentage of customers repurchased the product"?

4. ***Marketing expenditure:***

Example: "What has been the marketing expenditure"? "How much was spent on promotion"?

Notes

5. ***Competitors marketing expenditure:***

Example: "How much competitor spent, to market a similar product"?

1.3.4 Decision on Research Design

1. Should the research be exploratory or conclusive?

   Exploratory research:

Example: "Causes for decline in sales of a specific company's product in a specific territory under a specific salesman".

The researcher may explore all possibilities why sales in falling?

(a) Faulty product planning

(b) Higher price

(c) Less discount

(d) Less availability

(e) Inefficient advertising/salesmanship

(f) Poor quality of salesmanship

(g) less awareness

Not all factors are responsible for decline in sales.

Conclusive research: Narrow down the option. Only one or two factors are responsible for decline in sales. Therefore zero down, and use judgment and past experience.

2. Who should be interviewed for collecting data?

   If the study is undertaken to determine whether, children influence the brand, for ready - to eat cereal (corn flakes) purchased by their parents. The researcher must decide, if only adults are to be studied or children are also to be included. The researcher must decide if data is to be collected by observation method or by interviewing. If interviewed, "Is it a personal interview or telephonic interview or questionnaire?"

3. Should a few cases be studied or choose a large sample?

   The researcher may feel that, there are some cases available which are identical and similar in nature. He may decide to use these cases for formulating the initial hypothesis. If suitable cases are not available, then the researcher may decide to choose a large sample.

4. How to incorporate experiment in research?

   If it is an experiment, "Where and when measurement should take place?", should be decided.

Example: In a test of advertising copy, the respondents can first be interviewed to measure their present awareness, and their attitudes towards certain brands. Then, they can be shown a pilot version of the proposed advertisement copy, following this, their attitude also is to be measured once again, to see if the proposed copy had any effect on them.

If it is a questionnaire, (a) What are the contents of the questionnaire? (b) What type of questions to be asked? Like pointed questions, general questions etc. (c) In what sequence should it be

Notes

asked? (d) Should there be a fixed set of alternatives or should it be open ended. (e) Should the purpose be made clear to the respondents or should it be disguised are to be determined well in advance.

Task Prepare a questionnaire to find if the consumers appreciate your new product as compared to the older ones or not.

1.3.5 Select the Sample Types

The first task is to carefully select "What groups of people or stores are to be sampled". For example, collecting the data from a fast food chain. Here, it is necessary to define what is meant by fast food chain. Also precise geographical location should be mentioned.

Next step is to decide whether to choose probability sampling or non-probability sampling. Probability sampling is one, in which each element has a known chance of being selected.

Notes A non-probability sampling can be convenience or judgment sampling.

1.3.6 Determine the Sample Size

Smaller the sample size, larger the error, vice versa.

Sample size depends up on (a) Accuracy required (b) Time available (c) Cost involved.

Sample size depends on the size of the sample frame/universe. Example: Survey on the attitudes towards the use of shampoo with reference to a specific brand, where husbands, wives or combination of all of them are to be surveyed or a specific segment is to be surveyed.

Caution While selecting the sample, the sample unit has to be clearly specified

1.3.7 Organize the Fieldwork

This includes selection, training and evaluating the field sales force to collect the data

(a) How to analyzing the field work?

(b) What type of questionnaire - structured/unstructured to use?

(c) How to approach the respondents?

(d) Week, day and time to meet the specific respondents etc., are to be decided.

1.3.8 Analyze the Data and Report Preparation

This involves (a) Editing, (b) Tabulating, (c) Codifying etc.

1. The data collected should be scanned, to make sure that it is complete and all the instructions are followed. This process is called editing. Once these forms have been edited, they must be coded.
2. Coding means, assigning numbers to each of the answers, so that they can be analyzed.

Notes

3. The final step is called as data tabulation. It is the orderly arrangement of the data in a tabular form. Also at the time of analyzing the data, the statistical tests to be used must be finalized such as T-Test, Z-test, Chi-square Test, ANOVA, etc.

Self Assessment

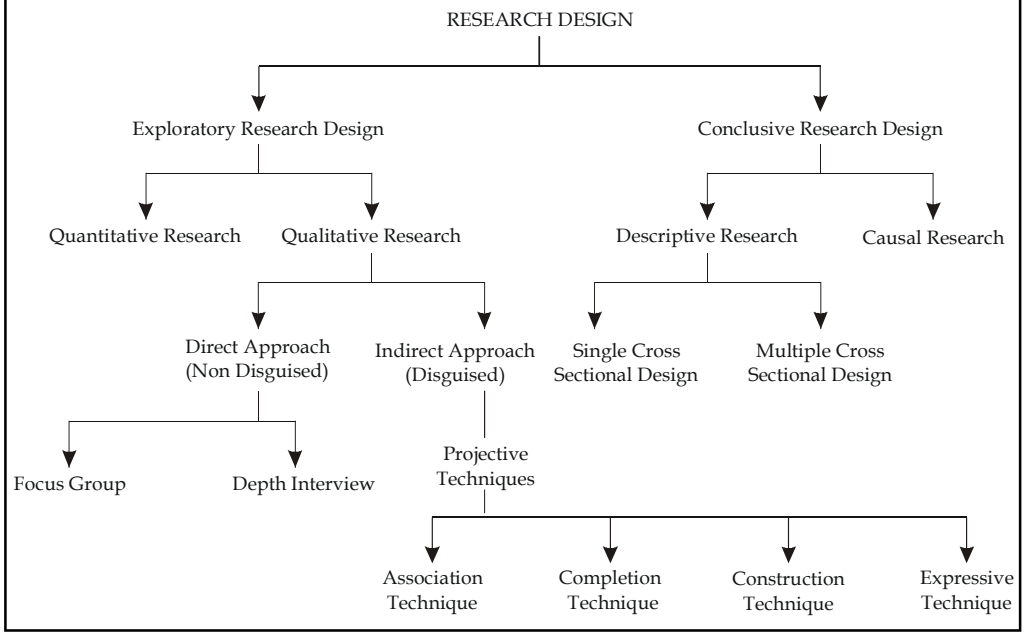
Fill in the blanks:

7. ..... is the orderly arrangement of the data in a tabular form.
8. While selecting the sample, the ..... has to be clearly specified.
9. A ..... sampling can be convenience or judgment sampling.
10. The ..... must decide if data is to be collected by observation method or by interviewing.
11. It is better for the researcher to generate as many alternatives as possible during problem .....
12. There are ..... steps in the research process.

1.4 Types of Research

There are different types of research. A detailed description of the same can be had from Figure 1.1 and the description that follows:

Figure 1.1: Research Methodology



1.4.1 Exploratory Research

This type of research is carried out at the very beginning when the problem is not clear or is vague. In exploratory research, all possible reasons which are very obvious are eliminated, thereby directing the research to proceed further with limited options.

Notes

Sales decline in a company may be due to:

1. Inefficient service
2. Improper price
3. Inefficient sales force
4. Ineffective promotion
5. Improper quality

The research executives must examine such questions to identify the most useful avenues for further research. Preliminary investigation of this type is called exploratory research. Expert surveys, focus groups, case studies and observation methods are used to conduct the exploratory survey.

1.4.2 Descriptive Research

The main purpose of descriptive research is to describe the state of view as it exists at present. Simply stated, it is a fact finding investigation. In descriptive research, definite conclusions can be arrived at, but it does not establish a cause and effect relationship. This type of research tries to describe the characteristics of the respondent in relation to a particular product.

1. Descriptive research deals with demographic characteristics of the consumer. For example, trends in the consumption of soft drink with respect to socio-economic characteristics such as age, family, income, education level etc. Another example can be the degree of viewing TV channels, its variation with age, income level, profession of respondent as well as time of viewing. Hence, the degree of use of TV to different types of respondents will be of importance to the researcher. There are three types of players who will decide the usage of TV: (i) Television manufacturers, (ii) Broadcasting agency of the programme, (iii) Viewers. Therefore, research pertaining to any one of the following can be conducted:
   (a) The manufacturer can come out with facilities which will make the television more user-friendly. Some of the facilities are – (i) Remote control, (ii) Child lock, (iii) Different models for different income groups, (iv) Internet compatibility etc., (v) Wall mounting etc.
   (b) Similarly, broadcasting agencies can come out with programmes, which can suit different age groups and income.
   (c) Ultimately, the viewers who use the TV must be aware of the programmes appearing in different channels and can plan their viewing schedule accordingly.
2. Descriptive research deals with specific predictions, for example, sales of a company's product during the next three years, i.e., forecasting.
3. Descriptive research is also used to estimate the proportion of population who behave in a certain way.

Example: "Why do middle income groups go to Food World to buy their products?"

A study can be commissioned by a manufacturing company to find out various facilities that can be provided in television sets based on the above discussion.

Similarly, studies can be conducted by broadcasting stations to find out the degree of utility of TV programmes.

Notes


Example: The following hypothesis may be formulated about the programmes:

1. The programmes in various channels are useful by way of entertainment to the viewers.
2. Viewers feel that TV is a boon for their children in improving their knowledge- especially, fiction and cartoon programmes.

1.4.3 Applied Research

Applied research aims at finding a solution for an immediate problem faced by any business organization. This research deals with real life situations.


Example: "Why have sales decreased during the last quarter"? Market research is an example of applied research. Applied research has a practical problem-solving emphasis. It brings out many new facts.

1. Use of fibre glass body for cars instead of metal.
2. To develop a new market for the product.

1.4.4 Pure/Fundamental Research or Basic Research

Gathering knowledge for knowledge's sake is known as basic research. It is not directly involved with practical problems. It does not have any commercial potential. There is no intention to apply this research in practice. Tata Institute of Fundamental Research conducts such studies.


*Example:*Theory of Relativity (by Einstein).

1.4.5 Conceptual Research

This is generally used by philosophers. In this type of research, the researcher should collect the data to prove or disapprove his hypothesis. The various ideologies or 'isms' are examples of conceptual research.


Did u know? Conceptual Research is related to some abstract idea or theory.

1.4.6 Causal Research

Causal research is conducted to determine the cause and effect relationship between the two variables.


Example: Effect of advertisement on sales.

1.4.7 Historical Research

The name itself indicates the meaning of the research. Historical study is a study of past records and data in order to understand the future trends and development of the organisation or market. There is no direct observation. The research has to depend on the conclusions or inferences drawn in the past.

Notes

Example: Investors in the share market study the past records or prices of shares which he/she intends to buy. Studying the share prices of a particular company enables the investor to take decision whether to invest in the shares of a company.

Crime branch police/CBI officers study the past records or the history of the criminals and terrorists in order to arrive at some conclusions.

The main objective of this study is to derive explanation and generalization from the past trends in order to understand the present and anticipate the future.

There are however, certain shortcomings of historical research:

1. Reliability and adequacy information is subjective and open to question
2. Accuracy of measurement of events is doubtful.
3. Verification of records are difficult.

Task List the records to be considered while conducting a historical research in analyzing the sales aspect of a television brand

1.4.8 Ex-post Facto Research

In this type of research, an examination of relationship that exists between independent and dependent variable is studied. We may call this empirical research. In this method, the researcher has no control over an independent variable. Ex-post facto literally means "from what is done afterwards". In this research, a variable "A" is observed. Thereafter, the researcher tries to find a causal variable "B" which caused "A". It is quite possible that "B" might not have been caused "A". In this type of analysis, there is no scope for the researcher to manipulate the variable. The researcher can only report "what has happened" and "what is happening".

1.4.9 Action Research

This type of research is undertaken by direct action. Action research is conducted to solve a problem. For example, test marketing a product is an example of action research. Initially, the geographical location is identified. A target sample is selected from among the population. Samples are distributed to selected samples and feedback is obtained from the respondent. This method is most common for industrial products, where a trial is a must before regular usage of the product.

1.4.10 Evaluation Research

This is an example of applied research. This research is conducted to find out how well a planned programme is implemented. Therefore, evaluation research deals with evaluating the performance or assessment of a project.

Example: "Rural Employment Programme Evaluation" or "Success of Midday Meal Programme".

Notes

1.4.11 Library Research

This is done to gather secondary data. This includes notes from the past data or review of the reports already conducted. This is a convenient method whereby both manpower and time are saved.

Self Assessment

Fill in the blanks:

13. ..... is conducted to solve a problem.
14. In ..... research, an examination of relationship that exists between independent and dependent variable is studied.
15. ..... research is generally used by philosophers.
16. Descriptive research deals with ..... characteristics of the consumer
17. Evaluation research is an example of ..... research
18. ..... research is done to gather secondary data.
19. Gathering knowledge for knowledge's sake is known as ..... research.
20. In exploratory research, all possible reasons which are ..... are eliminated

1.5 Summary

- Research originates in a decision process.
- Usually a research is said to begin with a question or a problem.
- In research process, management problem is converted into a research problem which is the major objective of the study.
- Research question is further subdivided, covering various facets of the problem that need to be solved.
- The role and scope of research has greatly increased in the field of business and economy as a whole.
- The study of research methods provides you with knowledge and skills you need to solve the problems and meet the challenges of today is modern pace of development.

1.6 Keywords

Ad Tracking: It is periodic or continuous in-market research to monitor a brand's performance using measures such as brand awareness, brand preference, and product usage.

Advertising Research: It is a specialized form of marketing research conducted to improve the efficacy of advertising.

Concept Testing: To test the acceptance of a concept by target consumers.

Copy Testing: It predicts in-market performance of an ad before it airs by analyzing audience levels of attention, brand linkage, motivation, entertainment, and communication, as well as breaking down the ad's flow of attention and flow of emotion.

Notes

Exploratory Research: Exploratory research provides insights into and comprehension of an issue or situation.

Marketing Research: Marketing research is about researching the whole of a company's marketing process.

Mystery Shopping: An employee or representative of the market research firm anonymously contacts a salesperson and indicates he or she is shopping for a product. The shopper then records the entire experience.

Product Research: This looks at what products can be produced with available technology, and what new product innovations near-future technology can develop.

1.7 Review Questions

1. An Indian company dealing in pesticides hires a qualified business management graduate to expand its marketing activities. Most of the current employees of the company are qualified chemists with science background. During their first review meeting the management graduate says that the "company should be involved in market research to get a better perspective of the problem on hand". On hearing this, one of the science graduate laughs and says "There is no such thing as marketing or business research, research is combined to science alone." What would be your response?
2. What would be the instances in which you might take causal research in your organization?
3. It is said that action research is conducted to solve a problem. Why are the other researches conducted then?
4. What type of research would you undertake in order find why middle income groups go to a particular retail store to buy their products?
5. Which research would you undertake if you have got a practical problem?
6. Which type of research would you conduct when the problem is not clear and all the possible reasons are eliminated? Why?
7. How does a research help the managers to determine the pattern of consumption?
8. Do you think that a market research helps the marketer to identify brand loyalty and establish it with further strength? Why/why not?
9. When records exist in all authenticated form, why is it so that their verification remains a big issue?
10. Is there any difference in pure research and ex-post facto research? Support you answer with suitable reasons.

Answers: Self Assessment

1. Social science
2. Business to Business (B2B)
3. Quantitative
4. Systematic, scientific
5. New knowledge
6. Purposeful
7. Data tabulation
8. sample unit
9. non-probability
10. researcher

Notes

11. formulation hypothesis
12. nine
13. Action research
14. Ex-post Facto
15. Conceptual
16. demographic
17. applied
18. Library
19. basic
20. very obvious

1.8 Further Readings

Books

Abrams, M.A., *Social Surveys and Social Action*, London: Heinemann, 1951.

Arthur, Maurice, *Philosophy of Scientific Investigation*, Baltimore: John Hopkins University Press, 1943.

Bernal, J.D., *The Social Function of Science*, London: George Routledge and Sons, 1939.

Chase, Stuart, *The Proper Study of Mankind: An inquiry into the Science of Human Relations*, New York, Harper and Row Publishers, 1958.

S. N. Murthy and U. Bhojanna, *Business Research Methods*, Excel Books.

Unit 2: Research Problem

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Formulate a research problem
- Identify the selection of the problem
- Report the understanding of problem
- State about necessity of defined problem
- Demonstrate the Self Questioning by researcher while defining the problem

Introduction

In all organizations, some kind of research is required to support decision-making, for example, examination of circulation records to determine if fund allocations should be changed. A manager exists in three time dimensions: past, present and future. The past specifies an accurate sense of what was achieved and what was not, while the present specifies what is being achieved. On the other hand, the future time dimension specifies what a manager should achieve.

Research is used to provide facts on the first two, which supports the decisions that will have an impact on the future. These decisions are made on the basis of collected data or facts. The importance of the decisions and their impact on the organization will determine the importance of research.

There is a famous saying that "problem well-defined is half solved". This statement is strikingly true in market research, because if the problem is not stated properly, the objectives will not be clear. If the objective is not clearly defined, the data collection becomes meaningless.

Research problem is a condition that causes a researcher to feel anxious, uneasy and confused. It involves the complete analysis of the problem area involving who, what, where, when and why of the problem situation.

Notes

2.1 Research Problem

A research problem refers to some difficulty which an organisation faces and wishes to obtain a solution for the same.

While doing research, defining the problem is very important because "problem clearly stated is half-solved". This shows how important it is to "define the problem correctly". While defining the problem, it should be noted that definition should be unambiguous. If the problem defining is ambiguous, then the researcher will not know "what data is to be collected" or "what technique is to be used" etc.

Example of an ambiguous definition: "Find out by how much sales has declined recently". Let us suppose that the research problem is defined in a broad and general way as follows:

"Why is the productivity in Korea much higher than that in India"? In this type of question, a number of ambiguities are there, such as:

1. What sort of productivity is to be specified; is it men, machine, materials?
2. To which type of industry is the productivity related to?
3. In which time-frame are we analysing the productivity?

Example of an unambiguous definition: On the contrary, a problem will be as follows:

"What are the factors responsible for increased labour productivity in Korean textile manufacturing industries during 1996-07 relative to Indian textile industries?"

Notes Problem formulation is the key to research process. For a researcher, problem formulation means converting the management problem to a research problem. In order to attain clarity, the M.R. manager and researcher must articulate clearly so that perfect understanding of each others is achieved.

In research process, the first and foremost step happens to be that of selecting and properly defining a research problem. A researcher must find the problem and formulate it so that it becomes susceptible to research. To define a problem correctly, a researcher must know: what a problem is?

Did u know? Like a medical doctor, a researcher must examine all the symptoms (presented to him or observed by him) concerning a problem before he can diagnose correctly.

Self Assessment

Fill in the blanks:

1. In order to attain clarity, the manager and researcher must ..... clearly.
2. Problem ..... is the key to research process.
3. To define a problem correctly, a researcher must know: .....

2.2 Selection of the Problem

Notes

The research problem undertaken for study must be carefully selected. The task is a difficult one, although it may not appear to be so. Help may be taken from a research guide in this connection. Nevertheless, every researcher must find out his own salvation for research problems cannot be borrowed. A problem must spring from the researcher's mind like a plant springing from its own seed. If our eyes need glasses, it is not the optician alone who decides about the number of the lens we require. We have to see ourself and enable him to prescribe for us the right number by cooperating with him. Thus, a research guide can at the most only help a researcher choose a subject.

Inevitably, selecting a problem is somewhat arbitrary, idiosyncratic, and personal. Avoid selecting the first problem that you encounter. Try to select the most interesting and personally satisfying choice from among two or three possibilities. The problem selection should matter to you. You should be eager and enthusiastic.

A good topic should be small enough for a conclusive investigation and large enough to yield interesting results.

Caution Remember that research must yield a publication for it to have meaning.

You may wish to query likely periodical editors to see if they might be interested in an article on your research topic.

In some cases, as with a thesis or a dissertation, some sort of preliminary study may be needed to see if the problem and the study are feasible and to identify snags. Such a PILOT STUDY can be quite valuable.

Task Analyse what problems you might encounter while selecting a problem?

Selection Criteria

1. Your genuine enthusiasm for the problem.
2. Controversial subject should not become the choice of an average researcher.
3. The degree to which research on this problem benefits the profession and society.
4. The degree to which research on this problem will assist your professional goals and career objectives.
5. Too narrow or too vague problems should be avoided.
6. The degree to which this research will interest superiors and other leaders in the field.
7. The degree to which the research builds on your experience and knowledge.
8. Ease of access to the population to be studied and the likelihood that they will be cooperative Affordability.

Notes

9. Likelihood of publication.
10. Relationship to theories or accepted generalizations in the field.
11. Degree to which ethical problems are involved.
12. Degree to which research is unique or fills a notable gap in the literature.
13. Degree to which the research builds on and extends existing knowledge before the final selection of a problem is done, a researcher must ask himself the following questions:
    (a) Whether he is well equipped in terms of his background to carry out the research?
    (b) Whether the study falls within the budget he can afford?
    (c) Whether the necessary cooperation can be obtained from those who must participate in research as subjects?

Self Assessment

Fill in the blanks:

4. A good topic should be small enough for a ..... investigation.
5. A ..... should always avoid selecting the first problem that he encounters.
6. The research problem undertaken for study must be ..... selected.

2.3 Understanding the Problem

Once the problem has been selected, the same has to be understood thoroughly and then the same has to be reframed into meaningful terms from an analytical point of view. The first step in research is to formulate the problem. A company manufacturing television sets might think that it is losing sales to a foreign company. A brief illustration aptly demonstrates how such problem can be ill-conceived. The management of a company felt, a drop in sales was because of the poor quality of product. Subsequently, research was undertaken with a view to improve the quality of the product. But despite an improvement in quality, sales did not pick up. In this case, we may say that the problem is ill-defined. The actual reason was ineffective sales promotion. The problem thus needs to be carefully identified.

Did u know? Marketing problem which needs research can be classified into two categories:

1. Difficulty related problems
2. Opportunity related problems, while the first category produces negative results such as, decline in market share or sales, the second category provides benefits.

Problem definition might refer to either a real-life situation or it may also refer to a set of opportunities. Market research problems or opportunities will arise under the following circumstances: (1) Unanticipated change (2) Planned change. Many factors in the environment can create problems or opportunities. Thus, change in the demographics, technological and legal changes affect the marketing function. Now the question is how the company responds to new technology, or product introduced by the competitor or how to cope with the changes in life-styles. It may be a problem and at the same time, it can also be viewed as an opportunity. In order to conduct research, the problem must be defined accurately.

Notes

While formulating the problem, clearly define:

1. Who is the focus?
2. What is the subject-matter of research?
3. To which geographical territory/area the problem refers to?
4. To which period does the study pertains to?

Example: "Why does the upper-middle class of Bangalore shop at Life-style during the Diwali season"?

Here all the above four aspects are covered. We may be interested in a number of variables due to which shopping is done at a particular place. The characteristic of interest to the researcher may be (1) Variety offered at life-style (2) Discount offered by way of promotion (3) Ambience at the life-style and (4) Personalised service offered. In some cases, the cause of the problem is obvious whereas in others the cause is not so obvious. The obvious causes are the products being on the decline. Not so obvious causes could be a bad first experience for the customer.

Self Assessment

Fill in the blanks:

7. Changes in the demographics, technological and legal changes affect the ..... function.
8. Opportunity related problems produce ..... results.
9. The first step in research is to formulate the ……..

2.4 Necessity of Defined Problem

Defining a research problem properly is a prerequisite for any study and is a step of the highest importance. A problem well defined is half solved. Defining the problem is often more essential than its solution because when the problem is formulated, an appropriate technique can be applied to generate alternative solutions. This statement signifies the need for defining a research problem. The problem to be investigated must be defined unambiguously for that will help to discriminate relevant data from the irrelevant ones. When you define a research problem you are trying to reduce the outcome of an answer. The question of course when you speak about "marketing research" is how I can target more customers that I can sell my product to. You are looking for specific answers such as: "What type of soda do all foreign born males between the ages of 25-35 drink?" This is defining the problem. What do you consider foreign born males? What constitutes soda? etc. This is important because companies and sales organization attempt to "target" their market instead of taking a shotgun approach. The process is to first make sure any information you obtain is credible and from a reputable organization. Then break down your problem and pick apart any inconsistencies you may see within you research project. Problem formulation is the key to research process. For a researcher, problem formulation means converting the management problem to a research problem. In order to attain clarity, the manager and researcher must articulate clearly so that perfect understanding of each others is achieved.

Notes

Notes A proper definition of research problem will enable the researcher to be on the track whereas an ill-defined problem may create hurdles.

What are the sources of problem identification?

Research students can adopt the following ways to identify the problems:

1. Research reports already published may be referred to define a specific problem.
2. Assistance of any research organisation, which handles a number of projects of the companies, can be sought to identify the problem.
3. Professors working in reputed academic institution can act as guides in problem identification.
4. Company employees and competitors can assist in identifying the problems.
5. Cultural and technological changes can act as a source for research problem identification.
6. Seminars/symposiums/focus groups can act as a useful source.

Self Assessment

Fill in the blanks:

10. ..... and ..... changes can act as a source for research problem identification.
11. Research reports already published may be referred to define a .....
12. When you define a research problem you are trying to ..... the outcome of an answer.
13. A problem well ..... is half solved.

2.5 Self Questioning by Researcher while Defining the Problem

1. Is the research problem correctly defined?
2. Is the research problem solvable?
3. Can relevant data be gathered through the process of marketing research?
4. Is the research problem significant?
5. Can the research be conducted within the available resources?
6. Is the time given to complete the project sufficient?
7. What exactly will be the difficulties in conducting the study, and hurdles to be overcome?
8. Am I competent, to carry the study out?

Managers often want the results of research in accordance with their expectation. This satisfies them immensely. If one were to closely look at the questionnaire, it is found that in most cases, there are stereotyped answers given by the respondents.

Notes

Caution Creativity aspect is fundamentally to be included by a researcher to look at problems in a different perspective.

Self Assessment

Fill in the blanks:

14. Managers often want the results of research in accordance with their .....
15. Assistance of any research organisation, which handles a number of projects of the companies, can be sought to ..... the problem.

2.6 Summary

- Proper problem formulation is the key to success in research.
- It is vital and any error in defining the problem incorrectly can result in wastage of time and money.
- Several elements of introspection will help in defining the problem correctly.
- The task of defining a research problem, very often, follows a sequential pattern.
- The problem is stated in a general way, the ambiguities are resolved, thinking and rethinking process results in a more specific formulation of the problem.
- It is done so that it may be a realistic one in terms of the available data and resources and is also analytically meaningful.
- All this results in a well defined research problem that is not only meaningful from an operational point of view.
- But is equally capable of paving the way for the development of working hypotheses and for means of solving the problem itself.

2.7 Keywords

Marketing Research Problem: It is a situation where your company intends to sell a product or service that fills a specific gap.

Objective of Research: It means to what the researcher aims to achieve.

Pilot Study: A small scale preliminary study conducted before the main research in order to check the feasibility or to improve the design of the research.

Problem Definition: The process in order to clear understanding (explanation) of what the problem is.

Research Problem: It focuses on the relevance of the present research.

2.8 Review Questions

1. The objective of research problem should be clearly defined; otherwise the data collection becomes meaningless. Discuss with suitable examples.

Notes

2. Cultural and technological changes can act as a source for research problem identification. Why/why not?
3. Defining a research problem properly is a prerequisite for any study. Why?
4. What precautions should be taken while formulating a problem?
5. If you are appointed to do a research for some problem with the client, what would you take as the sources for problem identification?
6. It may be a problem and at the same time, it can also be viewed as an opportunity. Why/why not?
7. In some cases, some sort of preliminary study may be needed. Which cases are being referred to and why?
8. A problem well defined is half solved. Comment.
9. While you define a research problem what do you try to do?
10. What do you think as the reason behind specialists suggesting to avoid selecting the first problem that you encounter?

Answers: Self Assessment

1. articulate
2. formulation
3. what a problem is
4. conclusive
5. researcher
6. carefully
7. marketing
8. negative
9. problem
10. Cultural, technological
11. specific problem
12. reduce
13. defined
14. expectation
15. identify

2.9 Further Readings

Books

C R Kotari, *Research Methodology*, Vishwa Prakashan.

Cooper and Schinder, *Business Research Methods*, TMH.

David Luck and Ronald Rubin, *Marketing Research*, PHI.

Naresh Amphora, *Marketing Research*, Pearson Education.

S. N. Murthy and U. Bhojanna, *Business Research Methods*, 3rd Edition, Excel Books.

Unit 3: Research Design

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Define research design
- Describe the need of research design
- Explain the different types of research design
- Identify the Secondary data and qualitative research
- Recognize the Descriptive research design
- Label the causal research design

Notes

Introduction

Research design is simply a plan for a study. This is used as a guide in collecting and analyzing the data. It can be called a blue print to carry out the study. It is like a plan made by an architect to build the house, if a research is conducted without a blue print, the result is likely to be different from what is expected at the start. The blue print includes (1) interviews to be conducted, observations to be made, experiments to be conducted data analysis to be made. (2) Tools used to collect the data such as questionnaire (3) what is the sampling methods used.

3.1 An Overview

Research design can be thought of as the structure of research - it is the "glue" that holds all of the elements in a research project together. A successful design stems from a collaborative process involving good planning and communication.

Research Design is mainly of three types namely, exploratory, descriptive and causal research.

Exploratory research is used to seek insights into general nature of the problem. It provides the relevant variable that need to be considered. In this type of research, there is no previous knowledge; research methods are flexible, qualitative and unstructured.

Notes The researcher in this method does not know "what he will find".

Descriptive research is a type of research, very widely used in marketing research. Generally in descriptive study there will be a hypothesis, with respect to this hypothesis, we ask questions like size, distribution, etc.

Causal research, this type of research is concerned with finding cause and effect relationship. Normally experiments are conducted in this type of research.

3.1.1 Need for Research Design

Before starting the research process, efficient and appropriate research design should be prepared. A research design is needed because of the following benefits it provides:

- It helps in smooth functioning of various research operations.
- It requires less effort, time and money.
- It helps to plan in advance the methods and techniques to be used for collecting and analysing data.
- It helps in obtaining the objectives of the research with the availability of staff, time and money.

The researcher should consider the following factors before creating a research design:

- The method for obtaining information source
- Skills of the researcher and the co-ordinating staff
- Problem objectives
- Nature of the problem
- Time and money available for the research work.

Notes

3.1.2 Types of Research Design

Exploratory, descriptive and causal research are some of the major types. Exploratory research is used to seek insights into general nature of the problem. It provides the relevant variable that need to be considered. In this type of research, there is no previous knowledge, research methods are flexible, qualitative and unstructured. The researcher in this method does not know "what he will find".

Descriptive research is a type of research, very widely used in marketing research. Generally in descriptive study there will be a hypothesis, with respect to this hypothesis, we ask questions like size, distribution, etc.

Causal research, this type of research is concerned with finding cause and effect relationship. Normally experiments are conducted in this type of research.

Self Assessment

Fill in the blanks:

1. ..... research is used to seek insights into general nature of the problem.
2. Research design helps to plan in advance the methods and techniques to be used for collecting and ……..data.

3.2 Exploratory Research

The major emphasis in exploratory research is on converting broad, vague problem statements into small, precise sub-problem statements, which is done in order to formulate specific hypothesis. The hypothesis is a statement that specifies, "how two or more variables are related?"

In the early stages of research, we usually lack from sufficient understanding of the problem to formulate a specific hypothesis. Further, there are often several tentative explanations.

Example: "Sales are down because our prices are too high",

"our dealers or sales representatives are not doing a good job",

"our advertisement is weak" and so on.

In this scenario, very little information is available to point out, what is the actual cause of the problem. We can say that the major purpose of exploratory research is to identify the problem more specifically. Therefore, exploratory study is used in the initial stages of research.

Under what circumstances is exploratory study ideal?

The following are the circumstances in which exploratory study would be ideally suited:

1. To gain an insight into the problem
2. To generate new product ideas
3. To list all possibilities. Among the several possibilities, we need to prioritize the possibilities which seem likely
4. To develop hypothesis occasionally

Did u know? Exploratory study is also used to increase the analyst's familiarity with the problem. This is particularly true, when the analyst is new to the problem area.

Notes


Example: A market researcher working for (new entrant) a company for the first time.

5. To establish priorities so that further research can be conducted.
6. Exploratory studies may be used to clarify concepts and help in formulating precise problems.


Example: The management is considering a change in the contract policy, which it hopes, will result in improved satisfaction for channel members.

An exploratory study can be used to clarify the present state of channel members' satisfaction and to develop a method by which satisfaction level of channel members is measured

7. To pre-test a draft questionnaire
8. In general, exploratory research is appropriate to any problem about which very little is known. This research is the foundation for any future study.

3.2.1 Characteristics of Exploratory Stage

1. Exploratory research is flexible and very versatile.
2. For data collection structured forms are not used.
3. Experimentation is not a requirement.
4. Cost incurred to conduct study is low.
5. This type of research allows very wide exploration of views.
6. Research is interactive in nature and also it is open ended.

3.2.2 Hypothesis Development at Exploratory Research Stage

1. Sometimes, it may not be possible to develop any hypothesis at all, if the situation is being investigated for the first time. This is because no previous data is available.
2. Sometimes, some information may be available and it may be possible to formulate a tentative hypothesis.
3. In other cases, most of the data is available and it may be possible to provide answers to the problem.

The examples given below indicate each of the above type:


Example:

Research Purpose	Research Question	Hypothesis
1. What product feature, if stated, will be most effective in the advertisement?	What benefit do people derive from this Ad appeal?	No hypothesis formulation is possible.
2. What new packaging is to be developed by the company (with respect to a soft drink)?	What alternatives exist to provide a container for soft drink?	Paper cup is better than any other forms, such as a bottle.
3. How can our insurance service be improved?	What is the nature of customer dissatisfaction?	Impersonalization is the problem.

Notes

In example 1: The research question is posed to determine "What benefit do people seek from the Ad?" Since no previous research is done on consumer benefit for this product, it is not possible to form any hypothesis.

In example 2: Some information is currently available about packaging for a soft drink. Here it is possible to formulate a hypothesis which is purely tentative. The hypothesis formulated here may be only one of the several alternatives available.

In example 3: The root cause of customer dissatisfaction is known, i.e. lack of personalised service. In this case, it is possible to verify whether this is a cause or not.

3.2.3 Formulation of Hypothesis in Exploratory Research

The quickest and the cheapest way to formulate a hypothesis in exploratory research is by using any of the four methods:

1. ***Literature Search:*** This refers to "referring to a literature to develop a new hypothesis". The literature referred are – trade journals, professional journals, market research finding publications, statistical publications etc. For example, suppose a problem is "Why are sales down?" This can quickly be analysed with the help of published data which should indicate, "whether the problem" is an "industry problem" or a "firm problem". Three possibilities exist to formulate the hypothesis.

   (a) The company's market share has declined but industry's figures are normal.

   (b) The industry is declining and hence the company's market share is also declining.

   (c) The industry's share is going up but the company's share is declining.

   If we accept the situation that our company's sales are down despite the market showing an upward trend, then we need to analyse the marketing mix variables.

Example:

(a) A TV manufacturing company feels that its market share is declining whereas the overall television industry is doing very well.

(b) Due to a trade embargo imposed by a country, textiles exports are down and hence sales of a company making garment for exports is on the decline.

The above information may be used to pinpoint the reason for declining sales.

2. ***Experience Survey:*** In experience surveys, it is desirable to talk to persons who are well informed in the area being investigated. These people may be company executives or persons outside the organisation. Here, no questionnaire is required. The approach adopted in an experience survey should be highly unstructured, so that the respondent can give divergent views.

Caution Since the idea of using experience survey is to undertake problem formulation, and not conclusion, probability sample need not be used. Those who cannot speak freely should be excluded from the sample.

Notes

Example:

(a) A group of housewives may be approached for their choice for a "ready to cook product".

(b) A publisher might want to find out the reason for poor circulation of newspaper introduced recently. He might meet (i) Newspaper sellers (ii) Public reading room (iii) General public (iv) Business community, etc.

These are experienced persons whose knowledge researcher can use.

3. ***Focus Group:*** Another widely used technique in exploratory research is the focus group. In a focus group, a small number of individuals are brought together to study and talk about some topic of interest. The discussion is co-ordinated by a moderator. The group usually is of 8-12 persons. While selecting these persons, care has to be taken to see that they should have a common background and have similar experiences in buying. This is required because there should not be a conflict among the group members on the common issues that are being discussed. During the discussion, future buying attitudes, present buying opinion, etc., are gathered.

   Most of the companies conducting the focus groups first screen the candidates to determine who will compose the particular group. Firms also take care to avoid groups, in which some of the participants have their friends and relatives, because this leads to a biased discussion. Normally, a number of such groups are constituted and the final conclusion of various groups are taken for formulating the hypothesis. Therefore a key factor in focus group is to have similar groups. Normally there are 4-5 groups. Some of them may even have 6-8 groups. The guiding criteria is to see whether the latter groups are generating additional ideas or repeating the same with respect to the subject under study. When this shows a diminishing return from the group, the discussions stopped. The typical focus group lasts for 1-30 hours to 2 hours. The moderator under the focus group has a key role. His job is to guide the group to proceed in the right direction.

   The following should be the characteristics of a moderator/facilitator:

   (a) *Listening:* He must have a good listening ability. The moderator must not miss the participant's comment, due to lack of attention.

   (b) *Permissive:* The moderator must be permissive, yet alert to the signs that the group is disintegrating.

   (c) *Memory:* He must have a good memory. The moderator must be able to remember the comments of the participants. Example: A discussion is centered around a new advertisement by a telecom company. The participant may make a statement early and make another statement later, which is opposite to what was said earlier. Example: The participant may say that s(he) never subscribed to the views expressed in the advertisement by the competitor, but subsequently may say that the "current advertisement of competitor is excellent".

   (d) *Encouragement:* The moderator must encourage unresponsive members to participate.

   (e) *Learning:* He should be a quick learner.

   (f) *Sensitivity:* The moderator must be sensitive enough to guide the group discussion.

   (g) *Intelligence:* He must be a person whose intelligence is above the average.

   (h) *Kind/firm:* He must combine detachment with empathy.

Notes

Notes **Variation of Focus Group**

1. *Respondent moderator group:* Under this method, the moderator will select one of the participants to act as a temporary moderator.
2. *Dualing moderator group:* In this method, there are two moderators. They purposely take opposing positions on a given topic. This will help the researcher to obtain the views of both groups.
3. *Two way focus group:* Under this method one group will listen to the other group. Later, the second group will react to the views of the first group.
4. *Dual moderator group:* Here, there are two moderators. One moderator will make sure that the discussion moves smoothly. The second moderator will ask a specific question.

4. ***Case Studies:*** Analysing a selected case sometimes gives an insight into the problem which is being researched. Case histories of companies which have undergone a similar situation may be available. These case studies are well suited to carry out exploratory research. However, the result of investigation of case histories are always considered suggestive, rather than conclusive. In case of preference to "ready to eat food", many case histories may be available in the form of previous studies made by competitors. We must carefully examine the already published case studies with regard to other variables such as price, advertisement, changes in the taste, etc.

3.2.4 Secondary Data

Secondary data is information gathered for purposes other than the completion of a research project. A variety of secondary information sources is available to the researcher gathering data on an industry, potential product applications and the market place. Secondary data is also used to gain initial insight into the research problem.

Secondary data analysis saves time that would otherwise be spent collecting data and, particularly in the case of quantitative data, provides larger and higher-quality databases than would be unfeasible for any individual researcher to collect on their own. In addition to that, analysts of social and economic change consider secondary data essential, since it is impossible to conduct a new survey that can adequately capture past change and/or developments.

Secondary data can be obtained from two different research strands:

1. ***Quantitative:*** Census, housing, social security as well as electoral statistics and other related databases.
2. ***Qualitative:*** Semi-structured and structured interviews, focus groups transcripts, field notes, observation records and other personal, research-related documents.

Notes Secondary data can also be helpful in the research design of subsequent primary research and can provide a baseline with which the collected primary data results can be compared to. Therefore, it is always wise to begin any research activity with a review of the secondary data.

Notes

Secondary data is classified in terms of its source - either internal or external. Internal, or in-house data, is secondary information acquired within the organization where research is being carried out. External secondary data is obtained from outside sources.

Internal Data Sources

Internal secondary data is usually an inexpensive information source for the company conducting research, and is the place to start for existing operations. Internally generated sales and pricing data can be used as a research source. The use of this data is to define the competitive position of the firm, an evaluation of a marketing strategy the firm has used in the past, or gaining a better understanding of the company's best customers.

There are three main sources of internal data. These are:

1. ***Sales and marketing reports:*** These can include such things as:
   (a) Type of product/service purchased
   (b) Type of end-user/industry segment
   (c) Method of payment
   (d) Product or product line
   (e) Sales territory
   (f) Salesperson
   (g) Date of purchase
   (h) Amount of purchase
   (i) Price
   (j) Application by product
   (k) Location of end-user
2. ***Accounting and financial records:*** These are often an overlooked source of internal secondary information and can be invaluable in the identification, clarification and prediction of certain problems. Accounting records can be used to evaluate the success of various marketing strategies such as revenues from a direct marketing campaign.

   There are several problems in using accounting and financial data. One is the timeliness factor – it is often several months before accounting statements are available. Another is the structure of the records themselves. Most firms do not adequately setup their accounts to provide the types of answers to research questions that they need. For example, the account systems should capture project/product costs in order to identify the company's most profitable (and least profitable) activities.

   Companies should also consider establishing performance indicators based on financial data. These can be industry standards or unique ones designed to measure key performance factors that will enable the firm to monitor its performance over a period of time and compare it to its competitors. Some example may be sales per employee, sales per square foot, expenses per employee (salesperson, etc.).
3. ***Miscellaneous reports:*** These can include such things as inventory reports, service calls, number (qualifications and compensation) of staff, production and R&D reports. Also the company's business plan and customer calls (complaints) log can be useful sources of information.

External Data Sources

Notes

There is a wealth of statistical and research data available today. Some sources are:

1. Federal government
2. Provincial/state governments
3. Statistics agencies
4. Trade associations
5. General business publications
6. Magazine and newspaper articles
7. Annual reports
8. Academic publications
9. Library sources
10. Computerized bibliographies
11. Syndicated services.

The two major advantages of using secondary data in market research are time and cost savings.

1. The secondary research process can be completed rapidly - generally in 2 to 3 week. Substantial useful secondary data can be collected in a matter of days by a skillful analyst.
2. When secondary data is available, the researcher need only locate the source of the data and extract the required information.
3. Secondary research is generally less expensive than primary research. The bulk of secondary research data gathering does not require the use of expensive, specialized, highly trained personnel.
4. Secondary research expenses are incurred by the originator of the information.

There are also a number of disadvantages of using secondary data. These include:

1. Secondary information pertinent to the research topic is either not available, or is only available in insufficient quantities.
2. Some secondary data may be of questionable accuracy and reliability. Even government publications and trade magazines statistics can be misleading.
3. Data may be in a different format or units than is required by the researcher.
4. Much secondary data is several years old and may not reflect the current market conditions. Trade journals and other publications often accept articles six months before appear in print. The research may have been done months or even years earlier.

Did u know? Many trade magazines survey their members to derive estimates of market size, market growth rate and purchasing patterns, then average out these results. Often these statistics are merely average opinions based on less than 10% of their members.

Notes

3.2.5 Qualitative Research

Qualitative research seeks out the 'why', not the 'how' of its topic through the analysis of unstructured information – things like interview transcripts, e-mails, notes, feedback forms, photos and videos. It doesn't just rely on statistics or numbers, which are the domain of quantitative researchers.

Qualitative research is used to gain insight into people's attitudes, behaviours, value systems, concerns, motivations, aspirations, culture or life-styles. It's used to inform business decisions, policy formation, communication and research. Focus groups, in-depth interviews, content analysis and semiotics are among the many formal approaches that are used, but qualitative research also involves the analysis of any unstructured material, including customer feedback forms, reports or media clips.

Qualitative research is used to help us understand how people feel and why they feel as they do. It is concerned with collecting in-depth information asking questions such as why do you say that?. Samples tend to be smaller compared with quantitative projects that include much larger samples. Depth interviews or group discussions are two common methods used for collecting qualitative information.

Thus we can say that Qualitative research is a type of scientific research. In general terms, scientific research consists of an investigation that:

1. seeks answers to a question
2. systematically uses a predefined set of procedures to answer the question
3. collects evidence
4. produces findings that were not determined in advance
5. produces findings that are applicable beyond the immediate boundaries of the study

Qualitative research shares these characteristics. Additionally, it seeks to understand a given research problem or topic from the perspectives of the local population it involves. Qualitative research is especially effective in obtaining culturally specific information about the values, opinions, behaviors, and social contexts of particular populations.

The strength of qualitative research is its ability to provide complex textual descriptions of how people experience a given research issue. It provides information about the "human" side of an issue - that is, the often contradictory behaviors, beliefs, opinions, emotions, and relationships of individuals. Qualitative methods are also effective in identifying intangible factors, such as social norms, socioeconomic status, gender roles, ethnicity, and religion, whose role in the research issue may not be readily apparent. When used along with quantitative methods, qualitative research can help us to interpret and better understand the complex reality of a given situation and the implications of quantitative data. Although findings from qualitative data can often be extended to people with characteristics similar to those in the study population, gaining a rich and complex understanding of a specific social context or phenomenon typically takes precedence over eliciting data that can be generalized to other geographical areas or populations. In this sense, qualitative research differs slightly from scientific research in general.

The three most common qualitative methods, explained in detail in their respective modules, are participant observation, in-depth interviews, and focus groups. Each method is particularly suited for obtaining a specific type of data.

1. Participant observation is appropriate for collecting data on naturally occurring behaviors in their usual contexts.

Notes

2. In-depth interviews are optimal for collecting data on individuals' personal histories, perspectives, and experiences, particularly when sensitive topics are being explored.
3. Focus groups are effective in eliciting data on the cultural norms of a group and in generating broad overviews of issues of concern to the cultural groups or subgroups represented.

Task Enlist the basic differences between quantitative and qualitative research methods.

Self Assessment

Fill in the blanks:

3. The major emphasis in exploratory research is on converting ....., vague problem statements into ..... and ..... sub-problem statements.
4. Exploratory research is ..... and very .....
5. In experience surveys, it is desirable to talk to persons who are well informed in the area being .....
6. Most of the companies conducting the ..... groups first screen the candidates to determine who will compose the particular group.
7. The moderator must not miss the ..... comment.
8. The moderator must encourage ..... members to participate.

3.3 Descriptive Research Design

The name itself reveals that, it is essentially a research to describe something. For example, it can describe the characteristics of a group such as – customers, organisations, markets, etc. Descriptive research provides "association between two variables" like income and place of shopping, age and preferences.

Descriptive inform us about the proportions of high and low income customers in a particular territory. What descriptive research cannot indicate is that it cannot establish a cause and effect relationship between the characteristics of interest. This is the distinct disadvantage of descriptive research.

Descriptive study requires a clear specification of "Who, what, when, where, why and how" of the research. For example, consider a situation of convenience stores (food world) planning to open a new outlet. The company wants to determine, "How people come to patronize a new outlet?" Some of the questions that need to be answered before data collection for this descriptive study are as follows:

1. Who? Who is regarded as a shopper responsible for the success of the shop, whose demographic profile is required by the retailer?
2. What? What characteristics of the shopper should be measured?
3. Is it the age of the shopper, sex, income or residential address?
4. When? When shall we measure?
5. Should the measurement be made while the shopper is shopping or at a later time?

Notes

6. Where? Where shall we measure the shoppers?
7. Should it be outside the stores, soon after they visit or should we contact them at their residence?
8. Why? Why do you want to measure them?
9. What is the purpose of measurement? Based on the information, are there any strategies which will help the retailer to boost the sales? Does the retailer want to predict future sales based on the data obtained?
10. Answer to some of the above questions will help us in formulating the hypothesis.
11. How to measure? Is it a 'structured' questionnaire, 'disguised' or 'undisguised' questionnaire?

3.3.1 When to use Descriptive Study?

1. To determine the characteristics of market such as:
   (a) Size of the market
   (b) Buying power of the consumer
   (c) Product usage pattern
   (d) To find out the market share for the product
   (e) To track the performance of a brand.
2. To determine the association of the two variables such as Ad and sales.
3. To make a prediction. We might be interested in sales forecasting for the next three years, so that we can plan for training of new sales representatives.
4. To estimate the proportion of people in a specific population, who behave in a particular way?

Example: What percentage of population in a particular geographical location would be shopping in a particular shop?

Hypothesis study at the descriptive research stage (to demonstrate the characteristics of the group).

Management problem	Research problem	Hypothesis
How should a new product be distributed?	Where do customers buy a similar product right now?	Upper class buyers use 'Shopper's Stop' and middle class buyers buy from local departmental stores
What will be the target segment?	What kind of people buy our product now?	Senior citizens buy our products. Young and married buy our competitors products.

3.3.2 Types of Descriptive Studies

There are two types of descriptive research:

1. Longitudinal study
2. Cross-sectional study

Notes

1. ***Longitudinal Study:*** These are the studies in which an event or occurrence is measured again and again over a period of time. This is also known as 'Time Series Study'. Through longitudinal study, the researcher comes to know how the market changes over time.

   Longitudinal studies involve panels. Panel once constituted will have certain elements. These elements may be individuals, stores, dealers, etc. The panel or sample remains constant throughout the period. There may be some dropouts and additions. The sample members in the panel are being measured repeatedly. The periodicity of the study may be monthly or quarterly etc.

Example: For longitudinal study, assume a market research is conducted on ready to eat food at two different points of time T1 and T2 with a gap of 4 months. Each of the above two times, a sample of 2000 household is chosen and interviewed. The brands used most in the household is recorded as follows.

Brands	At T1	At T2
Brand X	500(25%)	600(30%)
Brand Y	700(35%)	650(32.5%)
Brand Z	400(20%)	300(15%)
Brand M	200(10%)	250(12.5%)
All others	200(10%)	250(12.5%)
	200	100%

As can be seen between period T1 and T2 Brand X and Brand M has shown an improvement in market share. Brand Y and Brand Z has decrease in market share, where as all other cate gories remains the same. This shows that Brand A and M has gained market share at the cost of Y and Z.

There are two types of panels: (a) True panel (b) Omnibus panel.

(a) ***True panel:*** This involves repeat measurement of the same variables. Example: Perception towards frozen peas or iced tea. Each member of the panel is examined at a different time, to arrive at a conclusion on the above subject.

(b) ***Omnibus panel:*** In omnibus panel too, a sample of elements is being selected and maintained, but the information collected from the member varies. At a certain point of time, the attitude of panel members "towards an advertisement" may be measured. At some other point of time the same panel member may be questioned about the "product performance".

Advantages of Panel Data

(a) We can find out what proportion of those who bought our brand and those who did not. This is computed using the brand switching matrix.

(b) The study also helps to identify and target the group which needs promotional effort.

(c) Panel members are willing persons, hence a lot of data can be collected. This is because becoming a member of a panel is purely voluntary.

Notes

(d) The greatest advantage of panel data is that it is analytical in nature.

(e) Panel data is more accurate than cross-sectional data because it is free from the error associated with reporting past behaviour. Errors occur in past behaviour because of time that has elapsed or forgetfulness.

Disadvantages of Panel Data

(a) The sample may not be representative. This is because sometimes, panels may be selected on account of convenience.

(b) The panel members who provide the data, may not be interested to continue as panel members. There could be dropouts, migration, etc. Members who replace them may differ vastly from the original member.

(c) Remuneration given to panel members may not be attractive. Therefore, people may not like to be panel members.

(d) Sometimes the panel members may show disinterest and non-committed.

(e) A lengthy period of membership in the panel may cause respondents to start imagining themselves to be experts and professionals. They may start responding like experts and consultants and not like respondents. To avoid this, no one should be retained as a member for more than 6 months.

2. ***Cross-sectional Study:*** Cross-sectional study is one of the most important types of descriptive research, it can be done in two ways:

(a) *Field study:* This includes a depth study. Field study involves an in-depth study of a problem, such as reaction of young men and women towards a product.

Example: Reaction of Indian men towards branded ready-to-wear suit. Field study is carried out in real world environment settings. Test marketing is an example of field study.

(b) *Field survey:* Large samples are a feature of the study. The biggest limitations of this survey are cost and time. Also, if the respondent is cautious, then he might answer the questions in a different manner. Finally, field survey requires good knowledge like constructing a questionnaire, sampling techniques used, etc.

Example: Suppose the management believes that geographical factor is an important attribute in determining the consumption of a product, like sales of a woolen wear in a particular location. Suppose that the proposition to be examined is that, the urban population is more likely to use the product than the semi-urban population. This hypothesis can be examined in a cross-sectional study. Measurement can be taken from a representative sample of the population in both geographical locations with respect to the occupation and use of the products. In case of tabulation, researcher can count the number of cases that fall into each of the following classes:

(i) Urban population which uses the product - Category I

(ii) Semi-urban population which uses the product - Category II

(iii) Urban population which does not use the product - Category III

(iv) Semi-urban population which does not use the product - Category IV.

Here, we should know that the hypothesis need to be supported and tested by the sample data i.e., the proportion of urbanities using the product should exceed the semi-urban population using the product.

Notes

3.3.3 Survey

The survey is a research technique in which data are gathered by asking questions of respondents. Survey research is one of the most important areas of measurement in applied social research. The broad area of survey research encompasses any measurement procedures that involve asking questions of respondents. A "survey" can be anything form a short paper-and-pencil feedback form to an intensive one-on-one in-depth interview.

Types of Surveys

Surveys can be divided into two broad categories: the questionnaire and the interview. Questionnaires are usually paper-and-pencil instruments that the respondent completes. Interviews are completed by the interviewer based on the respondent says. Sometimes, it's hard to tell the difference between a questionnaire and an interview. For instance, some people think that questionnaires always ask short closed-ended questions while interviews always ask broad open-ended ones. But you will see questionnaires with open-ended questions (although they do tend to be shorter than in interviews) and there will often be a series of closed-ended questions asked in an interview.

Survey research has changed dramatically in the last ten years. We have automated telephone surveys that use random dialing methods. There are computerized kiosks in public places that allows people to ask for input. A whole new variation of group interview has evolved as focus group methodology. Increasingly, survey research is tightly integrated with the delivery of service. Your hotel room has a survey on the desk. Your waiter presents a short customer satisfaction survey with your check. You get a call for an interview several days after your last call to a computer company for technical assistance. You're asked to complete a short survey when you visit a web site.

Selecting the Survey Method

Selecting the type of survey you are going to use is one of the most critical decisions in many social research contexts. You'll see that there are very few simple rules that will make the decision for you – you have to use your judgment to balance the advantages and disadvantages of different survey types. Here, all I want to do is give you a number of questions you might ask that can help guide your decision.

Population Issues

He first set of considerations have to do with the population and its accessibility.

1. ***Can the population be enumerated?***

   For some populations, you have a complete listing of the units that will be sampled. For others, such a list is difficult or impossible to compile. For instance, there are complete listings of registered voters or person with active drivers licenses. But no one keeps a complete list of homeless people. If you are doing a study that requires input from homeless persons, you are very likely going to need to go and find the respondents personally. In such contexts, you can pretty much rule out the idea of mail surveys or telephone interviews.

2. ***Is the population literate?***

   Questionnaires require that your respondents can read. While this might seem initially like a reasonable assumption for many adult populations, we know from recent research that the instance of adult illiteracy is alarmingly high. And, even if your respondents can

Notes

read to some degree, your questionnaire may contain difficult or technical vocabulary. Clearly, there are some populations that you would expect to be illiterate. Young children would not be good targets for questionnaires.

3. ***Are there language issues?***

   We live in a multilingual world. Virtually every society has members who speak other than the predominant language. Some countries (like Canada) are officially multilingual. And, our increasingly global economy requires us to do research that spans countries and language groups. Can you produce multiple versions of your questionnaire? For mail instruments, can you know in advance the language your respondent speaks, or do you send multiple translations of your instrument? Can you be confident that important connotations in your instrument are not culturally specific? Could some of the important nuances get lost in the process of translating your questions?

4. ***Will the population cooperate?***

   People who do research on immigration issues have a difficult methodological problem. They often need to speak with undocumented immigrants or people who may be able to identify others who are. Why would we expect those respondents to cooperate? Although the researcher may mean no harm, the respondents are at considerable risk legally if information they divulge should get into the hand of the authorities. The same can be said for any target group that is engaging in illegal or unpopular activities.

5. ***What are the geographic restrictions?***

   Is your population of interest dispersed over too broad a geographic range for you to study feasibly with a personal interview? It may be possible for you to send a mail instrument to a nationwide sample. You may be able to conduct phone interviews with them. But it will almost certainly be less feasible to do research that requires interviewers to visit directly with respondents if they are widely dispersed.

Sampling Issues

The sample is the actual group you will have to contact in some way. There are several important sampling issues you need to consider when doing survey research.

1. ***What data is available?:*** What information do you have about your sample? Do you know their current addresses? Their current phone numbers? Are your contact lists up to date?

2. ***Can respondents be found?:*** Can your respondents be located? Some people are very busy. Some travel a lot. Some work the night shift. Even if you have an accurate phone or address, you may not be able to locate or make contact with your sample.

3. ***Who is the respondent?:*** Who is the respondent in your study? Let's say you draw a sample of households in a small city. A household is not a respondent. Do you want to interview a specific individual? Do you want to talk only to the "head of household" (and how is that person defined)? Are you willing to talk to any member of the household? Do you state that you will speak to the first adult member of the household who opens the door? What if that person is unwilling to be interviewed but someone else in the house is willing? How do you deal with multi-family households? Similar problems arise when you sample groups, agencies, or companies. Can you survey any member of the organization? Or, do you only want to speak to the Director of Human Resources? What if the person you would like to interview is unwilling or unable to participate? Do you use another member of the organization?

Notes

4. ***Can all members of population be sampled?:*** If you have an incomplete list of the population (i.e., sampling frame) you may not be able to sample every member of the population. Lists of various groups are extremely hard to keep up to date. People move or change their names. Even though they are on your sampling frame listing, you may not be able to get to them. And, it's possible they are not even on the list.

5. ***Are response rates likely to be a problem?:*** Even if you are able to solve all of the other population and sampling problems, you still have to deal with the issue of response rates. Some members of your sample will simply refuse to respond. Others have the best of intentions, but can't seem to find the time to send in your questionnaire by the due date. Still others misplace the instrument or forget about the appointment for an interview. Low response rates are among the most difficult of problems in survey research. They can ruin an otherwise well-designed survey effort.

Question Issues

Sometimes the nature of what you want to ask respondents will determine the type of survey you select.

1. ***What types of questions can be asked?***

   Are you going to be asking personal questions? Are you going to need to get lots of detail in the responses? Can you anticipate the most frequent or important types of responses and develop reasonable closed-ended questions?

2. ***How complex will the questions be?***

   Sometimes you are dealing with a complex subject or topic. The questions you want to ask are going to have multiple parts. You may need to branch to sub-questions.

3. ***Will screening questions be needed?***

   A screening question may be needed to determine whether the respondent is qualified to answer your question of interest. For instance, you wouldn't want to ask someone their opinions about a specific computer program without first "screening" them to find out whether they have any experience using the program. Sometimes you have to screen on several variables (e.g., age, gender, experience). The more complicated the screening, the less likely it is that you can rely on paper-and-pencil instruments without confusing the respondent.

4. ***Can question sequence be controlled?***

   Is your survey one where you can construct in advance a reasonable sequence of questions? Or, are you doing an initial exploratory study where you may need to ask lots of follow-up questions that you can't easily anticipate?

5. ***Will lengthy questions be asked?***

   If your subject matter is complicated, you may need to give the respondent some detailed background for a question. Can you reasonably expect your respondent to sit still long enough in a phone interview to ask your question?

6. ***Will long response scales be used?***

   If you are asking people about the different computer equipment they use, you may have to have a lengthy response list (CD-ROM drive, floppy drive, mouse, touch pad, modem, network connection, external speakers, etc.). Clearly, it may be difficult to ask about each of these in a short phone interview.

Notes

Content Issues

The content of your study can also pose challenges for the different survey types you might utilize.

1. ***Can the respondents be expected to know about the issue?***

   If the respondent does not keep up with the news (e.g., by reading the newspaper, watching television news, or talking with others), they may not even know about the news issue you want to ask them about. Or, if you want to do a study of family finances and you are talking to the spouse who doesn't pay the bills on a regular basis, they may not have the information to answer your questions.

2. ***Will respondent need to consult records?***

   Even if the respondent understands what you're asking about, you may need to allow them to consult their records in order to get an accurate answer. For instance, if you ask them how much money they spent on food in the past month, they may need to look up their personal check and credit card records. In this case, you don't want to be involved in an interview where they would have to go look things up while they keep you waiting (they wouldn't be comfortable with that).

Bias Issues

People come to the research endeavor with their own sets of biases and prejudices. Sometimes, these biases will be less of a problem with certain types of survey approaches.

1. ***Can social desirability be avoided?***

   Respondents generally want to "look good" in the eyes of others. None of us likes to look like we don't know an answer. We don't want to say anything that would be embarrassing. If you ask people about information that may put them in this kind of position, they may not tell you the truth, or they may "spin" the response so that it makes them look better. This may be more of a problem in an interview situation where they are face-to face or on the phone with a live interviewer.

2. ***Can interviewer distortion and subversion be controlled?***

   Interviewers may distort an interview as well. They may not ask questions that make them uncomfortable. They may not listen carefully to respondents on topics for which they have strong opinions. They may make the judgment that they already know what the respondent would say to a question based on their prior responses, even though that may not be true.

3. ***Can false respondents be avoided?***

   With mail surveys it may be difficult to know who actually responded. Did the head of household complete the survey or someone else? Did the CEO actually give the responses or instead pass the task off to a subordinate? Is the person you're speaking with on the phone actually who they say they are? At least with personal interviews, you have a reasonable chance of knowing who you are speaking with. In mail surveys or phone interviews, this may not be the case.

Administrative Issues

Last, but certainly not least, you have to consider the feasibility of the survey method for your study.

Notes

1. ***Costs:*** Cost is often the major determining factor in selecting survey type. You might prefer to do personal interviews, but can't justify the high cost of training and paying for the interviewers. You may prefer to send out an extensive mailing but can't afford the postage to do so.
2. ***Facilities:*** Do you have the facilities (or access to them) to process and manage your study? In phone interviews, do you have well-equipped phone surveying facilities? For focus groups, do you have a comfortable and accessible room to host the group? Do you have the equipment needed to record and transcribe responses?
3. ***Time:*** Some types of surveys take longer than others. Do you need responses immediately (as in an overnight public opinion poll)? Have you budgeted enough time for your study to send out mail surveys and follow-up reminders, and to get the responses back by mail? Have you allowed for enough time to get enough personal interviews to justify that approach?
4. ***Personnel:*** Different types of surveys make different demands of personnel. Interviews require interviewers who are motivated and well-trained. Group administered surveys require people who are trained in group facilitation. Some studies may be in a technical area that requires some degree of expertise in the interviewer.

Clearly, there are lots of issues to consider when you are selecting which type of survey you wish to use in your study. And there is no clear and easy way to make this decision in many contexts. There may not be one approach which is clearly the best. You may have to make tradeoffs of advantages and disadvantages. There is judgment involved. Two expert researchers may, or the very same problem or issue, select entirely different survey methods. But, if you select a method that isn't appropriate or doesn't fit the context, you can doom a study before you even begin designing the instruments or questions themselves.

3.3.4 Observation Studies

An observational study draws inferences about the possible effect of a treatment on subjects, where the assignment of subjects into a treated group versus a control group is outside the control of the investigator. This is in contrast with controlled experiments, such as randomized controlled trials, where each subject is randomly assigned to a treated group or a control group before the start of the treatment.

Observational studies are sometimes referred to as natural experiments or as quasi-experiments. These differences in terminology reflect certain differences in emphasis, but a shared theme is that the early stages of planning or designing an observational study attempt to reproduce, as nearly as possible, some of the strengths of an experiment.

Self Assessment

Fill in the blanks:

9. ..... studies are the studies in which an event or occurrence is measured again and again over a period of time.
10. Longitudinal study is also known as .....
11. True panel involves ..... measurement of the same variables.
12. The biggest limitations of field survey are ..... and .....

Notes

3.4 Difference between Exploratory Research and Descriptive Research

Table 3.1

Exploratory Research	Descriptive Research
It is concerned with the "Why" aspect of consumer behaviour i.e., it tries to understand the problem and not measure the result.	It is concerned with the "What", "When" or "How often" on the consumer behaviour.
This research does not require large samples.	This needs large samples of respondents.
Sample need not to represent the population.	Sample must be representative of population.
Due to imprecise statement, data collection is not easy.	Statement is precise. Therefore data collection is easy
Characteristics of interest to be measured is not clear.	Characteristics of interest to be measured is clear.
There is no need for a questionnaire for collecting the data.	There should be a properly designed questionnaire for data collection.
Data collection methods are: Focus group Literature Searching Case study	Use of panel data Longitudinal Cross-sectional studies

Self Assessment

Fill in the blanks:

13. …….research requires large samples.

14. In ……research, there is no need for a questionnaire for collecting the data.

3.5 Causal Research Design

Causal Research are the studies that engage in hypotheses testing usually explain the nature of certain relationships, or establish the differences among groups or the independence of two or more factors in a situation. A research design in which the major emphasis is on determining a cause-and-effect relationship. The research is used to measure what impact a specific change will have on existing norms and allows market researchers to predict hypothetical scenarios upon which a company can base its business plan.

Example: If a clothing company currently sells blue denim jeans, causal research can measure the impact of the company changing the product design to the colour white.

Following the research, company bosses will be able to decide whether changing the colour of the jeans to white would be profitable.

To summarise, causal research is a way of seeing how actions now will affect a business in the future. Nevertheless, it has to be remembered that not all causal research hypotheses can be

Notes

studied. There are many reasons for this, one of them being that true random assignment is not possible in many cases. The three main reasons why you can't test everything deal with:

1. ***Technology,*** or the impossibility by today's technology to be able to do certain tasks, such as assign gender.
2. ***Ethics,*** because we can't randomly assign that some people receive a virus to test its effects, or that some participants have to act as slaves and others as masters to test a hypothesis, and
3. ***Resources,*** if a researcher does not have the money or the equipment needed to perform a study, then it won't be done.

Causal design is the study of cause and effect relationships between two or more variables.

William J. Goode & Paul K. Hatt in Methods in Social Research define cause and effect relationship as:

"*when two or more cases of given phenomenon have one and only one condition in common, that condition may be regarded as the cause and effect of that phenomenon*."

The set of causes generated to predict their effects, can be deterministic or probabilistic in nature. The deterministic cause is the one which is essential and adequate for stimulating the occurrence of another event. While the probabilistic is the one that is essential, but is not the only one responsible for the stimulation of the occurrence of another event.

The objective is to determine which variable might be causing certain behaviour i.e., whether there is a cause and effect relationship between variables, causal research must be undertaken. This type of research is very complex and the researcher can never be completely certain that there are not other factors influencing the causal relationship, especially when dealing with people's attitudes and motivations. There are often much deeper psychological considerations that even the respondent may not be aware of.

In marketing decision making, all the conditions allowing the most accurate casual statements are not usually present but in these circumstances, casual inference will still be made by marketing managers. Because in doing so they would want to be able to make casual statements about the effects of their actions.

Example: The new advertising campaign a company developed has resulted in percentage increase in sales or the sales discount strategy a company followed has resulted in percentage increase in sales. In both of these examples, marketing managers are making a casual statement.

However, the scientific concept of casuality is complex and differs substantiality from the one held by the common person on the street. The common sense view holds that a single event (the cause) always results in another event (the effect) occurring. In science, we recognize that an event has a number of determining conditions or causes which act together to make the event probable. Note that the common sense notion of casuality is that the effect always follows the cause. This is deterministic causation in contrast to scientific notion which specifies the effect only as being probable. This is termed as probabilistic causation. The scientific notion holds that we can only infer casuality and never really prove it. That is the chance of an incorrect inference is always thought to exist. The world of marketing fits the scientific view of casuality. Marketing effects are probabilistically caused by multiple factors and we can only infer a casual relationship. The condition under which we can make casual inference are:

(a) Time and order of occurrence of variables.

(b) Concomitant variation

(c) Elimination of other possible causal factors.

Notes

Causal research design are used to provide a stronger basis for the existence of causal relationship between variables. The researcher is able to control the influence of one or more extraneous variables on the dependent variable. If it is not possible to control the influence of an extraneous variable on the dependent variable, that variable is called confounded variable.

Caution Gender cannot be randomly assigned, and therefore already you cannot test all causal hypotheses.

How to Prepare a Synopsis

Synopsis is an abstract form of research which underlines the research procedure followed and is presented before the guide for evaluating its potentiality. In one sentence it may be described as a condensation of the final report. The structure of synopsis varies and also depends on the guides' choice. However, for our understanding a common structure may be framed as under:

1. ***Defining the Problem:*** In defining the problem of the research objective, definition of key terms, general background information, limitations of the study and order of presentation should be mentioned in brief.
2. ***Review of Existing Literature:*** In this head, researcher should study the summary of different points of view on the subject matter as found in books, periodicals and approach to be followed at the time of writing.
3. ***Conceptual Framework and Methodology:*** Under this head the researcher should first make a statement of the hypothesis. Discussion on the research methodology used, duly pointing out the relationship between the hypothesis and objective of the study and finally discussions about the sources and means of obtaining data should also be made. In this head the researcher should also point out the limitations of methodology, if any, and the natural crises from which the research is bound to suffer for such obvious limitations.
4. ***Analysis of Data:*** Analysis of the data involves testing of hypothesis from data collected and key conclusions thus arrived.
5. ***General Conclusions:*** In general conclusions, the researcher should make a restatement of objectives. Conclusion with respect to the acceptance or rejection of hypothesis, conclusion with respect to the stated objectives, suggested areas of further research and final discussion of possible implications of the study for a model, group, theory and discipline.

Finally the researcher should mention about the bibliographies and appendices. The above format is drawn after a standard framework followed internationally in preparation of a synopsis. However, in our country, keeping in view the object of research, style and structure of synopsis varies and quite often it is found that the research guide exercises his own discretion in synopsis preparation than following some acceptable international norms. A standard format for preparation of synopsis commonly used in management and commerce research in India may be drawn as follows:

1. ***Introduction:*** This includes definition of the problem and its review from a historical perspective.
2. ***Objective of the Study:*** It defines the research purpose and its speciality from the existing available research in the related field.
3. ***Literature Review:*** It includes among other things, different sources from which the required abstract is drawn.

Notes

4. ***Methodology:*** It is intended to draw out the sequences followed in research and ways and manners of carrying out the survey and compilation of data.
5. ***Hypothesis:*** It is a formal statement relating to the research problem and it need to be tested based on the researchers' findings.
6. ***Model:*** It underlies the nature and structure of the model that the researcher is going to build in the light of survey findings.

Self Assessment

Fill in the blanks:

15. ……. research is a way of seeing how actions now will affect a business in the future.
16. Synopsis is an abstract form of research which underlines the research procedure followed and is presented before the guide for evaluating its …….

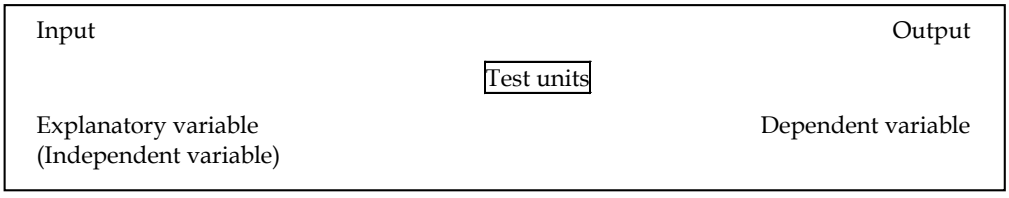
3.6 Experimentation

Experimentation Research is also known as causal research. Descriptive research, will suggest the relationship if any between the variable, but it will not establish cause and effect relationship between the variable. Example: The data collected may show that the no. of people who own a car and their income has risen over a period of time. Despite this, we cannot say "No. of car increase is due to rise in the income". May be, improved road conditions or increase in number of banks offering car loans have caused in increase in the ownership of cars.

To find the causal relationship between the variables, the researcher has to do an experiment.

Example:

1. Which print advertisement is more effective? Is it front page, middle page or the last page?
2. Among several promotional measure, such as Advertisement, personal selling, "which one is more effective"? Can we increase sales of our product by obtaining additional shelf space? What is experimentation? It is research process in which one or more variables are manipulated, which shows the cause and effect relationship. Experimentation is done to find out the effect of one factor on the other. The different elements of experiment are explained below.



Test Units

These are units, on which the experiment is carried out. It is done, with one or more independent variables controlled by a person to find out its effect, on a dependent variable.

Notes

Explanatory Variable

These are the variables whose effects, researcher wishes to examine. For example, explanatory variables may be advertising, pricing, packaging etc.

Dependent Variable

This is a variable which is under study. For example, sales, consumer attitude, brand loyalty etc.

Example: Suppose a particular colour TV manufacturer reduces the price of the TV by 20%. Assume that his reduction is passed on to the consumer and expect the sales will go up by 15% in next 1 year. This types of experiments are done by leading TV companies during festival season.

The causal research finds out, whether the price reduction causes an increase in sales.

Extraneous Variables

These are also called as blocking variables Extraneous variables affects, the result of the experiments.

Example:

1. Suppose a toffee manufacturing company is making an attempt to measure the response of the buyers, on two different types of packaging, at two different locations. The manufacturer needs to keep all other aspects the same, for each buyers group. If the manufacturer allows the extraneous variable namely the "Price", to vary between the two buyer groups, then he will not be sure, as to which particular packaging is preferred by the consumers. Here prices change is an extraneous factor.

   There are two possible courses of action with respect to extraneous variables.

   Extraneous variables may be physically controlled. Example: Price in the above example.

   In the second category, extraneous variables may be totally out of control of the researcher. In this case, we say that the experiment has been confounded i.e., it is not possible to make any conclusions with regard to that experiment. Such a variable is called as "Confounding variables".

2. Company introduces a product in two different cities. They would like to know the impact of their advertising on sales. Simultaneously competitors product in one of the cities is not available during this period due to strike in the factory. Now researcher cannot conclude that sales of their product in that city has increased due to advertisement. Therefore this experiment is confounded. In this case, strike is the confounding variable.

Types of Extraneous Variables

The following are the various types:

1. History
2. Maturation
3. Testing
4. Instrument variation

Notes

5. Selection bias

6. Experimental mortality

1. ***History:*** History refers to those events, external to the experiment, but occurs at the same time, as the experiment is being conducted. This may affect the result. Example: Let us say that, a manufacture makes a 20% cut in the price of a product and monitors sales in the coming weeks. The purpose of the research, is to find the impact of price on sales. Mean while if the production of the product declines due to shortage of raw materials, then the sales will not increase. Therefore, we cannot conclude that the price cut, did not have any influence on sales because the history of external events have occurred during the period and we cannot control the event. The event can only be identified.

2. ***Maturation:*** Maturation is similar to history. Maturation specifically refers to changes occurring within the test units and not due to the effect of experiment. Maturation takes place due to passage of time. Maturation refers to the effect of people growing older. People may be using a product. They may discontinue the product usage or switch over to alternate product.

Example:

(a) Pepsi is consumed when young. Due to passage of time the consumer becoming older, might prefer to consume Diet pepsi or even avoid it.

(b) Assume that training programme is conducted for sales man, the company wants to measure the impact of sales programme. If the company finds that, the sales have improved, it may not be due to training programme. It may be because, sales man have more experience now and know the customer better. Better understanding between sales man and customer may be the cause for increased sales.

Maturation effect is not just limited to test unit, composed of people alone. Organizations also changes, dealers grow, become more successful, diversify, etc.

3. ***Testing:*** Pre-testing effect occurs, when the same respondents are measured more than once. Responses given at a later part will have a direct bearing on the responses given during earlier measurement.

Example: Consider a respondent, who is given an initial questionnaire, intended to measure brand awareness. After exposing him, if a second questionnaire similar to the initial questionnaire is given to the respondent, he will respond quiet differently, because of respondent's familiarity with the earlier questionnaire.

Pretest suffers from internal validity. This can be understood through an example. Assume that a respondent's opinion is measured before and after the exposure to a TV commercial of Hyundai car with Shahrukh Khan as brand ambassador. When the respondent is replying the second time, He may remember, how he rated Hyundai during the first measurement. He may give the same rating to prove that, he is consistent. In that case, the difference between the two measurements will reveal nothing about the real impact.

Alternately some of respondents might give a different rating during second measurement. This may not be due to the fact that the respondent has changed his opinion about Hyundai and the brand ambassador. He has given different rating because, he does not want to be identified as a person with no change of opinion to the said commercial.

In both the cases of above, internal validity suffers.

Notes

4. ***Instrument Variation:*** Instrument variation effect is a threat to internal validity when human respondents are involved. For example, an equipment such as a vacuum cleaner is left behind, for the customer to use for two weeks. After two weeks the respondents are given a questionnaire to answer. The reply may be quite different from what was given by the respondent before the trail of the product. This may be because of two reasons:

   (a) Some of the questions have been changed

   (b) Change in the interviewer for pre-testing and post testing are different

   The measurement in experiments will depend upon the instrument used to measure. Also results may vary due to application of instruments, where there are several interviewers. Thus, it is very difficult to ensure that all the interviewers will ask the same questions with the same tone and develop the same rapport. There may be difference in response, because each interviewer conducts the interview differently.

5. ***Selection Bias:*** Selection bias occurs because 2 groups selected for experiment may not be identical. If the 2 groups are asked various questions, they will respond differently. If multiple groups are participating, this error will occur. There are two promotional advertisement A & B for "Ready to eat food". The idea is to find effectiveness of the two advertisements. Assume that the respondent exposed to 'A' are dominant users of the product. Now suppose 50% of those who saw 'Advertisement A' bought the product and only 10% of those who saw 'Advertisement B' bought the product. From the above, one should not conclude that advertisement 'A' is more effective than advertisement 'B'. The main difference may be due to food preference habits between the groups, even in this case, internal validity might suffer but to a lesser degree.

6. ***Experimental Mortality:*** Some members may leave the original group and some new members join the old group. This is because some members might migrate to another geographical area. This change in the members will alter the composition of the group.

Example: Assume that a vacuum cleaner manufacturer wants to introduce a new version. He interviews hundred respondents who are currently using the older version. Let us assume that, these 100 respondents have rated the existing vacuum cleaner on a 10 point scale (1 for lowest and 10 for highest). Let the mean rating of the respondents be 7.

Now the newer version is demonstrated to the same hundred respondents and equipment is left with them for 2 months. At the end of two months only 80 participant respond, since the remaining 20 refused to answer. Now if the mean score of 80 respondents is 8 on the same 10 point scale. From this can we conclude that the new vacuum cleaner is better?

The answer to the above question depends on the composition of 20 respondents who dropped out. Suppose the 20 respondents who dropped out had negative reaction to the product, then the mean score would not have been 8. It may even be lower than 7. The difference in mean rating does not give true picture. It does not indicate that the new product is better than the old product.

One might wonder, why not we leave the 20 respondent from the original group and calculate the mean rating of the remaining 80 and compare. But this method also will not solve the mortality effect. Mortality effect will occur in an experiment irrespective of whether the human beings or involved or not.

Concomitant Variable

Concomitant variable is the extent to which a cause "X" and the effect "Y" vary together in a predicted manner.

Notes

Example:

1. Electrical car is new to India. People may or may not hold positive attitude about electrical cars. Assume that, the company has undertaken a new advertising campaign "To change the attitude of the people towards this car", so that the sale of this car can increase. Suppose, in testing the result of this campaign, the company finds that both aims have been achieved i.e., the attitude of the people towards electrical car has become positive and also the sales have increased. Then we can say that there is a concomitant variation between attitude and sales. Both variables move in the same direction.
2. Assume that an education institute introduces a new elective which it claims is Job oriented. The college authorities advertise this course in leading news paper. They would like to know the perception of students to this course, and how many are willing to enroll. Now if on testing, it is found the perception towards this course is positive and majority of the respondent are willing to enroll, then we can say that, there is a concomitant variation between perception and enrolment. Both variables move in the same direction.

3.6.1 Experimental Designs

The various experimental designs are as follows:

1. After only design
2. Before-after design
3. Factorial design
4. Latin square design
5. Ex-post facto design

After only Design

In this design, dependent variable is measured, after exposing the test units to the experimental variable. This can be understood with the help of following example.

Assume M/s Hindustan Lever Ltd. wants to conduct an experiment on "Impact of free sample on the sale of toilet soaps". A small sample of toilet soap is mailed to a selected set of customers in a locality. After one month, 25 paise off on one cake of soap coupon is mailed to each of the customers to whom free sample has been sent earlier. An equal number of these coupons are also mailed, to people in another similar locality in the neighborhood. The coupons are coded, to keep an account of the number of coupons redeemed from each locality. Suppose, 400 coupons were redeemed from the experimental group and 250 coupons are redeemed from the control group. The difference of 150 is supposed to be the effect of the free samples. In this method conclusion can be drawn only after conducting the experiment.

Before-after Design

In this method, measurements are made before as well as after.

Example: Let us say that, an experiment is conducted to test an advertisement which is aimed at reducing the alcoholism.

Notes

Attitude and perception towards consuming liquor is measured before exposure to Ad. The group is exposed to an advertisement, which tells them the consequences, and attitude is again measured after several days. The difference, if any, shows the effectiveness of advertisement.

The above example of "Before-after" suffers from validity threat due to the following:

1. ***Before measure effect:*** It alerts the respondents to the fact that they are being studied. The respondents may discuss the topics with friends and relatives and change their behaviour.
2. ***Instrumentation effect:*** This can be due to two different instruments being used, one before and one after, change in the interviewers before and after, results in instrumentation effect.

Factorial Design

Factorial design permits the researcher to test two or more variables at the same time. Factorial design helps to determine the effect of each of the variables and also measure the interacting effect of the several variables.

Example: A departmental store wants to study the impact of price reduction for a product. Given that, there is also promotion (POP) being carried out in the stores (a) near the entrance (b) at usual place, at the same time. Now assume that there are two price levels namely regular price A_1 and reduced price A_2. Let there be three types of POP namely B_1, B_2, & B_3. There are 3 × 2 = 6 combinations possible. The combinations possible are B_1A_1, B_1A_2, B_2A_1, B_2A_2, B_3A_1, B_3A_2. Which of these combinations is best suited is what the researcher is interested. Suppose there are 60 departmental stores of the chain divided into groups of 10 stores. Now, randomly assign the above combination to each of these 10 stores as follows:

Combinations	Sales
B_1A_1	S_1
B_1A_2	S_2
B_2A_1	S_3
B_2A_2	S_4
B_3A_1	S_5
B_3A_2	S_6

S_1 TO S_6 represents the sales resulting out of each variable. The data gathered will provide details on product sales on account of two independent variables.

The two questions that will be answered are:

1. Is the reduced price more effective than regular price?
2. Is the display at the entrance more effective than the display at usual location? Also the research will tell us about the interaction effect of the two variables.

Out come of the experiment on sales is as follows:

1. Price reduction with display at the entrance.
2. Price reduction with display at usual place.
3. No display and regular price applicable
4. Display at the entrance with regular price applicable.

Notes

Latin Square Design

Researcher chooses 3 shelf arrangements in three stores. He would like to observe the sales generated in each stores at different period. Researcher must make sure that one type of shelf arrangement is used in each store only once.

In Latin square design, only one variable is tested. As an example of Latin square design assume that a super market chain is interested in the effect of in store promotion on sales. Suppose there are three promotions considered as follows:

1. No promotion
2. Free sample with demonstration
3. Window display

Which of the 3 will be effective? The out come may be affected by the size of the stores and the time period. If we choose 3 stores and 3 time periods, the total number of combination is 33 = 9. The arrangement is as follows:

Time period	Store		
	1	2	3
1	B	C	A
2	A	B	C
3	A	B	C

Latin square is concerned with effectiveness of each kind of promotion on sales.

Ex-post Facto Design

This is a variation of "after only design". The groups such as experiment and control are identified only after they are exposed to the experiment.

Let us assume that a magazine publisher wants to know the impact of advertisement on knitting in 'Women's Era' magazine. The subscribers of magazines are asked whether they have seen this advertisement on "knitting". Those who have read and not read, are asked about the price, design etc. of the product. The difference indicates the effectiveness of advertisement. In this design, the experimental group is set to receive the treatment rather than exposing it to the treatment by its choice.

Self Assessment

Fill in the blanks:

17. Explanatory variable are the variables whose effects, researcher wishes to .....
18. ……are units, on which the experiment is carried out.
19. ……..design helps to determine the effect of each of the variables and also measure the interacting effect of the several variables.

3.7 Summary

- There are primarily four types of research namely exploratory research, descriptive research, Casual and experimental research.

Notes

- Exploratory research helps the researcher to become familiar with the problem. It helps to establish the priorities for further research. It may or may not be possible to formulate Hypothesis during exploratory stage.
- To get an insight into the problem, literature search, experience surveys, focus groups, and selected case studies assist in gaining insight into the problem.
- The role of moderator or facilitator is extremely important in focus group. There are several variations in the formation of focus group.
- Descriptive research is rigid. This type of research is basically dependent on hypothesis.
- Descriptive research is used to describe the characteristics of the groups. It can also be used forecasting or prediction.
- Panel data is used in longitudinal studies. There are two different types of panels. True panel and Omnibus panel. In true panel same measurement are made during period of time. In Omnibus panel different measurement are made during a period of time.
- Cross-sectional studies involves field study and field survey, the difference being the size of sample.
- Causal research is conducted mainly to prove the fact that one factor "X" the cause was responsible for the effect "Y".
- While conducting experiment, the researcher must guard against extraneous source of error. This may confound the experiment.

3.8 Keywords

Causal Research: A research designed to determine cause and effect relationship.

Conclusive Research: This is a research having clearly defined objectives. In this type of research, specific courses of action are taken to solve the problem.

Concomitant Variation: It is the extent to which cause and effect vary together.

Descriptive Research: It is essentially a research to describe something.

Ex-post Facto Research: Study of the current state and factors causing it.

Extraneous Variable: These variables affect the response of test units. Also known as confounding variable.

Field Study: Field study involves an in-depth study of a problem, such as reaction of young men and women towards a product.

Literature Research: It refers to "referring to a literature to develop a new hypothesis".

Longitudinal Study: These are the studies in which an event or occurrence is measured again and again over a period of time.

3.9 Review Questions

1. Can all causal research hypotheses be studied? Why or why not?
2. For each of the situation mentioned below, state whether the research should be exploratory, descriptive or causal and why

   (a) To find out the relationship between promotion and sales.

Notes

(b) To find out the consumer reaction regarding use of new detergents which are economical

(c) To identify the target market demographics, for a shopping mall.

(d) Estimate the sales potential for ready-to-eat food in the northeastern parts of India.

3. In your analysis, what are the advantages and disadvantages of panel data?
4. What do you see as the reason behind Latin Square Design testing only one variable?
5. Do you see any benefit of factorial design over that of before-after design? Support your answer with reasons.
6. Is it necessary for the researcher to mention about the bibliographies and appendices? Why/why not?
7. Illustrate advantages of experience survey by the help of examples.
8. Why is an exploratory research used in the initial stages of research?
9. Which type of research would you use to generate new product ideas and why?
10. Which type of research study would you use to determine the characteristics of market?

Answers: Self Assessment

1.	Exploratory	2.	analyzing
3.	broad, small, precise	4.	flexible, versatile
5.	investigated	6.	focus
7.	participant	8.	unresponsive
9.	Longitudinal	10.	'Time Series Study'
11.	repeat	12.	cost, time
13..	Descriptive	14.	exploratory
15.	Causal	16.	potentiality
17.	examine	18.	Test units
19.	Factorial		

3.10 Further Readings

Books

Cooper and Schinder, *Business Research Methods*, TMH.

CR Kotari, *Research Methodology*, Vishwa Prakashan.

David Luck and Ronald Rubin, *Marketing Research*, PHI.

Naresh Amphora, *Marketing Research*, Pearson Education.

S. N. Murthy and U. Bhojanna, *Business Research Methods*, Excel Books.

William MC Trochim, *Research Methods*, Biztantra.

William Zikmund, *Business Research Methods*, Thomson.

Notes

Unit 4: Sampling Design

CONTENTS

Objectives

After studying this unit, you will be able to:

- Describe the conception of sampling
- Steps involved in the sampling design
- Identify the characteristics of good sampling design
- State the different types of sampling design
- Report about the probability and non-probability sampling
- Explain the various types of errors in sampling
- Tell about determining of sampling size

Notes

Introduction

Sampling is the process of selecting units (e.g., people, organizations) from a population of interest so that by studying the sample we may fairly generalize our results back to the population from which they were chosen. Each observation measures one or more properties (weight, location, etc.) of an observable entity enumerated to distinguish objects or individuals. Survey weights often need to be applied to the data to adjust for the sample design. Results from probability theory and statistical theory are employed to guide practice.

4.1 Sampling – An Introduction

A sample is a part of a target population, which is carefully selected to represent the population. Sampling frame is the list of elements from which the sample is actually drawn. Actually, sampling frame is nothing but the correct list of population.


Example: Telephone directory, Product finder, Yellow pages.

The sampling process comprises several stages:

1. Defining the population of concern
2. Specifying a sampling frame, a set of items or events possible to measure
3. Specifying a sampling method for selecting items or events from the frame
4. Determining the sample size
5. Implementing the sampling plan
6. Sampling and data collecting
7. Reviewing the sampling process

4.1.1 Distinction between Census and Sampling

Census refers to complete inclusion of all elements in the population. A sample is a sub-group of the population.

When is a Census Appropriate?

1. A census is appropriate if the size of population is small.


Example: A researcher may be interested in contacting firms in iron and steel or petroleum products industry. These industries are limited in number, so a census will be suitable.

2. Sometimes, the researcher is interested in gathering information from every individual.


Example: Quality of food served in a mess.

When is Sample Appropriate?

1. When the size of population is large.
2. When time and cost are the main considerations in research.
3. If the population is homogeneous.

Notes

4. Also, there are circumstances when a census is not possible.


Example: Reactions to global advertising by a company.

Self Assessment

Fill in the blanks:

1. A sample is a part of a ..... population.
2. Sampling ..... is the list of elements from which the sample is actually drawn.
3. A sample is appropriate when the size of population is ..... and .....
4. A census is appropriate if the size of population is .....

4.2 Steps of Sampling Design

Sampling process consists of seven steps. They are:

1. Define the population
2. Identify the sampling frame
3. Specify the sampling unit
4. Selection of sampling method
5. Determination of sample size
6. Specify sampling plan
7. Selection of sample

1. ***Define the population:*** Population is defined in terms of:
   (a) Elements
   (b) Sampling units
   (c) Extent
   (d) Time.


Example: If we are monitoring the sale of a new product recently introduced by a company, say (shampoo sachet) the population will be:

(a) ***Element*** - Company's product

(b) ***Sampling unit*** - Retail outlet, super market

(c) ***Extent*** - Hyderabad and Secunderabad

(d) ***Time*** - April 10 to May 10, 2006

2. ***Identify the sampling frame:*** Sampling frame could be
   (a) Telephone Directory
   (b) Localities of a city using the municipal corporation listing
   (c) Any other list consisting of all sampling units.

Notes

Example: You want to learn about scooter owners in a city. The RTO will be the frame, which provides you names, addresses and the types of vehicles possessed.

3. ***Specify the sampling unit:*** Individuals who are to be contacted are the sampling units. If retailers are to be contacted in a locality, they are the sampling units.

   Sampling unit may be husband or wife in a family. The selection of sampling unit is very important. If interviews are to be held during office timings, when the heads of families and other employed persons are away, interviewing would under-represent employed persons, and over-represent elderly persons, housewives and the unemployed.

4. ***Selection of sampling method:*** This refers to whether

   (a) probability or

   (b) non-probability methods are used.

5. ***Determine the sample size:*** This means we need to decide "how many elements of the target population are to be chosen?" The sample size depends upon the type of study that is being conducted. For example: If it is an exploratory research, the sample size will be generally small. For conclusive research, such as descriptive research, the sample size will be large.

   The sample size also depends upon the resources available with the company.

Did u know? Sample size depends on the accuracy required in the study and the permissible errors allowed.

6. ***Specify the sampling plan:*** A sampling plan should clearly specify the target population. Improper defining would lead to wrong data collection.

Example: This means that, if a survey of a household is to be conducted, a sampling plan should define a "household" i.e., "Does the household consist of husband or wife or both", minors etc., "Who should be included or excluded." Instructions to the interviewer should include "How he should obtain a systematic sample of households, probability sampling non-probability sampling". Advise him on what he should do to the household, if no one is available.

7. ***Select the sample:*** This is the final step in the sampling process.

4.2.1 Characteristics of a Good Sample Design

A good sample design requires the judicious balancing of four broad criteria - goal orientation, measurability, practicality and economy.

1. ***Goal orientation:*** This suggests that a sample design "should be oriented to the research objectives, tailored to the survey design, and fitted to the survey conditions". If this is done, it should influence the choice of the population, the measurement as also the procedure of choosing a sample.

2. ***Measurability:*** A sample design should enable the computation of valid estimates of its sampling variability. Normally, this variability is expressed in the form of standard errors in surveys. However, this is possible only in the case of probability sampling. In non-probability samples, such a quota sample, it is not possible to know the degree of precision of the survey results.

Notes

3. ***Practicality:*** This implies that the sample design can be followed properly in the survey, as envisaged earlier. It is necessary that complete, correct, practical, and clear instructions should be given to the interviewer so that no mistakes are made in the selection of sampling units and the final selection in the field is not different from the original sample design. Practicality also refers to simplicity of the design, i.e. it should be capable of being understood and followed in actual operation of the field work.
4. ***Economy:*** Finally, economy implies that the objectives of the survey should be achieved with minimum cost and effort. Survey objectives are generally spelt out in terms of precision, i.e. the inverse of the variance of survey estimates. For a given degree of precision, the sample design should give the minimum cost. Alternatively, for a given per unit cost, the sample design should achieve maximum precision (minimum variance).

It may be pointed out that these four criteria come into conflict with each other in most of the cases,

Caution The researcher should carefully balance the conflicting criteria so that he is able to select a really good sample design.

Self Assessment

Fill in the blanks:

5. A sampling plan should clearly specify the ..... population.
6. The sample size depends upon the ..... available with the company.

4.3 Types of Sample Design

Sampling is divided into two types:

Probability sampling: In a probability sample, every unit in the population has equal chances for being selected as a sample unit.

Non-probability sampling: In the non-probability sampling, the units in the population have unequal or negligible, almost no chances for being selected as a sample unit.

4.3.1 Probability Sampling Techniques

1. Random sampling.
2. Systematic random sampling.
3. Stratified random sampling.
4. Cluster sampling.
5. Multistage sampling.

Random Sampling

Simple random sample is a process in which every item of the population has an equal probability of being chosen.

Notes

There are two methods used in the random sampling:

1. ***Lottery method:*** Take a population containing four departmental stores: A, B, C and D. Suppose we need to pick a sample of two stores from the population using a simple random procedure. We write down all possible samples of two. Six different combinations, each containing two stores from the population, are AB, AD, AC, BC, BD, CD. We can now write down six sample combination on six identical pieces of paper, fold the piece of paper so that they cannot be distinguished. Put them in a box. Mix them and pull one at random. This procedure is the lottery method of making a random selection.
2. ***Using random number table:*** A random number table consists of a group of digits that are arranged in random order, i.e., any row, column, or diagonal in such a table contains digits that are not in any systematic order.

Notes There are three tables for random numbers

(a) Tippet's table

(b) Fisher and Yate's table

(c) Kendall and Raington table.

The table for random number is as follows:

40743	39672
80833	18496
10743	39431
88103	23016
53946	43761
31230	41212
24323	18054

Example: Taking the earlier example of stores. We first number the stores.

1 A 2 B 3 C 4 D

The stores A, B, C and D have been numbered as 1, 2, 3 and 4.

We proceed as follows, in order to select two shops out of four randomly:

Suppose, we start with the second row in the first column of the table and decide to read diagonally. The starting digit is 8. There are no departmental stores with the number 8 in the population. There are only four stores. Move to the next digit on the diagonal, which is 0. Ignore it, since it does not correspond to any of the stores in the population. The next digit on the diagonal is 1 which corresponds to store A. Pick A and proceed until we get two samples. In this case, the two departmental stores are 1 and 4. The sample derived from this consists of departmental stores A and D.

In random sampling, there are two possibilities (a) Equal probability (b) Varying probability.

(a) ***Equal Probability:*** This is also called as the random sampling with replacement.

Notes

Example: Put 100 chits in a box numbered 1 to 100. Pick one number at random. Now the population has 99 chits. Now, when a second number is being picked, there are 99 chits. In order to provide equal probability, the sample selected is being replaced in the population.

(b) ***Varying Probability:*** This is also called random sampling without replacement. Once a number is picked, it is not included again. Therefore, the probability of selecting a unit varies from the other. In our example, it is 1/100, 1/99, 1/98, 1/97 if we select four samples out of 100.

Systematic Random Sampling

There are three steps:

1. Sampling interval K is determined by the following formula:

$$K = \frac{\text{No. of units in the population}}{\text{No. of units desired in the sample}}$$

2. One unit between the first and Kth unit in the population list is randomly chosen.
3. Add Kth unit to the randomly chosen number.

Example: Consider 1,000 households from which we want to select 50 units.

$$K = \frac{1000}{50}$$

Calculate

To select the first unit, we randomly pick one number between 1 to 20, say 17. So our sample begins with 17, 37, 57…….. Please note that only the first item was randomly selected. The rest are systematically selected. This is a very popular method because we need only one random number.

Stratified Random Sampling

A probability sampling procedure in which simple random sub-samples are drawn from within different strata that are, more or less equal on some characteristics. Stratified sampling is of two types:

1. ***Proportionate stratified sampling:*** The number of sampling units drawn from each stratum is in proportion to the population size of that stratum.
2. ***Disproportionate stratified sampling:*** The number of sampling units drawn from each stratum is based on the analytical consideration, but not in proportion to the size of the population of that stratum.

Sampling process is as follows:

1. The population to be sampled is divided into groups (stratified).
2. A simple random sample is chosen.

Notes

Notes **Reason for Stratified Sampling**

Sometimes, marketing professionals want information about the component part of the population. Assume there are three stores. Each store forms a strata and the sampling from within each strata is being selected. The resultant might be used to plan different promotional activities for each store strata.

Suppose a researcher wishes to study the retail sales of products, such as tea in a universe of 1,000 grocery stores (Kirana shops included). The researcher can first divide this universe into three strata based on the size of the store. This benchmark for size could be only one of the following (a) floor space (b) volume of sales (c) variety displayed etc.

Size of stores	No. of stores	Percentage of stores
Large stores	2,000	20
Medium stores	3,000	30
Small stores	5,000	50
	10,000	100

Suppose we need 12 stores, then choose four from each strata, at random. If there was no stratification, simple random sampling from the population would be expected to choose two large stores (20% of 12) about four medium stores (30% of 12) and about six small stores (50% of 12).

As can be seen, each store can be studied separately using the stratified sample.

Selection by Proportionate Stratified Sample

Assume that there are 60 students in a class of a management school, of this, 10 has to be selected to take part in a Business quiz competition. Assume that the class has students specializing in marketing, finance and HR stream.

The first step is to subdivide the students of the class into 3 homogeneous groups or stratify the student population, by the area in which they are specializing.

Marketing Streaming			Finance Stream		HR Stream
1	32	8	11	33	34
2	36	12	13	35	37
3	40	15	17	38	39
4	43	18	20	41	42
5	46	19	21	44	45
7	47	22	24	49	48
9	60	23	25	59	58
10	57	28	26	60	56
14	50	27	29	52	51
16	53	31	30	55	54

Second step is to calculate the sampling fraction f = n/N

n = Sample size required

N = Population size

Notes

Third step - Determine how many are to be selected from marketing stream (say n1)

$n_1 = 30 \times 1/10 = 30 \times 1/10$

Sample to be selected from marketing strata $n_1 = 30 \times 1/10 = 3$

Now we can select 3 numbers from among 30 numbers at random say 7, 60, 22

Similarly we can select n_2 n_3

$n_2 = 20 \times 1/10 = 2$

The 2 numbers selected at random from finance stream are 13, 59

$N_3 = 10 \times 1/10 = 1$

Stratified sampling can be carried out with:

1. Same proportion across the strata proportionate stratified sample.
2. Varying proportion across the strata disproportionate stratified sample.

Example:

Size of stores	No. of stores (Population)	Sample Proportionate	Sample Disproportionate
Large	2,000	20	25
Medium	3,000	30	35
Small	5,000	50	40
	10,000	100	100

Estimation of universe mean with a stratified sample

Example:

Size of stores	Sample Mean Sales per store	No. of stores	Percent of stores
Large	200	2000	20
Medium	80	3000	30
Small	40	5000	50
		10,000	100

The population mean of monthly sales is calculated by multiplying the sample mean by its relative weight.

$200 \times 0.2 + 80 \times 0.3 + 40 \times 0.5 = 84$

Sample Proportionate

If N is the size of the population.

n is the size of the sample.

i represents 1, 2, 3,……..k [number of strata in the population]

\ Proportionate sampling

Notes

$$P = \frac{n_1}{N_1} = \frac{n_2}{N_2} = \text{.....} = \frac{n_k}{N_k} = \frac{n}{N}$$

$$= \frac{n_1}{N_1} = \frac{n}{N} = n_1 = \frac{n}{N} \times n_1 \text{ and so on}$$

N_1 is the sample size to be drawn from stratum 1

$n_1 + n_2 + \ldots\ldots\ldots\ldots n_k = n$ [Total sample size of the all strata]

Example: A survey is planned to analyse the perception of people towards their own religious practices. The population consists of various religions, viz., Hindu, Muslim, Christian, Sikh, Jain, assuming a total of 10,000. Hindu, Muslim, Christian, Sikh and Jains consists of 6,000, 2,000, 1,000, 500 and 500 respectively. Determine the sample size of each stratum by applying proportionate stratified sampling, if the sample size required is 200.

Solution:

Total population, N = 10,000

Population in the strata of Hindus N_1 = 6,000

Population in the strata of Muslims N_2 = 2,000

Population in the strata of Christians N_3 = 1,000

Population in the strata of Sikhs N_4 = 500

Population in the strata of Jains N_5 = 500

Proportionate Stratified Sampling

$$P = \frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \frac{n_4}{N_4} = \frac{n_5}{N_5} = \frac{n}{N}$$

Let us determine the sample size of strata N_1

$$\frac{n_1}{N_1} = d$$

$$= 20 \times 6$$

$$= 120$$

$$n_2 = \frac{n}{N} \times n_2 = \frac{2000}{10,000} \times 2,000$$

$$= 40$$

$$n_3 = \frac{n}{N} \times N_3 = \frac{200}{10,000} \times 1,000$$

$$= 20$$

$$n_4 = \frac{n}{N} \times N_4 = \frac{200}{10,000} \times 500$$

$$= 10$$

Notes

$$n_5 = \frac{n}{N} \times N_5 = 10$$

$$n = n_1 + n_2 + n_3 + n_4 + n_5$$

$$= 120 + 40 + 20 + 10 + 10$$

$$= 200.$$

Sample Disproportion

Let is the variance of the stratum i,

where i = 1, 2, 3…….k.

The formula to compute the sample size of the stratum i is the variance of the stratum i,

where size of stratum i

r_i = Sample size of stratum i

$$r_i = \frac{N_i}{N}$$

r = Ratio of the size of the stratum i with that of the population.

N_i = Population of stratum i

N = Total population.

Example: The Government of India wants to study the performance of women self help groups (WSHGs) in three regions viz. North, South and West. The total number of WSHGs is 1,500. The number of groups in North, South and West are 600, 500 and 400 respectively. The Government found more variation between WSHGs in the North, South and West regions. The variance of performance of WSHGs in these regions are 64, 25 and 16 respectively. If the disappropriate stratified sampling is to be used with the sample size of 100, determine the number of sampling units for each region.

Solution:

Total Population	$N = 1,500$
Size of the stratum 1,	$N_1 = 600$
Size of the stratum 2,	$N_2 = 500$
Size of the stratum 3,	$N_3 = 400$
Variance of stratum 1,	$\sigma^2 = 8^2 = 64$
Variance of stratum 2,	$\sigma^2 = 5^2 = 25$
Variance of stratum 3,	$\sigma^2 = 4^2 = 16$
Sample size	$n = 100$

Notes

Stratum Number	Size of the stratum N_i	$r_i = \frac{N_i}{N}$	σ_i	$r_i\sigma_{in}$	$r_i\sigma_{in} = \frac{r_i\sigma_{in}}{\sum_1^3 r_i\sigma_i}$
1	600	0.4	8	3.2	54
2	500	0.33	5	1.65	28
3	400	0.26	4	1.04	18
Total					**100**

Example: Let us consider a case of 3 strata, of income group with given stratum variance.

Stratum	No. of Households	Stratum Variance
0 - 5000	300	4.00
5001-10,000	450	9.00
> 10,000	750	2.25
Total	**1500**	

Find out the nos. From each stratum for a given sample size of 50?

Solution:

Disproportional Stratified Sampling

Stratum No (i)	No. of elements/ Households	Strata Variance	Stratum Standard Deviation	Sample size (m)	Sampling Ratio (n_i/N)
0 - 5000	300	4.00	2.0	10	0.033
5001-10000	450	9.00	3.0	22	0.049
> 10,000	750	2.25	1.5	18	0.024
Total	**1500**			**50**	

$$n_1\sigma_1 + n_2\sigma_2 + n_3\sigma_3 = (300 \times 2.0) + (450 \times 3.0) + (750 \times 1.5)$$

$$= 600 + 1350 + 1125 = 3075$$

$\therefore$ $$n_1 = \frac{50}{3075} \times 600 = 908$$

$$n_2 = \frac{50}{3075} \times 1350 = 22$$

$$n_3 = \frac{50}{3075} \times 1125 = 18$$

Stratified Sampling in Practice: The main reasons for using stratified sampling for managerial applications are:

1. It can obtain information about different parts of the universe, i.e., it allows to draw separate conclusion for each stratum.
2. It often provides universe estimates of greater precision than other methods of random sampling say simple random sampling.

However, the price paid for these advantages is high because of the complexity of design and analysis.

Notes

Cluster Sampling

The following steps are followed:

1. The population is divided into clusters.
2. A simple random sample of few clusters is selected.
3. All the units in the selected cluster are studied.

Step 1: The above mentioned cluster sampling is similar to the first step of stratified random sampling. But the two sampling methods are different. The key to cluster sampling is decided by how homogeneous or heterogeneous the clusters are.

A major advantage of simple cluster sampling is the case of sample selection. Suppose, we have a population of 20,000 units from which we wish to select 500 units. Choosing a sample of that size is a very time-consuming process, if we use Random Numbers table. Suppose, the entire population is divided into 80 clusters of 250 units each, we can choose two sample clusters (2 × 250 = 500) easily by using cluster sampling. The most difficult job is to form clusters. In marketing, the researcher forms clusters so that he can deal with each cluster differently.

Example: Assume there are 20 households in a locality.

Cross	Houses			
1	X_1	X_2	X_3	X_4
2	X_5	X_6	X_7	X_8
3	X_9	X_{10}	X_{11}	X_{12}
4	X_{13}	X_{14}	X_{15}	X_{16}

We need to select eight houses. We can choose eight houses at random. Alternatively, two clusters, each containing four houses can be chosen. In this method, every possible sample of eight houses would have a known probability of being chosen – i.e. chance of one in two. We must remember that in the cluster, each house has the same characteristics. With cluster sampling, it is impossible for certain random sample to be selected. For example, in the cluster sampling process described above, the following combination of houses could not occur: X_1 X_2 X_5 X_6 X_9 X_{10} X_{13} X_{14}. This is because the original universe of 16 houses have been redefined as a universe of four clusters. So only clusters can be chosen as a sample.

Example: Suppose, we want to have 7500 households from all over the country. In such a case, from the first stage, District, say 30 districts out of 600 are selected from all over the country.

I Stage - Cities: Suppose 5 cities are selected out of each 30 districts; and

II Stage - Wards/Localities: say 10 wards/localities are selected from each city

III Stage - Households: 50 households are selected from each ward/locality.

In stage I, we can employ stratified sampling

In stage II, we can use cluster sampling

In stage III, we can have simple random sampling.

Notes

Caution The use of various methods shall give individually contribute towards accuracy, cost, time, etc. This leads us to conclude that multistage sampling leads to saving of time, labour and money. Apart from this wherever an appropriate frame is not available, the use of multistage sampling has universal appeal.

Multistage Sampling

The name implies that sampling is done in several stages. This is used with stratified/cluster designs.

An illustration of double sampling is as follows.

The management of a newly-opened club is solicits new membership. During the first rounds, all corporates were sent details so that those who are interested may enroll. Having enrolled, the second round concentrates on how many are interested to enroll for various entertainment activities that club offers such as billiards, indoor sports, swimming, gym etc. After obtaining this information, you might stratify the interested respondents. This will also tell you the reaction of new members to various activities. This technique is considered to be scientific, since there is no possibility of ignoring the characteristics of the universe.

Task What are the advantages and disadvantages of multistage sampling? Enlist.

Area Sampling

This is a probability sampling, a special form of cluster sampling.

Example: If someone wants to measure the sales of toffee in retail stores, one might choose a city locality and then audit toffee sales in retail outlets in those localities.

The main problem in area sampling is the non-availability of lists of shops selling toffee in a particular area. Therefore, it would be impossible to choose a probability sample from these outlets directly. Thus, the first job is to choose a geographical area and then list out outlets selling toffee. Then follows the probability sample for shops among the list prepared.

Example: You may like to choose shops which sell the brand-Cadbury dairy milk. The disadvantage of the area sampling is that it is expensive and time-consuming.

4.3.2 Non-probability Sampling Techniques

1. Deliberate sampling
2. Shopping mall intercept sampling
3. Sequential sampling
4. Quota sampling
5. Snowball sampling
6. Panel samples

Notes

Deliberate or Purposive Sampling

This is also known as the judgment sampling. The investigator uses his discretion in selecting sample observations from the universe. As a result, there is an element of bias in the selection. From the point of view of the investigator, the sample thus chosen may be a true representative of the universe. However, the units in the universe do not enjoy an equal chance of getting included in the sample. Therefore, it cannot be considered a probability sampling.

Example: Test market cities are being selected, based on the judgment sampling, because these cities are viewed as typical cities matching with certain demographical characteristics. Judgment sample is also frequently used to select stores for the purpose of introducing a new display.

Shopping Mall Intercept Sampling

This is a non-probability sampling method. In this method the respondents are recruited for individual interviews at fixed locations in shopping malls.

Example: Shopper's Shoppe, Food World, Sunday to Monday.

This type of study would include several malls, each serving different socio-economic population.

Example: The researcher may wish to compare the responses of two or more TV commercials for two or more products. Mall samples can be informative for this kind of studies. Mall samples should not be used under following circumstances i.e., if the difference in effectiveness of two commercials varies with the frequency of mall shopping, change in the demographic characteristic of mall shoppers, or any other characteristic. The success of this method depends on "How well the sample is chosen".

Sequential Sampling

This is a method in which the sample is formed on the basis of a series of successive decisions. They aim at answering the research question on the basis of accumulated evidence. Sometimes, a researcher may want to take a modest sample and look at the results. Thereafter, s(he) will decide if more information is required for which larger samples are considered. If the evidence is not conclusive after a small sample, more samples are required. If the position is still inconclusive, still larger samples are taken. At each stage, a decision is made about whether more information should be collected or the evidence is now sufficient to permit a conclusion.

Example: Assume that a product needs to be evaluated.

A small probability sample is taken from among the current user. Suppose it is found that average annual usage is between 200 to 300 units. It is known that the product is economically viable only if the average consumption is 400 units. This information is sufficient to take a decision to drop the product. On the other hand, if the initial sample shows a consumption level of 450 to 600 units, additional samples are needed for further study.

Quota Sampling

Quota sampling is quite frequently used in marketing research. It involves the fixation of certain quotas, which are to be fulfilled by the interviewers.

Notes

Suppose, 2,00,000 students are appearing for a competitive examination. We need to select 1% of them based on quota sampling. The classification of quota may be as follows:

Example: Classification of Samples

Category	Quota
General merit	1,000
Sport	600
NRI	100
SC/ST	300
Total	**2,000**

Quota sampling involves the following steps:

1. The population is divided into segments on the basis of certain characteristics. Here, the segments are termed as cells.
2. A quota of unit is selected from each cell.

Snowball Sampling

This is a non-probability sampling. In this method, the initial group of respondents are selected randomly. Subsequent respondents are being selected based on the opinion or referrals provided by the initial respondents. Further referrals will lead to more referrals, thus leading to a snowball sampling. The referrals will have demographic and psychographic characteristics that are relatively similar to the person referring them.

Example: College students bring in more students on the consumption of Pepsi. The major advantage of snowball sampling is that it monitors the desired characteristics in the population.

Panel Samples

Panel samples are frequently used in marketing research. To give an example, suppose that one is interested in knowing the change in the consumption pattern of households. A sample of households is drawn. These households are contacted to gather information on the pattern of consumption. Subsequently, say after a period of six months, the same households are approached once again and the necessary information on their consumption is collected.

4.3.3 Distinction between Probability Sample and Non-probability Sample

Probability Sample

1. Here, each member of a universe has a known chance of being selected and included in the sample.
2. Any personal bias is avoided. The researcher cannot exercise his discretion in the selection of sample items.

Example: Random sample and cluster sample.

Notes

Non-probability Sample

In this case, the likelihood of choosing a particular universe element is unknown. The sample chosen in this method is based on aspects like convenience, quota etc.

Example: Quota sampling and Judgment sampling.

Difference between Cluster Sampling and Stratified Random Sampling

The major difference between cluster sampling and stratified sampling lies with the inclusion of the cluster or strata. In stratified random sampling, all the strata of the population is sampled while in cluster sampling, the researcher merely randomly selects a number of clusters from the collection of clusters of the entire population. Thus, only a number of clusters are sampled, all the other clusters are left unrepresented.

The other notable differences between Cluster and Stratified random sampling are as follows:

- When natural groupings are clear in a statistical population, cluster sampling technique is used. While Stratified sampling is a method where in, the member of a group are grouped into relatively homogeneous groups.
- Cluster sampling can be chosen if the group consists of homogeneous members. On the other hand, for heterogeneous members in the groups, stratified sampling is a good option.
- The benefit of cluster sampling over other sampling methods is, it is cheaper as compared to the other methods. While the benefits of stratified sampling are, this method ignores the irrelevant ones and focuses on the vital sub populations. Another advantage is, with stratified random sampling method is that for different sub populations, the researcher can opt for different sampling techniques. The stratified sampling method as well helps in improving the efficiency and accuracy of the estimation and facilitates greater balancing of statistical power of tests.
- The major disadvantage of cluster sampling is, it initiates higher sampling error. This sampling error may be represented as design effect. The disadvantages of stratified random sampling method are, it calls for choice of relevant stratification variables which can be tough at times. When there are homogeneous subgroups, random sampling method is not much useful. The implementation of random sampling method is expensive and If not provided with correct information about the population, then an error may be introduced.
- All strata are represented in the sample; but only a subset of clusters are in the sample.

Self Assessment

Fill in the blanks:

7. Sampling is divided into two types, viz. ..... and .....
8. There are ..... methods used in the random sampling.
9. ..... is also called as the random sampling with replacement.
10. ..... is also called random sampling without replacement.
11. Stratified sampling can be carried out with ..... proportion across the strata proportionate stratified sample.

Notes

4.4 Fieldwork

The fieldwork consists of informal conversations as well as formal standardized interviews, including projectives or questionnaires. Initially, a single person conducted the research. Changes in society have shifted research for the most part into teamwork. However, a single person can still conduct effective research. Traditionally, educational researchers began their research with a set of hypothesis, whereas the fieldworker's hypothesis emerges through the fieldwork.

Fieldwork in its inception may seem to be disorganized. The notes may be scattered, information is coming from all over the place. That is because the hypothesis has not yet emerged. Even though, at times the hypothesis may become very clear rapidly. Once the hypothesis became evident the fieldworker maintains an open mind thus allowing other hypothesis to emerge.

Another important difference between the types of research is the "nature of the proposition sought: his propositions are rarely of the A causes B type, the usual casual interrelationships between two or more variables dealt with in an experimental research".

Much of the naturalistic data is collected by using raw materials: notes stating the actual response given. In order to be accurate recorders are often used. Experienced researchers create their own techniques and develop the ability to remember the information that needs to be recorded.

Did u know? **How does a fieldworker know when the Enquiry should finish?**

The fieldworker knows when the inquiry should finish by analyzing the data as it is gathered. The end arrives when the fieldworker sees patterns and no new significant changes.

Three important points that must be included are:

1. The data can be subjective to quantitative analysis
2. Most practitioners of the method probably consider its products to have full status as actual studies
3. Can be credible regardless of abstraction.

Self Assessment

Fill in the blanks:

12. Fieldwork in its inception may seem to be .....
13. ..... researchers create their own techniques and develop the ability to remember the information that needs to be recorded.

4.5 Errors in Sampling

4.5.1 Sampling Error

The only way to guarantee the minimization of sampling error is to choose the appropriate sample size. As the sample keeps on increasing, the sampling error decreases. Sampling error is the gap between the sample mean and population mean.

Example: If a study is done amongst Maruti car-owners in a city to find the average monthly expenditure on the maintenance of car, it can be done by including all Maruti car-owners. It can

Notes

also be done by choosing a sample without covering the entire population. There will be a difference between the two methods with regard to monthly expenditure.

4.5.2 Non-sampling Error

One way of distinguishing between the sampling and the non-sampling error is that, while sampling error relates to random variations which can be found out in the form of standard error, non-sampling error occurs in some systematic way which is difficult to estimate.

4.5.3 Sampling Frame Error

A sampling frame is a specific list of population units, from which the sample for a study being chosen.

 *Example:*

1. An MNC bank wants to pick up a sample among the credit card holders. They can readily get a complete list of credit card holders, which forms their data bank. From this frame, the desired individuals can be chosen. In this example, sample frame is identical to ideal population namely all credit card holders. There is no sampling error in this case.
2. Assume that a bank wants to contact the people belonging to a particular profession over phone (doctors, lawyers) to market a home loan product. The sampling frame in this case is the telephone directory. This sampling frame may pose several problems: (1) People might have migrated. (2) Numbers have changed. (3) Many numbers were not yet listed. The question is "Are the residents who are included in the directory likely to differ from those who are not included"? The answer is yes. Thus in this case, there will be a sampling error.

4.5.4 Non-response Error

This occurs, because the planned sample and final sample vary significantly.

 *Example:* Marketers want to know about the television viewing habits across the country. They choose 500 households and mail the questionnaire. Assume that only 200 respondents reply. This does not show a non-response error, which depends upon the discrepancy. If those 200 who replied did not differ from the chosen 500, there is no non-response error.

Consider an alternative. The people who responded are those who had plenty of leisure time. Therefore, it is implied that non-respondents do not have adequate leisure time. In this case, the final sample and the planned sample differ. If it was assumed that all the 500 chosen have leisure time, but in the final analysis only 200 have leisure time and not others. Therefore, a sample with respect to leisure time leads to response error.

Guidelines to Increase the Response Rate

Every researcher likes to get maximum possible response from the respondents, and will be most delighted if cent percent respondent unfortunately, this does not happen. The non-response error can be reduced by increasing the response rate. Higher the response rate, more accurate and reliable is the data. In order to achieve this, some useful hints could be as follows:

Notes

1. Intimate the respondents in advance through a letter. This will improve the preparedness.
2. Personalized questionnaire should be accompanied by a covering letter.
3. Ensure/Assure that confidentiality will be maintained
4. Questionnaire length is to be restricted
5. Increase of personal interview, I.D. card is essential to prove the bona fide.
6. Monetary incentives are gifts will act as motivator
7. Reminder/Revisits would help.
8. Send self addressed/stamped envelope to return the completed questionnaire.

4.5.5 Data Error

This occurs during the data collection, analysis of data or interpretation. Respondents sometimes give distorted answers unintentionally for questions which are difficult, or if the question is exceptionally long and the respondent may not have answer. Data errors can also occur depending on the physical and social characteristics of the interviewer and the respondent. Things such as the tone and voice can affect the responses. Therefore, we can say that the characteristics of the interviewer can also result in data error. Also, cheating on the part of the interviewer leads to data error. Data errors can also occur when answers to open-ended questions are being improperly recorded.

Failure of the Interviewer to Follow Instructions

The respondent must be briefed before beginning the interview, "What is expected"? "To what extent he should answer"? Also, the interviewer must make sure that respondent is familiar with the subject. If these are not made clear by the interviewer, errors will occur.

Editing mistakes made by the editors in transferring the data from questionnaire to computers are other causes for errors.

The respondent could terminate his/her participation in data gathering, because it may be felt that the questionnaire is too long and tedious.

Self Assessment

Fill in the blanks:

14. The only way to guarantee the minimization of sampling error is to choose the appropriate .....
15. A ..... is a specific list of population units, from which the sample for a study being chosen.
16. The ..... error can be reduced by increasing the response rate.

4.6 Sample Size Decision

1. The first factor that must be considered in estimating sample size, is the error permissible.
2. Greater the desired precision, larger will be the sample size.
3. Higher the confidence level in the estimate, the larger the sample must be. There is a trade off between the degree of confidence and the degree of precision with a sample of fixed size.

Notes

4. The greater the number of sub-groups of interest within the sample, the greater its size must be.
5. Cost is a factor that determines the size of the sample.
6. The issue of response rate: The issue to be considered in deciding the necessary sample size is the actual number of questionnaires that must be sent out. Calculation-wise, we may send questionnaires to the required number of people, but we may not receive the response. For example, we may like to obtain the family income level from a mail survey, but the researcher may not receive response from everyone. If the researcher feels the response rate is 40%, then he needs to despatch that many extra questionnaires. A low percentage of response can cause serious problems to the researcher. This is known as the non-response error.

Non-response error may be due to (1) failure to locate, (2) flat refusal.

Failure to locate: People move to new destinations. However, if the sample frames used are of recent origin, this problem can be overcome.

Flat refusal: We do not know if those who did not respond hold different views or opinions from those who responded.

This implies that those who don't respond should be motivated. It can be done in any one of the following ways:

1. An advance letter informing the respondents that they will receive a questionnaire and requesting their cooperation. This will generally increase the rate of response.
2. Monetary incentive or gift given to respondents will yield a larger response rate.
3. Proper follow up is necessary after the potential respondent received the questionnaire.

Example: Determine the sample size if standard deviation of the population is 3.9, population mean is 36 and sample mean is 33 and the desired degree of precision is 99%.

Solution:

Given, $\sigma = 3.9$, $\mu = 36$, $\bar{x}$ and $z = 1\%$ (99% precision implies 1% level of significance)

i.e. $z_\alpha = 2.576$ (at 1% l.o.s) (Table value)

We know that sample size n can be obtained using the relation

$$n = \left(\frac{z_\alpha \sigma}{d}\right)^2 \qquad \text{where } d = \mu - \bar{X}$$

$$n = \left(\frac{2.576 \times 3.9}{36 - 33}\right)^2 = 11.21 \approx 11$$

Example: Determine the sample size if the standard deviation of population is 12 and the standard error (standard deviation of the sampling distribution) is 3.69.

Solution:

Given the standard deviation of population

$$\sigma = 12$$

Standard error = $\sigma X = 3.69$

Notes

We know that

$$\sigma X = \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow \quad \sigma X_2 = \frac{\sigma^2}{n}$$

$$\Rightarrow \quad n = \frac{\sigma^2}{\sigma X^2} = \left(\frac{12}{3.69}\right)^2$$

$$\Rightarrow \quad n = 10.57 \; 11$$

Example: Determine the sample size, if sample proportion p = 0.4 and standard error of proportion is 0.043.

Solution:

Given that $p = 0.4 \; 1 \Rightarrow q = 0.6 \; \sigma_p = 0.043$

We know that $\sigma_p = \sqrt{\frac{pq}{n}}$

$$\Rightarrow \quad \sigma_p^2 = \frac{pq}{n}$$

$$\Rightarrow \quad n = \frac{pq}{\sigma_p^2} = \frac{0.4 \times 0.6}{(0.043)^2}$$

$$= 129.79 \; 130$$

Example: Determine the sample size if the standard deviation of population is 8.66, sample mean is 45, population mean 43 and the desired degree of precision is 95%.

Solution:

Given that $\mu = 43, \; \bar{X} = 45$

$\sigma = 8.66 \qquad z = 5\%$ l.o.s

$$\Rightarrow \quad z_\alpha = 1.96$$

We know that sample size n can be obtained using the relation

$$n = \left(\frac{z_\alpha \sigma}{d}\right)^2$$

where $d = \mu - \bar{X}$

$$n = \left(\frac{1.96 \times 8.66}{43 - 45}\right)^2 = 72.03 \approx 72$$

Notes

Self Assessment

Fill in the blanks:

17. Greater the desired precision, _____will be the sample size

18. There is a trade off between the degree of confidence and the degree of _____with a sample of fixed size.

4.7 Sampling Distribution

A sampling distribution is the probability distribution of a given statistic based on a random sample of certain size n. It may be considered as the distribution of the statistic for all possible samples of a given size. The sampling distribution depends on the underlying distribution of the population, the statistic being considered, and the sample size used. The sampling distribution is frequently opposed to the asymptotic distribution, which corresponds to the limit case.

Example: Consider a normal population with mean and variance. Assume we repeatedly take samples of a given size from this population and calculate the arithmetic mean for each sample – this statistic is called the sample mean. Each sample will have its own average value, and the distribution of these averages will be called the "sampling distribution of the sample mean". This distribution will be normal $N(m, s^2/n)$ since the underlying population is normal.

The standard deviation of the sampling distribution of the statistic is referred to as the standard error of that quantity. For the case where the statistic is the sample mean, the standard error is:

$$\sigma_z = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population distribution of that quantity and n is the size (number of items) in the sample.

Self Assessment

Fill in the blanks:

19. Sampling distribution depends on the underlying distribution of the ....., the statistic being considered, and the sample size used.

20. The standard deviation of the sampling distribution of the statistic is referred to as the ..... of that quantity

4.8 Summary

- Sample is a representative of population while Census represents cent percent of population.
- The most important factors distinguishing whether to choose sample or census is cost and time. There are seven steps involved in selecting the sample.
- There are two types of sample, namely, Probability sampling and Non-probability sample.
- Probability sampling includes random sampling, stratified random sampling systematic sampling, cluster sampling, Multistage sampling.

Notes

- Random sampling can be chosen by Lottery method or using random number table.
- Samples can be chosen either with equal probability or varying probability.
- Random sampling can be systematic or stratified.
- In systematic random sampling, only the first number is randomly selected. Then by adding a constant "K" remaining numbers are generated.
- In stratified sampling, random samples are drawn from several strata, which has more or less same characteristics.
- In multistage sampling, sampling is drawn in several stages.

4.9 Keywords

Census: It refers to complete inclusion of all elements in the population. A sample is a sub-group of the population.

Deliberate Sampling: The investigator uses his discretion in selecting sample observations from the universe. As a result, there is an element of bias in the selection.

Multistage Sampling: The name implies that sampling is done in several stages

Quota Sampling: Quota sampling is quite frequently used in marketing research. It involves the fixation of certain quotas, which are to be fulfilled by the interviewers.

Random Sampling: Simple random sample is a process in which every item of the population has an equal probability of being chosen.

Sample Frame: Sampling frame is the list of elements from which the sample is actually drawn.

Stratified Random Sampling: A probability sampling procedure in which simple random sub-samples are drawn from within different strata, that are, more or less equal on some characteristics.

4.10 Review Questions

1. What do you analyse as the advantages and disadvantages of probability sampling?
2. Which method of sampling would you use in studies, where the level of accuracy can vary from the prescribed norms and why?
3. Shopping Mall Intercept Sampling is not considered a scientific approach. Why?
4. Quota sampling does not require prior knowledge about the cell to which each population unit belongs. Does this attribute serve as an advantage or disadvantage for Quota Sampling?
5. What suggestions would you give to reduce non sampling error?
6. One mobile phone user is asked to recruit another mobile phone user. What sampling method is this known as and why?
7. Sampling is a part of the population. True/False? Why/why not?
8. Determine the sample size if the standard deviation of population is 20 and the standard error is 4.1.
9. What do see as the reason behind purposive sampling being known as judgement sampling?

Notes

10. Suppose, the population consists of 45,000 households, divided into five (5) strata on the basis of monthly income. This can be illustrating as below:

0	-	1000
1001	-	5000
5001	-	7500
7501	-	10,000
Above 10,000		

Then

(a) Find out the number of units from each strata if the sample constitutes 1% of the population.

(b) If selection is for 150 items selecting equally from each strata, find out the number of sample units from each strata.

Answers: Self Assessment

1.	target	2.	frame
3.	large, homogeneous	4.	small
5.	target	6.	resources
7.	probability, non-probability	8.	two
9.	Equal Probability	10.	Varying Probability
11.	same	12.	disorganized
13.	Experienced	14.	sample size
15.	sampling frame	16.	non response
17.	larger	18.	precision
19.	population	20.	standard error

4.11 Further Readings

Books

Cooper and Schinder, *Business Research Methods*, TMH.

CR Kotari, *Research Methodology*, Vishwa Prakashan.

David Luck and Ronald Rubin, *Marketing Research*, PHI.

Naresh Amphora, *Marketing Research*, Pearson Education.

S.N. Murthy & U. Bhojanna, *Business Research Methods*, 3rd Edition, Excel Books.

William Zikmund, *Business Research Methods*, Thomson.

Unit 5: Measurement and Scaling Techniques

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Recognize the tools of sound measurement
- Explain the techniques of developing measurement tools
- Describe the meaning and techniques of scaling
- Differentiate among Comparative and non-comparative scales
- Describe the Multi-dimensional scaling techniques

Introduction

Measurement is assigning numbers or other symbols to characteristics of objects being measured, according to predetermined rules. Concept (or Construct) is a generalized idea about a class of objects, attributes, occurrences, or processes.

Relatively concrete constructs comprises of aspects such as Age, gender, number of children, education, income. Relatively abstract constructs take into accounts the aspects such as Brand loyalty, personality, channel power, satisfaction.

Notes

Scaling is the generation of a continuum upon which measured objects are located.

Scale is a quantifying measure – a combination of items that is progressively arranged according to value or magnitude. The purpose is to quantitatively represent an item's, person's, or event's place in the scaling continuum.

5.1 Measurement Scales: Tools of Sound Measurement

These are of four kinds of scales, namely:

1. Nominal scale
2. Ordinal scale
3. Interval scale
4. Ratio scale

5.1.1 Nominal Scale

In this scale, numbers are used to identify the objects. For example, University Registration numbers assigned to students, numbers on their jerseys.

The purpose of marking numbers, symbols, labels etc. in this type of scaling is not to establish an order but it is to simply put labels in order to identify events and count the objects and subjects. This measurement scale is used to classify individuals, companies, products, brands or other entities into categories where no order is implied. Indeed, it is often referred to as a categorical scale. It is a system of classification and does not place the entity along a continuum. It involves a simple count of the frequency of the cases assigned to the various categories, and if desired numbers can be nominally assigned to label each category.

Characteristics

1. It has no arithmetic origin.
2. It shows no order or distance relationship.
3. It distinguishes things by putting them into various groups.

Use: This scale is generally used in conducting in surveys and ex-post-facto research.

Example: Have you ever visited Bangalore?

Yes-1

No-2

'Yes' is coded as 'One' and 'No' is coded as 'Two'. The numeric attached to the answers has no meaning, and is a mere identification. If numbers are interchanged as one for 'No' and two for 'Yes', it won't affect the answers given by respondents. The numbers used in nominal scales serve only the purpose of counting.

The telephone numbers are an example of nominal scale, where one number is assigned to one subscriber. The idea of using nominal scale is to make sure that no two persons or objects receive the same number. Similarly, bus route numbers are the example of nominal scale.

Notes

"How old are you"? This is an example of a nominal scale.

"What is your PAN Card number?

Arranging the books in the library, subject wise, author wise - we use nominal scale.

Caution It should be kept in mind that nominal scale has certain limitation, viz.

1. There is no rank ordering.
2. No mathematical operation is possible.
3. Statistical implication - Calculation of the standard deviation and the mean is not possible. It is possible to express the mode.

5.1.2 Ordinal Scale (Ranking Scale)

The ordinal scale is used for ranking in most market research studies. Ordinal scales are used to ascertain the consumer perceptions, preferences, etc. For example, the respondents may be given a list of brands which may be suitable and were asked to rank on the basis of ordinal scale:

1. Lux
2. Liril
3. Cinthol
4. Lifebuoy
5. Hamam

Rank	Item	Number of respondents
I	Cinthol	150
II	Liril	300
III	Hamam	250
IV	Lux	200
V	Lifebuoy	100
Total		**1,000**

In the above example, II is mode and III is median.

Statistical implications: It is possible to calculate the mode and the median.

In market research, we often ask the respondents to rank the items, like for example, "A soft drink, based upon flavour or colour". In such a case, the ordinal scale is used. Ordinal scale is a ranking scale.

Rank the following attributes of 1-5 scale according to the importance in the microwave oven:

Attributes	Rank
A) Company Image	5
B) Functions	3
C) Price	2
D) Comfort	1
E) Design	4

Ordinal scale is used to arrange things in order. In qualitative researches, rank ordering is used to rank characteristics units from the highest to the lowest.

Notes

Characteristics

1. The ordinal scale ranks the things from the highest to the lowest.
2. Such scales are not expressed in absolute terms.
3. The difference between adjacent ranks is not equal always.
4. For measuring central tendency, median is used.
5. For measuring dispersion, percentile or quartile is used.

Scales involve the ranking of individuals, attitudes or items along the continuum of the characteristics being scaled.

From the information provided by ordinal scale, the researcher knows the order of preference but nothing about how much more one brand is preferred to another i.e., there is no information about the interval between any two brands. All of the information, a nominal scale would have given, is available from an ordinal scale. In addition, positional statistics such as the median, quartile and percentile can be determined. It is possible to test for order correlation with ranked data. The two main methods are Spearman's Ranked Correlation Coefficient and Kendall's Coefficient of Concordance which shall be discussed later in the unit.

Did u know? **What is the difference between nominal and ordinal scales?**

In nominal scale numbers can be interchanged, because it serves only for the purpose of counting. Numbers in Ordinal scale have meaning and it won't allow interchangeability.

1. Students may be categorized according to their grades of A, B, C, D, E, F etc. where A is better than B and so on. The classification is from the highest grade to the lowest grade.
2. Teachers are ranked in the University as professor, associate professors, assistant professors and lecturers, etc.
3. Professionals in good organizations are designated as GM, DGM, AGM, SR.MGR, MGR, Dy. MGR., Asstt. Mgr. and so on.
4. Ranking of two or more households according to their annual income or expenditure, e.g.

Households	A	B	C	D	E
Annual Income (Rs)	5,000	9,000	7,000	13,000	21,000

If highest income is given #1, than we write as

Households	**Order of Households on the Basis of Annual Income**
A	E (1)
B	D(2)
C	B(3)
D	C(4)
E	A(5)

One can ask respondents questions on the basis of one or more attributes such as flower, colour, etc., and ask about liking or disliking, e.g., whether the respondent likes soft drinks or not.

Notes

I strongly like it	+2
I like it	+1
I am indifferent	0
I dislike it	-1
I strongly dislike it	-2

In this manner, ranking can be obtained by asking the respondent their level of acceptability. One can then combine the individual ranking and get a collective ranking of the group.

Interval scale uses the principle of "equality of interval" i.e., the intervals are used as the basis for making the units equal assuming that intervals are equal.

It is only with an interval scaled data that researchers can justify the use of the arithmetic mean as the measure of average. The interval or cardinal scale has equal units of measurement thus, making it possible to interpret not only the order of scale scores but also the distance between them. However, it must be recognized that the zero point on an interval scale is arbitrary and is not a true zero. This, of course, has implications for the type of data manipulation and analysis we can carry out on data collected in this form. It is possible to add or subtract a constant to all of the scale values without affecting the form of the scale but one cannot multiply or divide the values. It can be said that two respondents with scale positions 1 and 2 are as far apart as two respondents with scale positions 4 and 5, but not that a person with score 10 feels twice as strongly as one with score 5. Temperature is interval scaled, being measured either in Centigrade or Fahrenheit. We cannot speak of 50°F being twice as hot as 25°F since the corresponding temperatures on the centigrade scale, 100°C and -3.9°C, are not in the ratio 2:1.

Interval scales may be either numeric or semantic.

Characteristics

1. Interval scales have no absolute zero. It is set arbitrarily.
2. For measuring central tendency, mean is used.
3. For measuring dispersion, standard deviation is used.
4. For test of significance, t-test and f-test are used.
5. Scale is based on the equality of intervals.

Use: Most of the common statistical methods of analysis require only interval scales in order that they might be used. These are not recounted here because they are so common and can be found in virtually all basic texts on statistics.

5.1.3 Interval Scale

Interval scale is more powerful than the nominal and ordinal scales. The distance given on the scale represents equal distance on the property being measured. Interval scale may tell us "How far the objects are apart with respect to an attribute?" This means that the difference can be compared. The difference between "1" and "2" is equal to the difference between "2" and "3".

Interval scale uses the principle of "equality of interval" i.e., the intervals are used as the basis for making the units equal assuming that intervals are equal.

It is only with an interval scaled data that researchers can justify the use of the arithmetic mean as the measure of average. The interval or cardinal scale has equal units of measurement thus, making it possible to interpret not only the order of scale scores but also the distance between

Notes

them. However, it must be recognized that the zero point on an interval scale is arbitrary and is not a true zero. This, of course, has implications for the type of data manipulation and analysis we can carry out on data collected in this form. It is possible to add or subtract a constant to all of the scale values without affecting the form of the scale but one cannot multiply or divide the values. It can be said that two respondents with scale positions 1 and 2 are as far apart as two respondents with scale positions 4 and 5, but not that a person with score 10 feels twice as strongly as one with score 5. Temperature is interval scaled, being measured either in Centigrade or Fahrenheit. We cannot speak of 50°F being twice as hot as 25°F since the corresponding temperatures on the centigrade scale, 100°C and -3.9°C, are not in the ratio 2:1.

Interval scales may be either numeric or semantic.

Characteristics

1. Interval scales have no absolute zero. It is set arbitrarily.
2. For measuring central tendency, mean is used.
3. For measuring dispersion, standard deviation is used.
4. For test of significance, t-test and f-test are used.
5. Scale is based on the equality of intervals.

Use: Most of the common statistical methods of analysis require only interval scales in order that they might be used. These are not recounted here because they are so common and can be found in virtually all basic texts on statistics.

Example:

1. Suppose we want to measure the rating of a refrigerator using interval scale. It will appear as follows:

   (a) Brand name Poor …… Good

   (b) Price High …….. Low

   (c) Service after-sales Poor …… Good

   (d) Utility Poor ……Good

   The researcher cannot conclude that the respondent who gives a rating of 6 is 3 times more favourable towards a product under study than another respondent who awards the rating of 2.

2. How many hours you spend to do class assignment every day?

   (a) < 30 min.

   (b) 30 min. to 1 hr.

   (c) 1 hr. to 1½ hrs.

   (d) > 1½ hrs.

   Statistical implications: We can compute the range, mean, median, etc.

Task Analyse the difference between interval and ordinal scales.

Notes

5.1.4 Ratio Scale

Ratio scale is a special kind of internal scale that has a meaningful zero point. With this scale, length, weight or distance can be measured. In this scale, it is possible to say, how many times greater or smaller one object is being compared to the other.

These scales are used to measure actual variables. The highest level of measurement is a ratio scale. This has the properties of an interval scale together with a fixed origin or zero point. Examples of variables which are ratio scaled include weights, lengths and times. Ratio scales permit the researcher to compare both differences in scores and in the relative magnitude of scores. For instance, the difference between 5 and 10 minutes is the same as that between 10 and 15 minutes, and 10 minutes is twice as long as 5 minutes.

Given that sociological and management research seldom aspires beyond the interval level of measurement, it is not proposed that particular attention be given to this level of analysis. Suffice it, to say that virtually all statistical operations can be performed on ratio scales.

Characteristics

1. This scale has an absolute zero measurement.
2. For measuring central tendency, geometric and harmonic means are used.

Use: Ratio scale can be used in all statistical techniques.

Example: Sales this year for product A are twice the sales of the same product last year.

Statistical implications: All statistical operations can be performed on this scale.

Self Assessment

Fill in the blanks:

1. ..... scale may tell us "How far the objects are apart with respect to an attribute?"
2. Ratio scale is a special kind of internal scale that has a meaningful .....

5.2 Techniques of Developing Measurement Tools

The scale construction techniques are used for measuring the attitude of a group or an individual. In other words, scale construction technique helps in estimate the interest or behaviour of an individual or a group towards others or another's environment rather than oneself. While performing a scale construction technique, you need to consider various decisions related to the attitude of the individual or group. A few of these decisions are:

- Determining the level of the involved data; identifying whether it is nominal, ordinal, interval or ratio.
- Identifying the useful statistical analysis for the scale construction.
- Identifying the scale construction technique to be used.
- Selecting the physical layout of the scales.
- Determining the scale categories that need to be used.

There are two primary scale construction techniques, comparative and non-comparative. The comparative technique is used to determine the scale values of multiple items by performing

Notes

comparisons among the items. In the non-comparative technique, scale value of an item is determined without comparing with another item. Furthermore, these two techniques are also of many types. The various types of comparative techniques are:

1. ***Pairwise comparison scale:*** This is an ordinal level scale construction technique, where a respondent is provided with two items and then asked him to select his/her choice.
2. ***Rasch model scale:*** In this technique, multiple respondents are simultaneously involved with several items and from their responses comparisons are derived to determine the scale values. Rank-order scale: This is also an ordinal level scale constructing technique, where a respondent is provided with multiple items, which he needs to rank accordingly.
3. ***Constant sum scale:*** In this scale construction technique, a respondent is usually provided with a constant amount of money, credits or points that he needs to allocate to various items for determining the scale values of the items.

The various types of non-comparative techniques are:

1. ***Continuous rating scale:*** In this technique, respondents generally use a series of numbers known as scale points for rating an item. This technique is also known as graphic rating scaling.
2. ***Likert scale:*** This technique allows the respondents to rate the items on a scale of five to seven points depending upon the amount of their agreement or disagreement on the item.
3. ***Semantic differential scale:*** In this technique, respondents are asked to rate the different attributes of an item on a seven-point scale.

Self Assessment

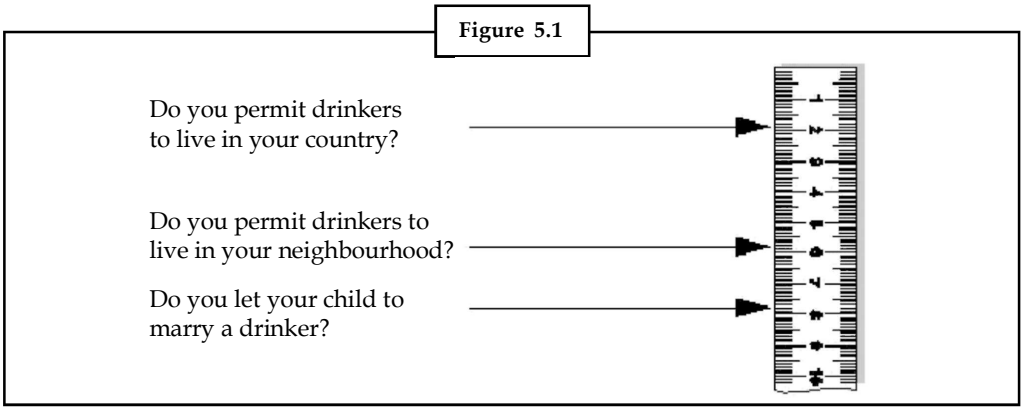
Fill in the blanks:

3. Scale construction techniques are used for measuring the …….of a group.
4. The comparative technique is used to determine the scale values of …….items by performing comparisons among the items.

5.3 Scaling – Meaning

Scaling is a process or set of procedures, which is used to assess the attitude of an individual. Scaling is defined as the assignment of objects to numbers according to a rule. The objects in the definition are text statements, which can be the statements of attitude or principle. Attitude of an individual is not measured directly by scaling. It is first migrated to statements and then the numbers are assigned to them. Figure below shows the how to scale the attitude of individuals.

Figure 5.1



Notes

In the above figure, we are going to assess the attitude of an individual by analysing his thoughts about drinkers. You can see that as you move down, the attitude or behaviour of people towards drinkers become more provisional. If an individual agrees with a statement in the list, then it is more likely that he will also agree with all of the assertions above that statement. Thus in this example, the rule is growing one. So this is called scaling. Scaling is done in the research process to test the hypothesis. Sometimes, you can also use scaling as the part of probing research.

Self Assessment

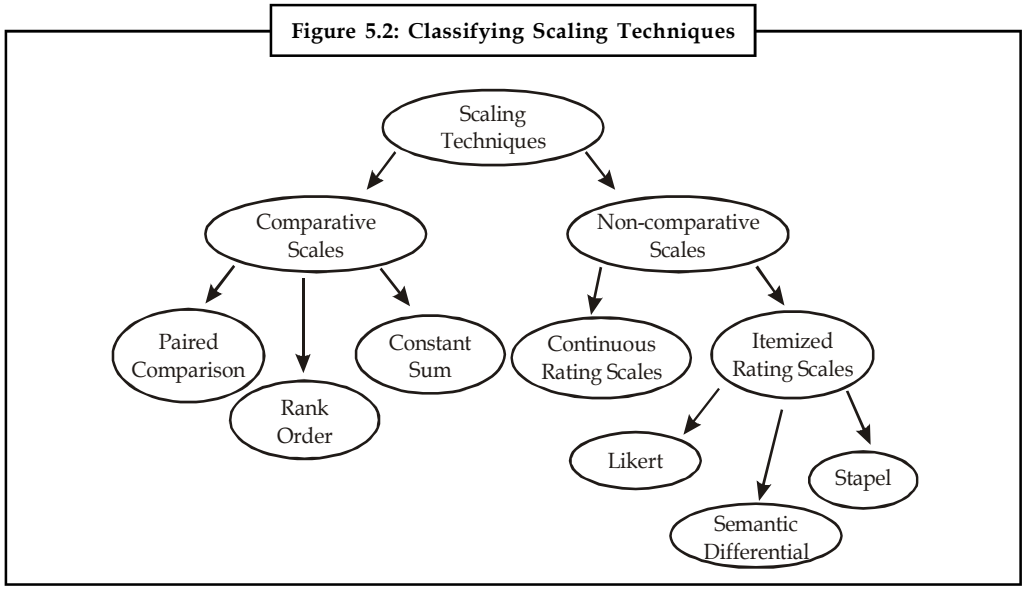
Fill in the blanks:

5. ……of an individual is not measured directly by scaling.
6. Scaling is done in the research process to test the…….

5.4 Comparative and Non-comparative Scaling Techniques

1. ***Comparative Scales:*** It involve the direct comparison of two or more objects.
2. ***Non-comparative Scales:*** Objects or stimuli are scaled independently of each other.

Figure 5.2: Classifying Scaling Techniques



5.4.1 Comparative Scaling Techniques

Paired Comparison

Example: Here a respondent is asked to show his preferences from among five brands of coffee – A, B, C, D and E with respect to flavours. He is required to indicate his preference in pairs. A number of pairs are calculated as follows. The brands to be rated are presented two at a time, so each brand in the category is compared once to every other brand. In each pair, the respondents were asked to divide 100 points on the basis of how much they liked one compared to the other. The score is totally for each brand.

No. of pairs = $\frac{N(N-1)}{2}$

Notes

In this case, it is $\frac{5(5-1)}{2}$

A&B	B&D
A&C	B&E
A&D	C&D
A&E	C&E
B&C	D&E

If there are 15 brands to be evaluated, then we have 105 paired comparison(s) and that is the limitation of this method.

Example: For each pair of professors, please indicate the professor from whom you prefer to take classes with a 1.

	Cunningham	**Day**	**Parker**	**Thomas**
Cunningham		0	0	0
Day	1		1	0
Parker	1	0		0
Thomas	1	1	1	0
# of times Preferred	3	1	2	0

Rank Order Scaling

1. Respondents are presented with several objects simultaneously
2. Then asked to order or rank them according to some criterion
3. Data obtained are ordinal in nature-Arranged or ranked in order of magnitude
4. Commonly used to measure preferences among brands and brand attributes

Example: Please rank the instructors listed below in order of preference. For the instructor you prefer the most, assign a "1", assign a "2" to the instructor you prefer the 2nd most, assign a "3" to the instructor that you prefer 3rd most, and assign a "4" to the instructor that you prefer the least.

Instructor	**Ranking**
Cunningham	1
Day	3
Parker	2
Thomas	4

Constant Sum Scaling

1. Respondents are asked to allocate a constant sum of units among a set of stimulus objects with respect to some criterion
2. Units allocated represent the importance attached to the objects

Notes

3. Data obtained are interval in nature
4. Allows for fine discrimination among alternatives

Example: Listed below are 4 marketing professors, as well as 3 aspects that students typically find important. For each aspect, please assign a number that reflects how well you believe each instructor performs on the aspect. Higher numbers represent higher scores. The total of all the instructors' scores on an aspect should equal 100.

Instructor	Availability	Fairness	Easy Tests
Cunningham	30	35	25
Day	30	25	25
Parker	25	25	25
Thomas	15	15	25
Sum Total	100	100	100

5.4.2 Non-comparative Scale

Continuous Rating Scale

VERY POOR …….....VERY GOOD

0 10 20 30 40 50 60 70 80 90 100

Likert Scale

It is known as summated rating scale. This consists of a series of statements concerning an attitude object. Each statement has '5 points', Agree and Disagree on the scale. They are also called summated scales, because scores of individual items are summated to produce a total score for the respondent. The Likert Scale consists of two parts-item part and evaluation part. Item part is usually a statement about a certain product, event or attitude. Evaluation part is a list of responses like "strongly agree" to "strongly disagree". The five point-scale is used here. The numbers like +2, +1, 0, –1, –2 are used. Now, let us see with an example how the attitude of a customer is measured with respect to a shopping mall.

Table 5.1: Evaluation of Globus-the Super Market by Respondent

S.No.	Likert scale items	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1.	Salesmen at the shopping mall are courteous	-	-	-	-	-
2.	Shopping mall does not have enough parking space	-	-	-	-	-
3.	Prices of items are reasonable.	-	-	-	-	-
4.	Mall has wide range of products to choose	-	-	-	-	-
5.	Mall operating hours are inconvenient	-	-	-	-	-
6.	The arrangement of items in the mall is confusing	-	-	-	-	-

Notes

The respondents' overall attitude is measured by summing up his (her) numerical rating on the statement making up the scale. Since some statements are favourable and others unfavourable, it is the one important task to be done before summing up the ratings. In other words, "strongly agree" category attached to favourable statement and "strongly disagree" category attached to unfavourable. The statement must always be assigned the same number, such as +2, or –2. The success of the Likert Scale depends on "How well the statements are generated?" The higher the respondent's score, the more favourable is the attitude. For example, if there are two shopping malls, ABC and XYZ and if the scores using the Likert Scale are 30 and 60 respectively, we can conclude that the customers' attitude towards XYZ is more favourable than ABC.

Caution The Likert Scale must contain an equal number of favourable and unfavourable statements.

Semantic Differential Scale

This is very similar to the Likert Scale. It also consists of a number of items to be rated by the respondents. The essential difference between Likert and Semantic Differential Scale is as follows:

It uses "Bipolar" adjectives and phrases. There are no statements in the Semantic Differential Scale.

Each pair of adjective is separated by a seven point scale.

Notes Some individuals have favourable descriptions on the right side, while some have on the left side. The reason for the reversal is to have a combination of both favourable and unfavourable statements.

Semantic Differential Scale Items

Please rate the five real estate developers mentioned below on the given scales for each of the five aspects. Developers are

S. No.	Scale items	–3	–2	–1	0	+1	+2	+3	-
1.	Not reliable	_	_	_	_	_	_	_	Reliable
2.	Expensive	_	_	_	_	_	_	_	Not expensive
3.	Trustworthy	_	_	_	_	_	_	_	Not trustworthy
4.	Untimely delivery	_	_	_	_	_	_	_	Timely delivery
5.	Strong Brand Image	_	_	_	_	_	_	_	Poor brand image

The respondents were asked to tick one of the seven categories which describes their views on attitude. Computation is being done exactly the same way as in the Likert Scale. Suppose, we are trying to evaluate the packaging of a particular product. The seven point scale will be as follows:

"I feel ……..

1. Delighted
2. Pleased
3. Mostly satisfied

Notes

4. Equally satisfied and dissatisfied
5. Mostly dissatisfied
6. Unhappy
7. Terrible.

Thurstone Scale

This is also known as an equal appearing interval scale. The following are the steps to construct a Thurstone Scale:

Step 1: To generate a large number of statements, relating to the attitude to be measured.

Step 2: These statements (75 to 100) are given to a group of judges, say 20 to 30, who were asked to classify them according to the degree of favourableness and unfavourableness.

Step 3: 11 piles are to be made by the judges. The piles vary from "most unfavourable" in pile 1 to neutral in pile 6 and most favourable statement in pile 11.

Step 4: Study the frequency distribution of ratings for each statement and eliminate those statements, which different judges have given widely scattered ratings.

Step 5: Select one or two statements from each of the 11 piles for the final scale. List the selected statements in random order to form the scale.

Step 6: The respondents whose attitudes are to be scaled were given the list of statements and asked to indicate their agreement or disagreement with each statement. Some may agree with one statement while some may agree with more than one statement.

Example:

1. Crime and violence in movies:
   (a) All movies with crime and violence should be prohibited by law.
   (b) Watching crime and violence in movies is a waste of time.
   (c) Most movies with crime are bad and harmful.
   (d) The direction and theme in most crime movies are monotonous.
   (e) Watching a movie with crime and violence does not interfere with my routine life.
   (f) I have no opinion one way or the other, about watching movies with crime and violence.
   (g) I like to watch movies with crime and violence.
   (h) Most movies with crime and violence are interesting and absorbing.
   (i) Crime movies act as a knowledge bank gained by the audience.
   (j) People learn "how to be safe and protect oneself" by seeing a movie on crime.
   (k) Watching crime in a movie does not harm our life-style.

   Conclusion: A respondent might agree with statements 8, 9 and 10. Such agreement represents a favourable attitude towards crime and violence. On the contrary, if items 1, 3, 4 are chosen by respondents, it shows that respondents are unfavourably disposed towards crime in movies. If the respondent chooses 1, 5 and 11, it could be interpreted to indicate that s(he) is not consistent in his(her) attitude about the subject.

Notes

2. Suppose, we are interested in the attitude of certain socio-economic class of respondents towards savings and investments. The final list of statements would be as follows:

   (a) One should live for the present and not the future. So, savings are absolutely not required.

   (b) There are many attractions to spend the money saved.

   (c) It is better to spend savings than risk them in investments.

   (d) Investments are unsafe as the money is also blocked.

   (e) You earn to spend and not to invest.

   (f) It is not possible to save these days.

   (g) A certain amount of income should be saved and invested.

   (h) The future is uncertain and investments will protect us.

   (i) Some amount of savings and investments are a must for every individual.

   (j) One should try to save more so that most of it could be invested.

   (k) All savings should be invested for the future.

   Conclusion: A respondent agreeing to statements 8, 9 and 11 would be considered having a favourable attitude towards savings and investments. The person agreeing with statements 2, 3 and 4 is an individual with an unfavourable attitude. Also, if a respondent chooses statements 1, 3, 7 or 9, his attitude is not considered consistent.

Multidimensional Scaling

This is used to study consumer attitudes, particularly with respect to perceptions and preferences. These techniques help identify the product attributes that are important to the customers and to measure their relative importance. Multi-Dimensional Scaling is useful in studying the following:

1. (a) What are the major attributes considered while choosing a product (soft drinks, modes of transportation)? (b) Which attributes do customers compare to evaluate different brands of the product? Is it price, quality, availability etc.?
2. Which is the ideal combination of attributes according to the customer? (i.e., which two or more attributes consumer will consider before deciding to buy.)
3. Which advertising messages are compatible with the consumer's brand perceptions?

Notes The multidimensional scaling is used to describe similarity and preference of brands. The respondents were asked to indicate their perception, or the similarity between various objects (products, brands, etc.) and preference among objects. This scaling is also known as perceptual mapping.

There are two ways of collecting the input data to plot perceptual mapping:

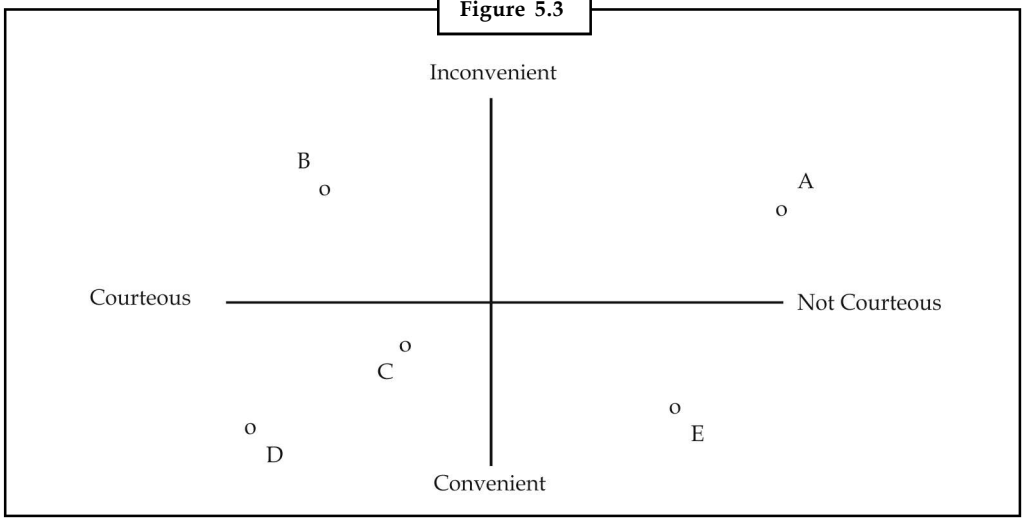
1. ***Non-attribute method:*** Here, the researcher asks the respondent to make a judgment about the objects directly. In this method, the criteria for comparing the objects is decided by the respondent himself.

Notes

2. ***Attribute method:*** In this method, instead of respondents selecting the criteria, they were asked to compare the objects based on the criteria specified by the researcher.

For example, to determine the perception of a consumer: Assume there are five insurance companies to be evaluated on two attributes namely (1) convenient locality (2) courteous personal service. Customers' perception regarding the five insurance companies are as follows:

Figure 5.3



A, B, C, D and E are five insurance companies.

1. According to the map, B & E are dissimilar insurance companies.
2. C is being located very conveniently.
3. A is a less convenient in location compared to E.
4. D is a less convenient in location than C.
5. E is a less convenient location compared to D.

Did u know? **What tools are used in MDS?**

Software such as SPSS, SAS and Excel are the packages used in MDS. Brand positioning research is one of SPSS's important features. SAS is a business intelligence software. Excel is also used to a certain extent.

Stapel Scales

1. Modern versions of the Stapel scale place a single adjective as a substitute for the semantic differential when it is difficult to create pairs of bipolar adjectives.
2. The advantage and disadvantages of a Stapel scale, as well as the results, are very similar to those for a semantic differential.

However, the stapel scale tends to be easier to conduct and administer.

Notes

Table 5.2: Basic Non-comparative Scales

Scale	Basic Characteristics	Examples	Advantages	Disadvantages
Continuous Rating Scale	Place a mark on a continuous line	Reaction to TV commercials	Easy to construct	Cumbersome scoring unless computerized
Itemized Rating Scales				
Likert Scale	Degree of agreement on a numbered scale	Measurement of attitudes, perceptions	Easy to construct, administer, & understand	More time consuming
Semantic Differential	Numbered scale with bipolar labels	Brand, product, & company images	Versatile	Difficult to construct appropriate bipolar adjectives
Stapel Scale	Unipolar numbered scale, no neutral point	Measurement of attitudes & images	Easy to construct, can administer over telephone	Confusing difficult to apply

Self Assessment

Fill in the blanks:

7. The advantage and disadvantages of a Stapel scale, as well as the results, are very similar to those for a ..... differential.
8. ..... Scaling is used to study consumer attitudes, particularly with respect to perceptions and preferences
9. Thurstone Scale is also known as an ..... scale.
10. Semantic Differential Scale is very similar to the ..... Scale.
11. The Likert Scale consists of two parts - ..... and .....
12. In ..... Scaling respondents are presented with several objects simultaneously.
13. Comparative Scales involve the direct comparison of .....objects.

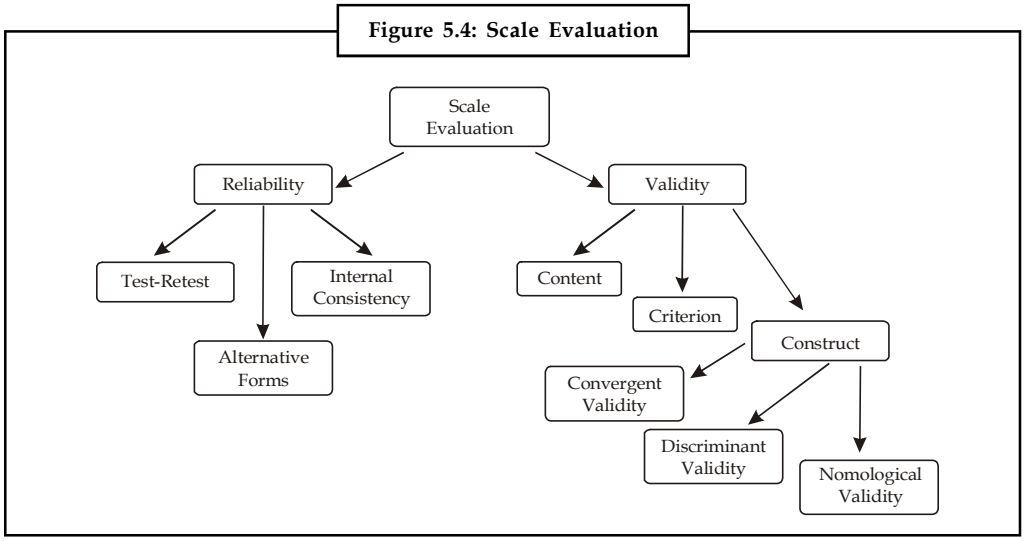
5.5 Criteria for the Good Test

There are two criteria to decide whether the scale selected is good or not. They are:

1. Reliability; and
2. Validity

Notes

Figure 5.4: Scale Evaluation



5.5.1 Reliability Analysis

Reliability means the extent to which the measurement process is free from errors. Reliability deals with accuracy and consistency. The scale is said to be reliable, if it yields the same results when repeated measurements are made under constant conditions.

Example: Attitude towards a product or brand preference.

Reliability can be ensured by using the same scale on the same set of respondents, using the same method. However, in actual practice, this becomes difficult as:

1. Extent to which a scale produces consistent results
2. Test-retest Reliability: Respondents are administered scales at 2 different times under nearly equivalent conditions
3. Alternative-form Reliability: 2 equivalent forms of a scale are constructed, then tested with the same respondents at 2 different times
4. Internal Consistency Reliability:
   (a) The consistency with which each item represents the construct of interest
   (b) Used to assess the reliability of a summated scale
   (c) Split-half Reliability
5. Items constituting the scale divided into 2 halves, and resulting half scores are correlated: Coefficient alpha (most common test of reliability)
6. Average of all possible split-half coefficients resulting from different splitting of the scale items.

Notes

5.5.2 Validity Analysis

The paradigm of validity focused in the question "Are we measuring, what we think, we are measuring?" Success of the scale lies in measuring "What is intended to be measured?" Of the two attributes of scaling, validity is the most important.

There are several methods to check the validity of the scale used for measurement:

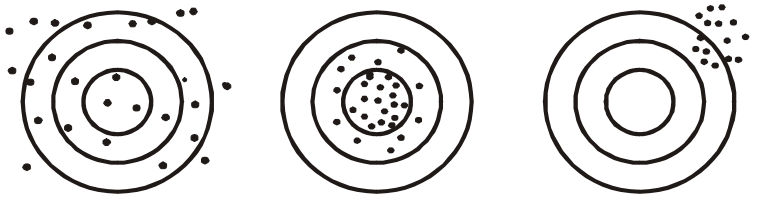
1. ***Construct Validity:*** A sales manager believes that there is a clear relation between job satisfaction for a person and the degree to which a person is an extrovert and the work performance of his sales force. Therefore, those who enjoy high job satisfaction, and have extrovert personalities should exhibit high performance. If they do not, then we can question the construct validity of the measure.
2. ***Content Validity:*** A researcher should define the problem clearly. Identify the item to be measured. Evolve a suitable scale for this purpose. Despite these, the scale may be criticised for being lacking in content validity. Content validity is known as face validity. An example can be the introduction of new packaged food. When new packaged food is introduced, the product representing a major change in taste. Thousands of consumers may be asked to taste the new packaged food. Overwhelmingly, people may say that they liked the new flavour. With such a favourable reaction, the product when introduced on a commercial scale may still meet with failure. So, what is wrong? Perhaps a crucial question that was omitted. The people may be asked if liked the new packaged food, to which the majority might have "yes" but the same respondents were not asked, "Are you willing to give up the product which you are consuming currently?" In this case, the problem was not clearly identified and the item to be 'measured' was left out.
3. ***Predictive Validity:*** This pertains to "How best a researcher can guess the future performance from the knowledge of attitude score"?

Example: An opinion questionnaire, which is the basis for forecasting the demand for a product has predictive validity. The procedure for predictive validity is to first measure the attitude and then predict the future behaviour. Finally, this is followed by the measurement of future behaviour at an appropriate time. Compare the two results (past and future). If the two scores are closely associated, then the scale is said to have predictive validity.

4. ***Criterion Validity:***
   (a) Examines whether measurement scale performs as expected in relation to other variables selected as meaningful criteria, i.e., predicted and actual behavior should be similar
   (b) Addresses the question of what construct or characteristic the scale is actually measuring
5. ***Convergent Validity:*** Extent to which scale correlates positively with other measures of the same construct.
6. ***Discriminant Validity:*** Extent to which a measure does not correlate with other constructs from which it is supposed to differ.
7. ***Nomological Validity:*** Extent to which scale correlates in theoretically predicted ways with measures of different but related constructs.

Notes

Figure 5.5: Reliability and Validity on Target



Self Assessment

Fill in the blanks:

14. An ..... questionnaire, which is the basis for forecasting the demand for a product has predictive validity.
15. Those who enjoy high job satisfaction, and have extrovert personalities should exhibit ..... performance.
16. Reliability deals with ..... and .....
17. There are two criteria to decide whether the scale selected is good or not, viz. ..... and .....

5.6 Summary

- Measurement can be made using nominal, ordinal, interval or ratio scale.
- The scales show the extent of likes/dislikes, agreement disagreement or belief towards an object.
- Each of the scale has certain statistical implications.
- There are four types of scales used in market research namely paired comparison, Likert, semantic differential and thurstone scale.
- Likert is a five point scale whereas semantic differential scale is a seven point scale.
- Bipolar adjectives are used in semantic differential scale.
- Thurstone scale is used to assess attitude of the respondents group regarding any issue of public interest.
- Validity and reliability of the scale is verified before the scale is used for measurement.
- Validity refers to "Does the scale measure what it intends to measure".
- There are three methods to check the validity which type of validity is required depends on "What is being measured".

Notes

5.7 Keywords

Interval Scale: Interval scale may tell us "How far the objects are apart with respect to an attribute?"

Likert Scale: This consists of a series of statements concerning an attitude object. Each statement has '5 points', Agree and Disagree on the scale.

Ordinal Scale: The ordinal scale is used for ranking in most market research studies.

Ratio Scale: Ratio scale is a special kind of internal scale that has a meaningful zero point.

Reliability: It means the extent to which the measurement process is free from errors.

5.8 Review Questions

1. What do you analyse as the merits of Thurstone Scale?
2. What might be the limitations of Thurstone Scale?
3. Which do you find to be more favorable of the attribute and non-attribute method of perceptual mapping and why?
4. In your opinion, what might be the uses of multi dimensional scaling?
5. One of the limitations of MDS can be that it keeps changing from time to time. What else than this do you see as the major drawbacks it has?
6. What can be the reasons for which you think that maintaining reliability can become difficult?
7. Does measurement scale always perform as expected in relation to other variables selected as meaningful criteria? Why/why not?
8. On an average, how many cups of tea do you drink in a day and why? Reply technically.
9. Explain the construction of
   (a) Likert scale
   (b) Semantic differential scale
   (c) Thurstone scale
10. Despite reliability, a scale may not have content validity. Comment.
11. Identify the type of scale, you will use in each of the following (ordinal, nominal, internal, ratio). Justify your answer.

Answers: Self Assessment

1.	Interval	2.	zero point
3.	attitude	4.	multiple
5	Attitude	6.	hypothesis
7.	semantic	8.	Multidimensional
9.	equal appearing interval	10.	Likert
11.	item part, evaluation part	12.	Rank Order

13. two or more
14. opinion
15. high
16. accuracy, consistency
17. reliability, validity

Notes

5.9 Further Readings

Books

A Parasuraman, *Marketing Research*, Dhruv Grewal, Biztantra.

Cisnal Peter, *Marketing Research*, MCGE.

CR Kotari, *Research Methodology*, Vishwa Prakashan.

David Luck and Ronald Rubin, *Marketing Research*, PHI.

GC Beri, *Marketing Research*, TMH.

Hague & Morgan, *Marketing Research in Practice*, Kogan page.

Paneerselvam, R, *Research Methods*, PHI.

S.N. Murthy & U. Bhojanna, *Business Research Methods*, 3rd Edition, Excel Books.

Tull and Donalds, *Marketing Research*, MMIL.

Notes

Unit 6: Primary Data and Questionnaire

CONTENTS

Objectives

After studying this unit, you will be able to:

- Recognize the methodology of collecting primary data
- Define a questionnaire and its characteristics
- Generalize the steps involved in questionnaire designing
- Identify to design survey research

Notes

Introduction

The data directly collected by the researcher, with respect to the problem under study, is known as primary data. Primary data is also the firsthand data collected by the researcher for the immediate purpose of the study.

Primary data is the data that is collected by the researchers for the purpose of investigation. This data is original in character and generated by surveys. Primary data is the information collected during the course of experiment in an experimental research. It can also be obtained through observations or through direct communication with the persons associated with the selected subject by performing surveys or descriptive research.

6.1 Methodology for Collection of Primary Data

Many times due to inadequacy of data or stale information, the need arises for collecting a fresh first hand information. In marketing research, there are broadly two ways by which primary information can be gathered namely, observation and communication.

Benefits & Limitations of Primary data

Benefits of Primary data cannot be neglected. A research can be conducted without secondary data but a research based on only secondary data is least reliable and may have biases because secondary data has already been manipulated by human beings. In statistical surveys it is necessary to get information from primary sources and work on primary data: for example, the statistical records of female population in a country cannot be based on newspaper, magazine and other printed sources. One such source is old and secondly they contain limited information as well as they can be misleading and biased.

1. ***Validity:*** Validity is one of the major concerns in a research. Validity is the quality of a research that makes it trustworthy and scientific. Validity is the use of scientific methods in research to make it logical and acceptable. Using primary data in research can improves the validity of research. First hand information obtained from a sample that is representative of the target population will yield data that will be valid for the entire target population.
2. ***Authenticity:*** Authenticity is the genuineness of the research. Authenticity can be at stake if the researcher invests personal biases or uses misleading information in the research. Primary research tools and data can become more authentic if the methods chosen to analyze and interpret data are valid and reasonably suitable for the data type. Primary sources are more authentic because the facts have not been overdone. Primary source can be less authentic if the source hides information or alters facts due to some personal reasons. There are methods that can be employed to ensure factual yielding of data from the source.
3. ***Reliability:*** Reliability is the certainty that the research is enough true to be trusted on. For example, if a research study concludes that junk food consumption does not increase the risk of cancer and heart diseases. This conclusion should have to be drawn from a sample whose size, sampling technique and variability is not questionable. Reliability improves with using primary data. In the similar research mentioned above if the researcher uses experimental method and questionnaires the results will be highly reliable. On the other hand, if he relies on the data available in books and on internet he will collect information that does not represent the real facts.

Notes

One limitation of primary data collection is that it consumes a lot of time. The researchers will need to make certain preparations in order to handle the different demands of the processes and at the same time, manage time effectively. Besides time consumption, the researchers will collect large volumes of data when they collect primary data. Since they will interact with different people, they will end up with large volumes of data, which they will need to go through when analyzing and evaluating their findings. The primary data also require the greater proportion of workforce to be engaged in the collection of information and analysis, which enhances complexity of operations. There is requirement of large amount of resources to collect primary data.

There are several methods of collecting the primary data, which are as follows:

- Observation Method
- Interview Method
- Through Questionnaires
- Through Schedules

Other methods such as warranty cards, distributor audits, pantry audits, consumer panels, using mechanical devices, through projective techniques, depth interviews and content analysis.

Observation and questioning are two broad approaches available for primary data collection. The major difference between the two approaches is that in the questioning process, the respondents play an active role because of their interaction with the researcher.

Self Assessment

Fill in the blanks:

1. ……..and ……are two broad approaches available for primary data collection.
2. The major difference between the observation and questioning approaches is that in the …….process

6.2 Observation Method

In the observation method, only present/current behaviour can be studied. Therefore, many researchers feel that this is a great disadvantage. A causal observation could enlighten the researcher to identify the problem. Such as the length of the queue in front of a food chain, price and advertising activity of the competitor etc. Observation is the least expensive mode of data collection.

Example: Suppose a Road Safety Week is observed in a city and the public is made aware of advance precautions while walking on the road. After one week, an observer can stand at a street corner and observe the number of people walking on the footpath and those walking on the road during a given period of time. This will tell him whether the campaign on safety is successful or unsuccessful.

Sometimes, observation will be the only method available to the researcher.

Example: Behaviour or attitude of the children, and also of those who are inarticulate.

Notes

6.2.1 Types of Observation Methods

There are several methods of observation of which any one or a combination of some of them, could be used by the observer. Some of these are:

- Structured or unstructured method
- Disguised or undisguised method
- Direct-indirect observation
- Human-mechanical observation

Structured-Unstructured Observation

Whether the observation should be structured or unstructured depends on the data needed.

Example: A manager of a hotel wants to know "how many of his customers visit the hotel with their families and how many come as single customers. Here, the observation is structured, since it is clear "what is to be observed". He may instruct his waiters to record this. This information is required to decide requirements of the chairs and tables and also the ambience.

Suppose, the manager wants to know how single customers and those with families behave and what their attitudes are like. This study is vague, and it needs a non-structured observation.

Example: To distinguish between structured and unstructured observations, consider a study, investigating the amount of search that goes into the purchase of a soap. On the one hand, the observers could be instructed to stand at one end of a supermarket and record each sample customer's search. This may be observed and recorded as follows: "The purchaser first paused after looking at HLL brand. He looked at the price on of the product, kept the product back on the shelf, then picked up a soap cake of HLL and glanced at the picture on the pack and its list of ingredients, and kept it back. He then checked the label and price for P&G product, kept that back down again, and after a slight pause, picked up a different flavour soap of M/s Godrej Company and placed it in his trolley and moved down the aisle". On the other hand, observers might simply be told to record the "first soap cake examined", by checking the appropriate boxes in the observation form. The 'second situation' represents more structured than the first.

Did u know? The observation method is the only method applicable to study the growth of plants and crops.

Caution To use a more structured approach, it would be necessary to decide precisely what is to be observed and the specific categories and units that would be used to record the observations.

Disguised-Undisguised Observation

In disguised observation, the respondents do not know that they are being observed. In non-disguised observation, the respondents are well aware that they are being observed. In disguised observation, observers often pose as shoppers. They are known as "mystery shoppers". They are paid by research organisations. The main strength of disguised observation is that it allows for registering the true of the individuals.

Notes

In the undisguised method, observations may be restrained due to induced error by the objects of observation. The ethical aspect of disguised observations is still open to question and debate.

Direct-Indirect Observation

In direct observation, the actual behaviour or phenomenon of interest is observed. In indirect observation, the results of the consequences of the phenomenon are observed. Suppose, a researcher is interested in knowing about the soft drinks consumption of a student in a hostel room. He may like to observe empty soft drink bottles dropped into the bin. Similarly, the observer may seek the permission of the hotel owner to visit the kitchen or stores. He may carry out a kitchen/stores audit, to find out the consumption of various brands of spice items being used by the hotel. It may be noted that the success of an indirect observation largely depends on "how best the observer is able to identify physical evidence of the problem under study".

Human-Mechanical Observation

Most of the studies in marketing research are based on human observation, wherein trained observers are required to observe and record their observation. In some cases, mechanical devices such as eye cameras are used for observation. One of the major advantages of electrical/ mechanical devices is that their recordings are free from any subjective bias.

Did u know? It is easier to record structured observation than non-structured observation.

6.2.2 Advantages of Observation Method

1. The original data can be collected at the time of occurrence of the event.
2. Observation is done in natural surroundings. Therefore, the facts emerge more clearly, whereas in a questionnaire, experiments have environmental as well as time constraints.
3. Sometimes, the respondents may not like to part with some of the information. Such information can be obtained by the researcher through observation.
4. Observation can also be done on those who cannot articulate.
5. Any bias on the part of the researcher is greatly reduced in the observation method.

6.2.3 Limitations of Observation Method

1. The observer might wait for longer period at the point of observation. And yet the desired event may not take place. Observation is required over a long period of time and hence may not occur.
2. For observation, an extensive training of observers is required.
3. This is an expensive method.
4. External observation provides only superficial indications. To delve beneath the surface is very difficult. Only overt behaviour can be observed.
5. Two observers may observe the same event, but may draw different inferences.
6. It is very difficult to gather information on (1) Opinions (2) Intentions.

Notes

Tasks What observation technique would you use to gather the following information?

1. What kind of influence do children have on the purchase behaviour of their parents?
2. How do discounts influence the purchase behaviour of customers buying colour TV?
3. A study to find out the potential location for a snack bar in a city.

Self Assessment

Fill in the blanks:

3. A causal observation could enlighten the researcher to …….the problem.
4. In ……observation, the actual behaviour or phenomenon of interest is observed.
5. ……observation provides only superficial indications.

6.3 Survey Research Design

Survey is used most often to describe a method of gathering information from samples of individuals. For example, sample of voters are questioned before elections to determine how the public perceives the candidate and the party. A manufacturer does a survey of the potential market before introducing a new product. Government commissions conduct a survey to gather the factual information, it needs to evaluate existing legislation, etc.

6.3.1 Steps Involved in Designing Survey Method

1. Select and formulate research problem.
2. Select an appropriate survey method.
3. Design the survey method/research design.
4. Conduct survey and collect data.
5. Analyze and report.

Notes **Cover Note**

Researcher needs to send a polite short cover note, especially with mailed questionnaires and it should include the following:

- Introduction to the researcher.
- What the research is all about?
- Why is he conducting the study?
- What will happen with the results?
- Who to contact if respondent has any queries?
- How to return the questionnaire to the researcher?

Notes

6.3.2 Characteristics of Survey

1. Survey is conducted in a natural setting.
2. Survey seeks responses directly form the respondents.
3. Survey is widely used in non-experimental social science research.
4. Often use questionnaire or interview method for data collection.
5. Survey involves real world samples.
6. Often it is quantitative method, but can also be qualitative.
7. It is systematic, follows specific set of rules, a formal and orderly logic of sequence.
8. It is impartial, select sample units without any prejudice and preference.

6.3.3 Purpose of Survey

There are two purposes of survey, they are as follows:

1. ***Information gathering:*** It collects information for a specific purpose. For example, pools, census, customer satisfaction, attitude, etc.
2. ***Theory testing and building:*** Surveys are also used for the purpose of testing and building theory. For example, personality and social psychology theories.

6.3.4 Advantages of Survey

- Access to wide range of participants.
- Collection of large amount of data.
- May be more ethical than experimental designs.

6.3.5 Disadvantages of Survey

- Lack of control.
- Data may be superficial.
- Costly to obtain representative data.

Self Assessment

Fill in the blanks:

6. Survey is widely used in ……social science research
7. Survey seeks responses ……form the respondents.
8. A manufacturer does a survey of the …….market before introducing a new product.

6.4 Survey Methods

6.4.1 Personal Interviews

An interview is called personal when the Interviewer asks the questions face-to-face with the Interviewee. Personal interviews can take place at home, at a shopping mall, on the street, and so on.

Notes

Advantages

- The ability to let the Interviewee see, feel and/or taste a product.
- The ability to find the target population. For example, you can find people who have seen a film much more easily outside a theater in which it is playing than by calling phone numbers at random.
- Longer interviews are sometimes tolerated. Particularly with in-home interviews that have been arranged in advance. People may be willing to talk longer face-to-face than to someone on the phone.

Disadvantages

- Personal interviews usually cost more per interview than other methods.
- Change in the characteristics of the population might make sample non-representative.

6.4.2 Telephone Surveys

It is a process of collecting information from sample respondents by calling them over telephone. Surveying by telephone is the most popular interviewing method.

Advantages

- People can usually be contacted faster over the telephone than with other methods.
- You can dial random telephone numbers when you do not have the actual telephone numbers of potential respondents.
- Skilled interviewers can often invite longer or more complete answers than people will give on their own to mail, e-mail surveys.

Disadvantages

- Many telemarketers have given legitimate research a bad name by claiming to be doing research when they start a sales call.
- The growing number of working women often means that no one is at home during the day. This limits calling time to a "window" of about 6-9 p.m. (when you can be sure to interrupt dinner or a favorite TV program).
- You cannot show sample products by phone.

6.4.3 Computer Direct Interviews

These are methods in which the respondents key in (enter) their answers directly into a computer.

Advantages

- It eliminates data entry and editing costs.
- Answers are more accurate to sensitive questions through a computer than to a person or paper questionnaire.
- Interviewer bias is eliminated. Different interviewers can ask questions in different ways, leading to different results. The computer asks the questions the same way every time.

Notes

- Ensuring skip patterns are accurately followed. The Survey System can ensure people are not asked questions they should skip based on their earlier answers. These automatic skips are more accurate than relying on an Interviewer reading a paper questionnaire.
- Response rates are usually higher as it looks novel and interesting to some people.

Disadvantages

- The interviewees must have access to a computer or it must be provided for them.
- As with mail surveys, computer direct interviews may have serious response rate problems in populations due to literacy levels being low.

6.4.4 E-mail Surveys

Email Questionnaire is a new type of questionnaire system that revolutionizes the way on-line questionnaires are conducted. Unlike other on-line questionnaire systems that need a web server to construct, distribute and manage results, Email Questionnaire is totally email based. It works with the existing email system making on-line questionnaire surveys available to anyone with an Internet connection.

Advantages

- Speed: An email questionnaire can gather several thousand responses within a day or two.
- There is practically no cost involved once the set up has been completed.
- Pictures and sound files can be attached.
- The novelty element of an email survey often stimulates higher response levels than ordinary mail surveys.

Disadvantages

- Researcher must possess or purchase a list of email addresses.
- Some people will respond several times or pass questionnaires along to friends to answer.
- Many people dislike unsolicited email even more than unsolicited regular mail.
- Findings cannot be generalised with email surveys. People who have email are different from those who do not, even when matched on demographic characteristics, such as age and gender.
- Email surveys cannot automatically skip questions or randomize question.

6.4.5 Internet/Intranet (Web Page) Survey

Web surveys are rapidly gaining popularity. They have major speed, cost, and flexibility advantages, but also significant sampling limitations. These limitations restrict the groups that can be studied using this technique.

Caution Software selection is especially important in internet survey so it should be selected with proper care and after analyzing through feasibility studies

Notes

Advantages

- Web page surveys are extremely fast. A questionnaire posted on a popular Web site can gather several thousand responses within a few hours. Many people who will respond to an email invitation to take a Web survey will do so the first day, and most will do so within a few days.
- There is practically no cost involved once the set up has been completed.
- Pictures can be shown. Some Web survey software can also show video and play sound.
- Web page questionnaires can use complex question skipping logic, randomizations and other features which is not possible with paper questionnaires. These features can assure better data.
- Web page questionnaires can use colors, fonts and other formatting options not possible in most email surveys.
- A significant number of people will give more honest answers to questions about sensitive topics, such as drug use or sex, when giving their answers to a computer, instead of to a person or on paper.
- On an average, people give longer answers to open-ended questions on Web page questionnaires than they do on other kinds of self-administered surveys.

Disadvantages

- Current use of the Internet is far from universal. Internet surveys do not reflect the population as a whole. This is true even if a sample of Internet users is selected to match the general population in terms of age, gender and other demographics.
- People can easily quit in the middle of a questionnaire. They are not as likely to complete a long questionnaire on the Web as they would be if talking with a good interviewer.
- Depending on your software, there is often no control over people responding multiple times to bias the results.

6.4.6 Mail Questionnaire

Mail questionnaire is a paper questionnaire, which is sent to selected respondents to fill and post filled questionnaire back to the researcher.

Advantages

1. Easier to reach a larger number of respondents throughout the country.
2. Since the interviewer is not present face to face, the influence of interviewer on the respondent is eliminated.
3. This is the only kind of survey you can do if you have the names and addresses of the target population, but not their telephone numbers.
4. Mail surveys allow the respondent to answer at their leisure, rather than at the often inconvenient moment they are contacted for a phone or personal interview. For this reason, they are not considered as intrusive as other kinds of interviews.
5. Where the questions asked are such that they cannot be answered immediately, and needs some thinking on the part of the respondent, the respondent can think over leisurely and give the answer.

Notes

6. Saves cost (cheaper than interview).
7. No need to train interviewers.
8. Personal and sensitive questions are well answered in this method.
9. The questionnaire can include pictures - something that is not possible over the phone.

Limitations

1. It is not suitable when questions are difficult and complicated. Example, Do you believe in value price relationship?
2. When the researcher is interested in a spontaneous response, this method is unsuitable. Because thinking time allowed to the respondent will influence the answer.

   Example, "Tell me spontaneously, what comes to your mind if I ask you about cigarette smoking".
3. In case of a mail questionnaire, it is not possible to verify whether the respondent himself/herself has filled the questionnaire. If the questionnaire is directed towards the housewife, say, to know her expenditure on kitchen items, she alone is supposed to answer it. Instead, if her husband answers the questionnaire, the answer may not be correct.
4. Any clarification required by the respondent regarding questions is not possible.

Example: Prorated discount, product profile, marginal rate, etc., may not be understood by the respondents.

5. If the answers are not correct, the researcher cannot probe further.
6. Poor response (30%) - Not all will reply.
7. In populations of lower educational and literacy levels, response rates to mail surveys are often too small to be useful.

Additional Consideration for the Preparation of Mail Questionnaire

1. It should be shorter than the questionnaire used for a personal interview.
2. The wording should be extremely simple.
3. If a lengthy questionnaire has to be made, first write a letter requesting the co-operation of the respondents.
4. Provide clear guidance, wherever necessary.
5. Send a pre-addressed and stamped envelope to receive the reply.

Self Assessment

Fill in the blanks:

9. Telephone Surveys is a process of collecting information from …….respondents by calling them over telephone.
10. In case of a mail questionnaire, it is not possible to …….whether the respondent himself/herself has filled the questionnaire.
11. ……are methods in which the respondents enter their answers directly into a computer

6.5 Questionnaire

Notes

What is Questionnaire?

A questionnaire is a research instrument consisting of a series of questions and other prompts for the purpose of gathering information from respondents. The questionnaire was invented by Sir Francis Galton.

Notes Importance and Limitations of Questionnaire in MR

Questionnaires have advantages over some other types of data collection. Questionnaires are cheap, do not require as much effort from the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data. However, such standardized answers may frustrate users. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them. Thus, for some demographic groups conducting a survey by questionnaire may not be practical.

Characteristics of Questionnaire

1. It must be simple. The respondents should be able to understand the questions.
2. It must generate replies that can be easily be recorded by the interviewer.
3. It should be specific, so as to allow the interviewer to keep the interview to the point.
4. It should be well arranged, to facilitate analysis and interpretation.
5. It must keep the respondent interested throughout.

6.5.1 Process of Questionnaire Designing

The following are the seven steps involved in designing a questionnaire:

Step 1: Determine What Information is Required

The first question to be asked by the market researcher is "what type of information does he need from the survey?" This is valid because if he omits some information on relevant and vital aspects, his research is not likely to be successful. On the other hand, if he collects information which is not relevant, he is wasting his time and money.

At this stage, information required, and the scope of research should be clear. Therefore, the steps to be followed at the planning stage are:

1. Decide on the topic for research.
2. Get additional information on the research issue, from secondary data and exploratory research. The exploratory research will suggest "what are the relevant variables?"
3. Gather what has been the experience with similar study.
4. The type of information required. There are several types of information such as

   (a) awareness, (b) facts, (c) opinions, (d) attitudes, (e) future plans, (f) reasons.

Facts are usually sought out in marketing research.

Notes

Example: Which television programme did you see last Saturday? This requires a reasonably good memory and the respondent may not remember. This is known as recall loss. Therefore, questioning the distant past should be avoided. Memory of events depends on (1) Importance of the events, and (2) Whether it is necessary for the respondent to remember. In the above case, both the factors are not fulfilled. Therefore, the respondent does not remember. On the contrary, a birthday or wedding anniversary of individuals is remembered without effort since the event is important. Therefore, the researcher should be careful while asking questions about the past. First, he must make sure that the respondent has the answer.

Example: Do you go to the club? He may answer 'yes', though it is untrue. This may be because the respondent wants to impress upon the interviewer that he belongs to a well-to-do family and can afford to spend money on clubs. To obtain facts, the respondents must be conditioned (by good support) to part with the correct facts.

Mode of Collecting the Data

The questionnaire can be used to collect information either through personal interview, mail or telephone. The method chosen depends on the information required and also the type of respondent. If the information is to be collected from illiterate individuals, a questionnaire would be the wrong choice.

Step 2: Different Types of Questionnaire

1. Structured and Non-disguised
2. Structured and Disguised
3. Non-structured and Disguised
4. Non-structured and Non-disguised

1. ***Structured and Non-disguised Questionnaire:*** Here, questions are structured so as to obtain the facts. The interviewer will ask the questions strictly in accordance with the prearranged order. For example, what are the strengths of soap A in comparison with soap B?

   (a) Cost is less

   (b) Lasts longer

   (c) Better fragrance

   (d) Produces more lather

   (d) Available in more convenient sizes

   Structured and non-disguised questionnaire is widely used in market research. Questions are presented with exactly the same wording and same order to all respondents. The reason for standardizing the question is to ensure that all respondents reply the same question. The purpose of the question is clear. The researcher wants the respondent to choose one of the five options given above. This type of questionnaire is easy to administer. The respondents have no difficulty in answering, because it is structured, the frame of reference is obvious.

   In a non-disguised type, the purpose of the questionnaire is known to the respondent.

Notes

Example: "Subjects attitude towards Cyber laws and the need for government legislation to regulate it".

Certainly, not needed at present

Certainly not needed

I can't say

Very urgently needed

Not urgently needed

2. ***Structured and disguised Questionnaire:*** This type of questionnaire is least used in marketing research. This type of questionnaire is used to know the peoples' attitude, when a direct undisguised question produces a bias. In this type of questionnaire, what comes out is "what does the respondent know" rather than what he feels. Therefore, the endeavour in this method is to know the respondent's attitude.

   Currently, the "Office of Profit" Bill is:

   (a) In the Lok Sabha for approval.

   (b) Approved by the Lok Sabha and pending in the Rajya Sabha.

   (c) Passed by both the Houses, pending the presidential approval.

   (d) The bill is being passed by the President.

   Depending on which answer the respondent chooses, his knowledge on the subject is classified.

   In a disguised type, the respondent is not informed of the purpose of the questionnaire. Here the purpose is to hide "what is expected from the respondent?"

Example: "Tell me your opinion about Mr. Ben's healing effect show conducted at Bangalore?"

"What do you think about the Babri Masjid demolition?"

3. ***Non-Structured and Disguised Questionnaire:*** The main objective is to conceal the topic of enquiry by using a disguised stimulus. Though the stimulus is standardized by the researcher, the respondent is allowed to answer in an unstructured manner. The assumption made here is that individual's reaction is an indication of respondent's basic perception. Projective techniques are examples of non-structured disguised technique. The techniques involve the use of a vague stimulus, which an individual is asked to expand or describe or build a story, three common types under this category are (a) Word association (b) Sentence completion (c) Story telling.

4. ***Non structured and Non disguised Questionnaire:*** Here the purpose of the study is clear, but the responses to the question are open-ended. Example: "How do you feel about the Cyber law currently in practice and its need for further modification"? The initial part of the question is consistent. After presenting the initial question, the interview becomes very unstructured as the interviewer probes more deeply. Subsequent answers by the respondents determine the direction the interviewer takes next. The question asked by the interviewer varies from person to person. This method is called "the depth interview". The major advantage of this method is the freedom permitted to the interviewer. By not restricting the respondents to a set of replies, the experienced interviewers will be above

Notes

to get the information from the respondent fairly and accurately. The main disadvantage of this method of interviewing is that it takes time, and the respondents may not cooperate. Another disadvantage is that coding of open-ended questions may pose a challenge. For example: When a researcher asks the respondent "Tell me something about your experience in this hospital". The answer may be "Well, the nurses are slow to attend and the doctor is rude. 'Slow' and 'rude' are different qualities needing separate coding. This type of interviewing is extremely helpful in exploratory studies.

Step 3: Type of Questions

Open-ended Questions

These are questions where respondents are free to answer in their own words. Example: "What factor do you consider while buying a suit"? If multiple choices are given, it could be colour, price, style, brand, etc., but some respondents may mention attributes which may not occur to the researcher.

Therefore, open-ended questions are useful in exploratory research, where all possible alternatives are explored. The greatest disadvantage of open-ended questions is that the researcher has to note down the answer of the respondents verbatim. Therefore, there is a likelihood of the researcher failing to record some information.

Another problem with open-ended question is that the respondents may not use the same frame of reference.

Example: "What is the most important attribute in a job?"

Ans: Pay

The respondent may have meant "basic pay" but interviewer may think that the respondent is talking about "total pay including dearness allowance and incentive". Since both of them refer to pay, it is impossible to separate two different frames.

Dichotomous Question

These questions have only two answers, 'Yes' or 'no', 'true' or 'false' 'use' or 'don't use'.

Do you use toothpaste? Yes …….. No ……

There is no third answer. However sometimes, there can be a third answer:

Example: "Do you like to watch movies?"

Ans: Neither like nor dislike.

Dichotomous questions are most convenient and easy to answer. A major disadvantage of dichotomous question is that it limits the respondent's response. This may lead to measurement error.

Close-Ended Questions

There are two basic formats in this type:

- Make one or more choices among the alternatives.
- Rate the alternatives.

Notes

Choice Among Alternatives

Which of the following words or phrases best describes the kind of person you feel would be most likely to use this product, based on what you have seen in the commercial?

1. Young …… old …….

   Single …… Married ……..

   Modern …… Old fashioned ……...

2. Rating Scale

   (i) Please tell us your overall reaction to this commercial?

      (a) A great commercial; would like to see again.

      (b) Just so-so, like other commercials.

      (c) Another bad commercial.

      (d) Pretty good commercial.

   (ii) Based on what you saw in the commercial, how interested do you feel, you would be buying the products?

      (a) Definitely

      (b) Probably I would buy

      (c) I may or may not buy

      (d) Probably I would not buy

      (e) Definitely I would not buy.

Closed-ended questionnaires are easy to answer. It requires less effort on the part of the interviewer. Tabulation and analysis is easier. There are lesser errors, since the same questions are asked to everyone. The time taken to respond is lesser. We can compare the answer of one respondent to another respondent.

Notes One basic criticism of closed-ended questionnaires is that middle alternatives are not included in this, such as "don't know". This will force the respondents to choose among the given alternative.

Step 4: Wordings of Questions

Wordings of particular questions could have a large impact on how the respondent interprets them. Even a small shift in the wording could alter the respondent's answer.

Example: "Don't you think that Brazil played poorly in the FIFA cup?" The answer will be 'yes'. Many of them, who do not have any idea about the game, will also most likely say 'yes'. If the question is worded in a slightly different manner, the response will be different.

Example: "Do you think that, Brazil played poorly in the FIFA cup?" This is a straightforward question. The answer could be 'yes', 'no' or 'don't know' depending on the knowledge the respondents have about the game.

Notes


Example: "Do you think anything should be done to make it easier for people to pay their phone bill, electricity bill and water bill under one roof"?


Example: "Don't you think something might be done to make it easier for people to pay their phone bill, electricity bill, water bill under one roof"?

A change of just one word as above, can generate different responses by respondents.

Guidelines towards the use of correct wording:

Is the vocabulary simple and familiar to the respondents?


Example: Instead of using the word 'reasonably', 'usually', 'occasionally', 'generally', 'on the whole'.


Example: "How often do you go to a movie?" "Often, may be once a week, once a month, once in two months or even more."

Avoid Double-Barreled Questions

These are questions, in which the respondent can agree with one part of the question, but not agree with the other or cannot answer without making a particular assumption.


Example: "Do you feel that firms today are employee-oriented and customer-oriented?" There are two separate issues here - [yes] [no]


Example: "Are you happy with the price and quality of branded shampoo?" [yes] [no]

Avoid Leading and Loading Questions

1. ***Leading Questions:*** A leading question is one that suggests the answer to the respondent. The question itself will influence the answer, when respondents get an idea that the data is being collected by a company. The respondents have a tendency to respond positively.


Example: "How do you like the programme on 'Radio Mirchy'? The answer is likely to be 'yes'. The unbiased way of asking is "which is your favorite F.M. Radio station? The answer could be any one of the four stations namely (1) Radio City (2) Mirchy (3) Rainbow (4) Radio-One.


Example: Do you think that offshore drilling for oil is environmentally unsound? The most probable response is 'yes'. The same question can be modified to eliminate the leading factor.

What is your feeling about the environmental impact of offshore drilling for oil? Give choices as follows:

(a) Offshore drilling is environmentally sound.

(b) Offshore drilling is environmentally unsound.

(c) No opinion.

Notes

2. ***Loaded Questions:*** A leading question is also known as a loaded question. In a loaded question, special emphasis is given to a word or a phrase, which acts as a lead to respondent.

Example: "Do you own a Kelvinator refrigerator." A better question would be "what brand of refrigerator do you own?" "Don't you think the civic body is 'incompetent'?" Here the word incompetent is 'loaded'.

(a) Are the Questions Confusing? If there is a question unclear or is confusing, then the respondent becomes more biased rather that getting enlightened. Example: "Do you think that the government publications are distributed effectively"? This is not the correct way, since respondent does not know what is the meaning of the word effective distribution. This is confusing. The correct way of asking questions is "Do you think that the government publications are readily available when you want to buy?" Example: "Do you think whether value price equation is attractive"? Here, respondents may not know the meaning of value price equation.

(b) *Applicability:* "Is the question applicable to all respondents?" Respondents may try to answer a question even though they don't qualify to do so or may lack from any meaningful opinion.

Example:

1. "What is your present education level"
2. "Where are you working" (assuming he is employed)?
3. "From which bank have you taken a housing loan" (assuming he has taken a loan).

Avoid Implicit Assumptions

An implicit alternative is one that is not expressed in the options. Consider following two questions:

1. Would you like to have a job, if available?
2. Would you prefer to have a job, or do you prefer to do just domestic work?

Even though, we may say that these two questions look similar, they vary widely. The difference is that Q-2 makes explicit the alternative implied in Q-1.

Split Ballot Technique

This is a procedure used wherein (1) The question is split into two halves and (2) Different sequencing of questions is administered to each half. There are occasions when a single version of questions may not derive the correct answer and the choice is not obvious to the respondent.

Example: "Why do you use Ayurvedic soap"? One respondent might say "Ayurvedic soap is better for skin care". Another may say "Because the dermatologist has recommended". A third might say "It is a soap used by my entire family for several years". The first respondent answers the reason for using it at present. The second respondent answers how he started using. The third respondent "the family tradition for using". As can be seen, different reference frames are used. The question may be balanced and rephrased.

Notes

Complex Questions?

In which of the following do you like to park your liquid funds?

i. Debenture

ii. Preferential share

iii. Equity linked MF

iv. IPO

v. Fixed deposit

If this question is posed to the general public, they may not know the meaning of liquid fund. Most of the respondents will guess and tick one of them.

Are the Questions Too Long? Generally as a thumb rule, it is advisable to keep the number of words in a question not exceeding 20. The question given below is too long for the respondent to comprehend, leave alone answer.

Example: Do you accept that the people whom you know, and associate yourself have been receiving ESI and P.F. benefits from the government accept a reduction in those benefits, with a view to cut down government expenditure, to provide more resources for infrastructural development?

Yes..... No..... Can't say.....

Participation at the Expense of Accuracy

Sometimes the respondent may not have the information that is needed by the researcher.

Example: The husband is asked a question "How much does your family spend on groceries in a week"? Unless the respondent does the grocery shopping himself, he will not know how much has been spent. In a situation like this, it will be helpful to ask a 'filtered question'. An example of a filtered question can be, "Who buys the groceries in your family"?

Example: "Do you have the information of Mr. Ben's visit to Bangalore"? Not only should the individual have the information but also s(he) should remember the same. The inability to remember the information is known as "recall loss".

Task Give one example for each of the following type of the questions:

1. Leading question
2. Double-barreled question
3. Close-ended question
4. Fixed alternative question
5. Split-ballot question

Step 5: Sequence and Layout

Notes

Some guidelines for sequencing the questionnaire are as follows:

Divide the questionnaire into three parts:

1. Basic information
2. Classification
3. Identification information.

Items such as age, sex, income, education, etc., are questioned in the classification section. The identification part involves body of the questionnaire. Always move from general to specific questions on the topic. This is known as funnel sequence. Sequencing of questions is illustrated below:

(1) Which TV shows do you watch?

Sports ..... News .....

(2) Which among the following are you most interested in?

Sports ..... News .....

Music ..... Cartoon .....

(3) Which show did you watch last week?

World Cup Football .....

Bournvita Quiz Contest .....

War News in the Middle East .....

Tom and Jerry cartoon show .....

The above three questions follow a funnel sequence. If we reverse the order of question and ask "which show was watched last week"?, the answer may be biased. This example shows the importance of sequencing.

Layout: How the questionnaire looks or appears.

Example: Clear instructions, gaps between questions, answers and spaces are part of layout. Two different layouts are shown below:

Layout - 1 How old is your bike?

..... Less than 1 year ..... 1 to 2 years ..... 2 to 4 years ..... more than 4 years.

Layout - 2 How old is your bike?

..... Less than 1 year

..... 1 to 2 years.

..... 2 to 4 years.

..... More than 4 years.

From the above example, it is clear that layout - 2 is better. This is because likely respondent error due to confusion is minimised.

Therefore, while preparing a questionnaire start with a general question. This is followed by a direct and simple question. This is followed by more focused questions. This will elicit maximum information.

Notes

Forced and Unforced Scales

Suppose the questionnaire is not provided with 'don't know' or 'no option', then the respondent is forced to choose one side or the other. A 'don't know' is not a neutral response. This may be due to genuine lack of knowledge.

Balanced and Unbalanced Scales

In a balanced scale, the number of favourable responses are equal to the number of unfavorable responses. If the researcher knows that there is a possibility of a favourable response, it is best to use unbalanced scale.

Use Funnel Approach

Funnel sequencing gets the name from its shape, starting with broad questions and progressively narrowing down the scope. Move from general to specific examples.

1. How do you think this country is getting along in its relations with other countries?
2. How do you think we are doing in our relations with the US?
3. Do you think we ought to be dealing with US?
4. If yes, what should be done differently?
5. Some say we are very weak on the nuclear deal with the US, while, some say we are OK. What do you feel ?

The first question introduces the general subject. In the next question, a specific country is mentioned. The third and fourth questions are asked to seek views. The fifth question is to seek a specific opinion.

Step 6: Pretesting of Questionnaire

Pretesting of a questionnaire is done to detect any flaws that might be present. For example, the word used by researcher must convey the same meaning to the respondents. Are instructions clear skip questions clear? One of the prime conditions for pretesting is that the sample chosen for pretesting should be similar to the respondents who are ultimately going to participate. Just because a few chosen respondents fill in all the questions going does not mean that the questionnaire is sound.

How Many Questions to be Asked? The questionnaire should not be too long as the response will be poor. There is no rule to decide this. However, the researcher should consider that if he were the respondent, how would he react to a lengthy questionnaire. One way of deciding the length of the questionnaire is to calculate the time taken to complete the questionnaire. He can give the questionnaire to a few known people to seek their opinion.

Step 7: Revise and Preparation of Final Questionnaire

Final questionnaire may be prepared after pretesting the questionnaire with the small group of respondents. Questionnaire should be revised for the following:

i. To correct the spellings.

ii. To place the questions in proper order to avoid the contextual bias.

iii. To remove the words which are not familiar to respondents.

Notes

iv. To add or remove questions arise in the process of pretest, if any.

v. To purge the words with double meaning, etc.

Self Assessment Questions

Fill in the blanks:

12. Generally as a thumb rule, it is advisable to keep the number of words in a question not exceeding …….

13. In a ……..scale, the number of favourable responses are equal to the number of unfavorable responses.

14. A major disadvantage of dichotomous question is that it ----------the respondent's response..

15. Open-ended questions are useful in ……research, where all possible alternatives are explored.

6.6 Summary

- Primary data may pertain to life style, income, awareness or any other attribute of individuals or groups.
- There are mainly two ways of collecting primary data namely: (a) Observation (b) By questioning the appropriate sample.
- Observation method has a limitation i.e., certain attitudes, knowledge, motivation, etc. cannot be measured by this method. For this reason, researcher needs to communicate.
- Communication method is classified based on whether it is structured or disguised.
- Structured questionnaire is easy to administer. This type is most suited for descriptive research. If the researcher wants to do exploratory sturdy, unstructured method is better.
- In unstructured method questions will have to be framed based on the answer by the respondent. Questionnaire can be administered either in person or online or Mail questionnaire. Each of these methods have advantages and disadvantages.
- Questions in a questionnaire may be classified into (a) Open question (b) Close ended questions (c) Dichotomous questions, etc.
- While formulating questions, care has to be taken with respect to question wording, vocabulary, leading, loading and confusing questions should be avoided. Further it is desirable that questions should not be complex, nor too long.
- It is also implied that proper sequencing will enable the respondent to answer the question easily. The researcher must maintain a balanced scale and must use a funnel approach.
- Pretesting of the questionnaire is preferred before introducing to a large population.

6.7 Keywords

Computer Direct Interview: This is the method in which the respondents key in (enter) their answers directly into a computer.

Dichotomous Question: These questions have only two answers, like 'Yes' or 'no'

Disguised Observation: The observation under which the respondents do not know that they are being observed.

Notes

Loaded Question: A question in which special emphasis is given to a word or a phrase, which acts as a lead to respondent.

Non-disguised Observation: The observation in which the respondents are well aware that they are being observed.

6.8 Review Questions

1. What is primary data?
2. What are the various methods available for collecting primary data?
3. What are the advantages and disadvantages of a structured questionnaire?
4. What are the several methods used to collect data by observation method?
5. What are the advantages and limitations of collecting data by observation method?
6. What are the various methods of survey research?
7. What is a questionnaire? What are its importance and characteristics?
8. Explain the steps involved in designing a questionnaire.
9. Explain Open ended and Closed ended questions in a questionnaire.
10. One method of sequencing the question in a questionnaire is to proceed from general to specific. What is the logical reason behind this?

Answers: Self Assessment

1. Observation, questioning
2. questioning
3. identify
4. direct
5. External
6. non-experimental
7. directly
8. potential
9. sample
10. verify
11. Computer Direct Interviews
12. 20
13. balanced
14. limits
15. exploratory

6.9 Further Readings

Books

Abrams, M.A., *Social Surveys and Social Action*, London: Heinemann, 1951.

Arthur, Maurice, *Philosophy of Scientific Investigation*, Baltimore: John Hopkins University Press, 1943.

Bernal, J.D., *The Social Function of Science*, London: George Routledge and Sons, 1939.

Chase, Stuart, *The Proper Study of Mankind: An inquiry into the Science of Human Relations*, New York, Harper and Row Publishers, 1958.

S. N. Murthy and U. Bhojanna, *Business Research Methods*, Excel Books.

Unit 7: Secondary Data

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Recognize the notion of secondary data
- Differentiate the various types of secondary data.
- Identify the special techniques of secondary data
- Generalize the advantages and disadvantages of secondary data

Introduction

In research, secondary data is collecting and possibly processed by people other than the researcher in question. Common sources of secondary data for social science include censuses, large surveys, and organizational records. In sociology primary data is data you have collected yourself and secondary data is data you have gathered from primary sources to create new research. In terms of historical research, these two terms have different meanings. A primary source is a book or set of archival records. A secondary source is a summary of a book or set of records.

7.1 Secondary Data

Secondary data are statistics that already exist. They have been gathered not for immediate use. This may be described as "those data that have been compiled by some agency other than the user". Secondary data can be classified as:

Notes

1. Internal secondary data
2. External secondary data

7.1.1 Internal Secondary Data

Internal secondary data is a part of the company's record, for which research is already conducted. Internal data are those that are found within the organisation.

Example: Sales in units, credit outstanding, call reports of sales persons, daily production report, monthly collection report, etc.

7.1.2 External Secondary Data

The data collected by the researcher from outside the company. This can be divided into four parts:

1. Census data
2. Individual project report being published
3. Data collected for sale on a commercial basis called syndicated data
4. Miscellaneous data

Did u know? Census data is the most important data among the sources of data.

The following are some of the data that can obtained by census records:

- Census of the wholesale trade
- Census of the retail trade
- Population Census
- Census of manufacturing industries
- Individual project report being published
- Encyclopedia of business information sources
- Product finder
- Thomas registers etc.

7.1.3 Benefits and Limitations of Secondary Data

Benefits

It is far cheaper to collect secondary data than to obtain primary data. For the same level of research budget a thorough examination of secondary sources can yield a great deal of information than can be had through a primary data collection exercise.

The time involved in searching secondary sources is much less than that needed to complete primary data collection.

Secondary sources of information can yield more accurate data than that obtained through primary research. This is not always true but where a government or international agency has

Notes

undertaken a large scale survey, or even a census, this is likely to yield far more accurate results than custom designed and executed surveys when these are based on relatively small sample sizes.

It should not be forgotten that secondary data can play a substantial role in the exploratory phase of the research when the task at hand is to define the research problem and to generate hypotheses. The assembly and analysis of secondary data almost invariably improve the researcher's understanding of the marketing problem, the various lines of inquiry that could or should be followed and the alternative courses of actions which might be pursued.

Secondary sources help define the population. Secondary data can be extremely useful both in defining the population and in structuring the sample to be taken. For instance, government statistics on a country's agriculture will help decide how to stratify a sample and, once sample estimates have been calculated, these can be used to project those estimates to the population.

Limitations

1. ***Definition:*** The researcher, when making use of secondary data, may misinterpret the definitions used by those responsible for its preparation and draw erroneous conclusions
2. ***Measurement error:*** When a researcher conducts fieldwork she/he is possibly able to estimate inaccuracies in measurement through the standard deviation and standard error, but these are sometimes not published in secondary sources. The problem is sometimes not so much 'error' but differences in the levels of accuracy required by decision makers.
3. ***Source bias:*** Researchers face the problem of vested interests when they consult secondary sources. Those responsible for their compilation may have reasons for wishing to present a more optimistic or pessimistic set of results for their organization i.e., exaggerated figures or inflated estimates may be stated.
4. ***Reliability:*** The reliability of published statistics may vary over time. Because the systems of collecting data or geographical or administrative boundaries may be changed, or the basis for stratifying a sample may have altered. Other aspects of research methodology that affect the reliability of secondary data is the sample size, response rate, questionnaire design and modes of analysis without any indication of this to the reader of published statistics.
5. ***Time scale:*** The time period during which secondary data was first compiled may have a substantial effect upon the nature of the data for example: Most censuses take place at ten-year intervals, so data from this and other published sources may be out-of-date at the time the researcher wants to make use of the statistics.

Self Assessment

Fill in the blanks:

1. Secondary data can be classified as …….and ……secondary data
2. Those data that have been compiled by some agency other than the user are known as ……..data.
3. Internal secondary data is a part of the …….record
4. External Secondary Data can be divided into ……parts.
5. …….Secondary Data is the data collected by the researcher from outside the company.

Notes

7.2 Special Techniques of Market Research or Syndicated Data

These techniques involve data collection on a commercial basis i.e., data collected by this method is sold to interested clients on payment.

Example: Examples of organizations involved in collecting syndicated data are A.C. Nielson, ORG Marg, IMRB, etc. They provide business relationship survey called BRS which estimates the following:

1. Rating
2. Profile of the company etc.
3. These people also provide TRP rating namely television rating points on a regular basis. This provides:
   (i) Viewership figures
   (ii) Duplication between programmes, etc. Some of the interesting studies made by IMRB are SNAP- Study of Nations Attitude and Awareness Programme. In this study, the various groups of India's population and their lifestyles, attitudes of Indian housewives were detailed.

Notes Organizations involved in collecting syndicated data provide NRS called National Readership Survey to the sponsors and advertising agencies.

There is also a study called FSRP which covers children in the age group of 10-19 years. Beside their demographics and psychographics, the study covers areas such as:

- Children as decision-makers
- Role models of Indian children
- Pocket money and its usage
- Media reviews
- Favoured personalities and characteristics
- Brand awareness and advertising recall.

Syndicated sources consist of market research firms offering syndicated services. These market research organisations collect and update information on a continuous basis. Since data is syndicated, its cost is spread over a number of client organisations and hence is cheaper.

Example: A client firm can give certain specific question to be included in the questionnaire, which are used routinely to collect syndicated data. The client will have to pay extra charges for these. The data generated from additional questions and analysis will be revealed only to the firms submitting the questions.

Therefore we can say that the customization of secondary data is possible. Some areas of syndicated services are newspapers, periodical readership, popularity of TV channels, etc. Data from syndicated sources are available on a weekly or monthly basis.

Notes

Syndicated data may be classified as:

(a) Consumer purchase data

(b) Retailer and wholesaler data

(c) Advertising data.

Most of these data collection methods as mentioned above are also known as syndicated data. Syndicated data can be classified into:

7.2.1 Consumer Purchase Data or Panel Type Data

This is one type of syndicated data. In this method there are consumer panels. Members of this panel will be representative of the entire population. Panel members keep diaries in which they record all purchases made by them. Products purchased range from packaged food to personal care products. Members submit the dairies every month to the organisations for which they are paid. This panel data can be used to find out the sales of the product. These panel data also provides an insight into repeat purchases, effect of free samples, coupon redemption etc.

The consumer panel data also provides profile of the target audience. Nowadays, diaries are replaced by hand-held scanners.

Notes Panels provide data on consumer buying habits on petrol, auto parts, sports goods etc.

Limitations

- Low-income groups are not represented
- Some people do not want to take the trouble of keeping records of their purchases. Therefore, relevant data is not available.

Advantages

- Use of scanner tied to the central computer helps the panel members to record their purchases early (almost immediately).
- It also provides reliability and speed.
- Panel can consist only of senior citizens or only children.

We also have the Consumer Mail Panel (CMP). This consists of members who are willing to answer mail questionnaires. A large number of such households are kept on the panel. This serves as a universe through which panels are selected.

7.2.2 Retail and Wholesale Data

Marketing research is done in a retail store. These are organisations that provide continuous data on grocery products. The procedure does not involve questioning people and also does not rely on their memory. This requires cooperation from the retailer to allow auditing to be carried out. Generally, retail audit involves counting of stocks between two consecutive visits.

Notes

It involves inspection of goods delivered between visits. If the stock of any product in the shop is accurately counted during both the visits and data on deliveries are accurately taken from the records, the collection of sales of a product over that period can be determined accurately as follows:

Initial stock + Deliveries between successive visits - Second time stock = Sales

If this information is obtained from different shops from the representative sample of shops, then the accurate estimates of sales of the product can be made. To do this, some shops can be taken as a "Panel of shops" representing the universe.

Advantages

- It provides information between audits on consumer purchase over the counter in specific units. For example, KGs, bottles, No's, etc.
- It provides data on shop purchases i.e., the purchases made by the retailer between audits.
- It is a very reliable method.

Disadvantages

- Experience is needed by the market researcher.
- Cooperation is required from the retail shop.
- It is time consuming.

Did u know? With the help of wholesale and retail data, the manufacturer comes to know how competitor is doing.

7.2.3 Advertising Data

Since a large amount of money is being spent on advertising, data needs to be collected on advertising. One way of recording is by using passive meter. This is attached to a TV set records when the set was 'On'. It will record "How long a channel is viewed". By this method, data regarding audience interest in a channel can be ascertained. One thing to be noticed from the above is that it only tells you that someone is viewing television at home. But it does not tell you "who is viewing at home". To find out "who is viewing" a new instrument called 'People's Meter' is introduced. This is a remote-controlled instrument buttons. Each household is given a specific button. When that button is pressed, it signals the control box that a specific person is viewing. This information is recorded electronically and sent to a computer that stores this information which is subsequently analysed.

Miscellaneous Secondary Data

This data includes trade associations such as FICCI, CEI, Institution of Engineers, Chamber of Commerce, libraries such as public library, university libraries, etc., literature, state and central government publications, private sources such as All India Management Association (AIMA), Financial Express and financial dailies, world bodies and international organizations such as IMF, ADB, etc.

Notes

Task List some major secondary sources of information for the following:

(a) Market research manager of a tea manufacturing company has to prepare a comprehensive report on the tea industry as a whole.

(b) M.T.R. has several product ideas on ready-to-eat products. It wishes to convert ideas into products and enter the market. Before entering, the company needs to find necessary information to assess the market potential.

(c) An MNC wishes to open a showroom in a Metro. The first step that the company would like to take is to collect the information about suitability.

Self Assessment

Fill in the blanks:

6. Syndicated sources consist of …….firms offering syndicated services
7. Retail and Wholesale data provides information between ……..on consumer purchase over the counter in specific units.
8. The consumer panel data provides profile of the ……audience.
9. Data from syndicated sources are available on a ……..or ……..basis.
10. The cost of syndicated data is ……since it is spread over a number of client organisations

7.3 Advantages and Disadvantages of Secondary Data

Advantages

- It is economical, without the need to hire field staff.
- It saves time (normally 2 to 3 months). If data is available on hand it can be tabulated in minutes.
- They provide information, which retailers may not be willing to reveal to researcher.
- No training is required to collect this data, unlike primary data.

Disadvantages

Because secondary data has been collected for some other projects, it may not fit in with the problem that is being defined. In some cases, the feed is so poor that the data becomes completely inappropriate. It may be ill-suited because of the following three reasons:

- Unit of measurement
- Definition of a class
- Recency

Unit of Measurement

It is common for secondary data to be expressed in units.

Notes

Example: Size of the retail establishments, for instance, can be expressed in terms of gross sales, profits, square feet area and number of employees. Consumer incomes can be expressed in variables the individual, family, household etc. Secondary data available may not fit in easily.

Assume that the class intervals are quite different from those which are needed.

Example: Data available with respect to age group is as follows:

<18 year

18-24 years

25-34 years

35-44 years

Suppose the company needs a classification less than 20, 20-30 and 30-40, the above classification of secondary data cannot be used.

Problem of Accuracy

The accuracy of secondary data available is highly questionable. A number of errors are possible in the collection and analysis of the data. Accuracy of secondary data depends upon:

(a) Who has collected the data?

(b) How is the data collected?

(a) ***Who has collected the data:*** The reliability of the source determines the accuracy of the data. Assume that a publisher of a private periodical conducts a survey of his readers. The main aim of the survey is to find out the opinion of readers about advertisements appearing in it. This survey is done by the publisher in the hope that other firms will buy this data before inserting advertisements.

Assume that a professional MR agency has conducted a similar survey and has sold its syndicated data on many periodicals.

If you are an individual who wants information on a particular periodical you buy the data from MR agency rather from the periodical's publisher. The reason for this is trust of the MR agency. The reasons for trusting the MR agency are as follows:

1. Being an independent agency there is no bias. The MR agency is likely to provide an unbiased data.
2. The data quality of MR agency will be good since they are professionals.

(b) ***How the data collected?***

1. What instruments were used?
2. What type of sampling was done?
3. How large was the sample?
4. What was the time period of data collection? Example: days of the week, time of the day.

Notes

Caution Before using the secondary data, the source of data must be verified in order to ensure accuracy and reliability of data.

Recency

This pertains to "how old was the information?" If it is five years old, it may be useless. Therefore, the publication lag is a problem.

Self Assessment

Fill in the blanks:

11. There is no need to hire ……for the purpose of secondary data.
12. Secondary data may be ill-suited because of the three reasons which are……..., Definition of a class and Recency
13. The reliability of the source determines the …….of the data.
14. It is common for secondary data to be expressed in……..
15. ……pertains to "how old was the information?"

7.4 Summary

- Secondary data are statistics that already exists.
- Secondary data may not be readily used because these data are collected for some other purpose.
- Secondary data has its own advantages and disadvantages.
- There are two types of secondary data (1) Internal and (2) External secondary data.
- Census is the most important among secondary data.
- Syndicated data is an important form of secondary data
- Syndicated data may be classified into (a) Consumer purchase data (b) Retailer and wholesale data (c) Advertising data. Each has advantages and disadvantages.

7.5 Keywords

External Data: The data collected by the researcher from outside the company.

Internal Data: Internal data are those that are found within the organisation.

Panel Type Data: This is one type of syndicated data in which there are consumer panels.

Secondary Data: Secondary data is collecting and possibly processed by people other than the researcher in question.

Syndicated Data: Data collected by this method is sold to interested clients on payment.

Notes

7.6 Review Questions

1. What is meant by secondary data?
2. Differentiate between internal and external secondary data.
3. What are the sources of secondary data?
4. What are the types of secondary data?
5. What are the special techniques of secondary data?
6. What is the classification of syndicated data?
7. What are the advantages and limitations of syndicated data?
8. What are the advantages and disadvantages of secondary data?
9. Discuss the sources of secondary data for the study on "consumer purchasing a white good".
10. What is Omnibus Survey?

Answers: Self Assessment

1.	Internal, External	2.	secondary
3.	company's	4.	four
5.	External	6.	market research
7.	audits	8.	target
9.	weekly, monthly	10.	cheaper
11.	field staff	12.	Unit of measurement
13.	accuracy	14.	units
15.	Recency		

7.7 Further Readings

Books

A Parasuraman, Dhruv Grewal, *Marketing Research*, Biztantra.

Cooper and Schinder, Business *Research Methods*, TMH.

CR Kotari, *Research Methodology*, Vishwa Prakashan.

David Luck and Ronald Rubin, *Marketing Research*, PHI.

GA Churchil, *Marketing Research*, Iacobucci, Thomson.

Naresh Amphora, *Marketing Research*, Pearson Education.

OR Krishna Swamy, *Methodology of Research in Social Sciences*, HPH.

S.N Murthy and U. Bhojanna, *Business Research Methods*, Excel Books, 3rd Edition.

William MC Trochim, *Research Methods*, Biztantra.

William Zikmund, *Business Research Methods*, Thomson.

Unit 8: Descriptive Statistics

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Recognize the Meaning and Characteristics various measures of Central Tendency
- Define the Arithmetic Mean
- Describe the Median
- State the impression of Mode
- Explain the Measures of dispersion

Introduction

Let's take a look at the most basic form of statistics, known as descriptive statistics. This branch of statistics lays the foundation for all statistical knowledge. Descriptive Statistics are used to describe the basic features of the data gathered from an experimental study in various ways. A descriptive statistics is distinguished from inductive statistics. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. It is necessary to be familiar with primary methods of describing data in order to understand phenomena and make intelligent decisions. There may be two objectives for formulating a summary statistic: (1) to choose a statistic that shows how different units seem similar. Statistical textbooks call one solution to this objective, a measure of central tendency and (2) to choose another statistic that shows how they differ. This kind of statistic is often called measure dispersion.

Notes

8.1 Measure of Central Tendency

Summarisation of the data is a necessary function of any statistical analysis. As a first step in this direction, the huge mass of unwieldy data are summarised in the form of tables and frequency distributions. In order to bring the characteristics of the data into sharp focus, these tables and frequency distributions need to be summarised further. A measure of central tendency or an average is very essential and an important summary measure in any statistical analysis.

Notes An average is a single value which can be taken as representative of the whole distribution.

Functions of an Average

1. ***To present huge mass of data in a summarised form:*** It is very difficult for human mind to grasp a large body of numerical figures. A measure of average is used to summarise such data into a single figure which makes it easier to understand and remember.
2. ***To facilitate comparison:*** Different sets of data can be compared by comparing their averages. For example, the level of wages of workers in two factories can be compared by mean (or average) wages of workers in each of them.
3. ***To help in decision-making:*** Most of the decisions to be taken in research, planning, etc., are based on the average value of certain variables.

Example: If the average monthly sales of a company are falling, the sales manager may have to take certain decisions to improve it.

Characteristics of a Good Average

A good measure of average must posses the following characteristics:

1. It should be rigidly defined, preferably by an algebraic formula, so that different persons obtain the same value for a given set of data.
2. It should be easy to compute.
3. It should be easy to understand.
4. It should be based on all the observations.
5. It should be capable of further algebraic treatment.
6. It should not be unduly affected by extreme observations.
7. It should not be much affected by the fluctuations of sampling.

Self Assessment

Fill in the blanks:

1. ..... of the data is a necessary function of any statistical analysis.
2. Different sets of data can be compared by comparing their .....

8.2 Various Measures of Average

Notes

Various measures of average can be classified into the following three categories:

1. ***Mathematical Averages:***
   (a) Arithmetic Mean or Mean
   (b) Geometric Mean
   (c) Harmonic Mean
   (d) Quadratic Mean
2. ***Positional Averages:***
   (a) Median
   (b) Mode
3. ***Commercial Average:***
   (a) Moving Average
   (b) Progressive Average
   (c) Composite Average

The above measures of central tendency will be discussed in the order to their popularity. Out of these, the Arithmetic Mean, Median and Mode, being most popular, are discussed in that order.

8.2.1 Arithmetic Mean

Before the discussion of arithmetic mean, we shall introduce certain notations. It will be assumed that there are n observations whose values are denoted by X_1, X_2, ... X_n, respectively. The sum of these observations $X_1 + X_2 + ... + X_n$ will be denoted in abbreviated form as,

$$\sum_{i=1}^{n} X_i$$

where S (called sigma) denotes summation sign. The subscript of X, i.e., 'i' is a positive integer, which indicates the serial number of the observation. Since there are n observations, variation in i will be from 1 to n. This is indicated by writing it below and above S, as written earlier. When there is no ambiguity in range of summation, this indication can be skipped and we may simply write $X_1 + X_2 + ... + X_n = SX_i$.

Arithmetic Mean is defined as the sum of observations divided by the number of observations. It can be computed in two ways:

1. Simple arithmetic mean and
2. Weighted arithmetic mean

In case of simple arithmetic mean, equal importance is given to all the observations while in weighted arithmetic mean, the importance given to various observations is not same.

Calculation of simple arithmetic mean can be done in following ways:

1. ***When Individual Observations are Given***

   Let there be n observations X_1, X_2 ..... X_n. Their arithmetic mean can be calculated either by direct method or by short cut method. The arithmetic mean of these observations will be denoted by $\overline{X}$.

   (a) *Direct Method:* Under this method, $\overline{X}$ is obtained by dividing sum of observations by number of observations, i.e.,

Notes

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

(b) *Shortcut Method:* This method is used when the magnitude of individual observations is large. The use of shortcut method is helpful in the simplification of calculation work.

Let A be any assumed mean. We subtract A from every observation. The difference between an observation and A, i.e., $X_i - A$ is called the deviation of *i*th observation from A and is denoted by d_i. Thus, we can write; $d_1 = X_1 - A$, $d_2 = X_2 - A$, ... $d_n = X_n - A$. On adding these deviations and dividing by n we get

$$\frac{\sum d_i}{n} = \frac{\sum (X_i - A)}{n} = \frac{\sum X_i - nA}{n} = \frac{\sum X_i}{n} - A$$

or $$\bar{d} = \bar{X} - A \left(\text{where } \bar{d} = \frac{\Sigma d_i}{n} \right)$$

On rearranging, we get

$$\bar{X} = A + \bar{d} = A + \frac{\Sigma d_i}{n}$$

This result can be used for the calculation of $\bar{X}$.

Notes Theoretically we can select any value as assumed mean. However, for the purpose of simplification of calculation work, the selected value should be as nearer to the value of $\bar{X}$ as possible.

Example: The following figures relate to monthly output of cloth of a factory in a given year:

Months	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Output (in '000 metres)	80	88	92	84	96	92	96	100	92	94	98	86

Calculate the average monthly output.

Solution:

(a) *Using Direct Method:*

$$\bar{X} = \frac{80 + 88 + 92 + 84 + 96 + 92 + 96 + 100 + 92 + 94 + 98 + 96}{12}$$

= 91.5 ('000 mtrs)

(b) *Using Shortcut Method:*

Let A = 90.

X_i	80	88	92	84	96	92	96	100	92	94	98	86	**Total**
$d_i = X_i - A$	–10	–2	2	–6	6	2	6	10	2	4	8	–4	Σd_i **= 18**

$$\bar{X} = 90 + \frac{18}{12}$$

= 90 + 1.5 = 91.5 thousand mtrs

2. ***When Data are in the form of an Ungrouped Frequency Distribution*** **Notes**

Let there be n values X_1, X_2, ... X_n out of which X_1 has occurred f_1 times, X_2 has occurred f_2 times, ... X_n has occurred fn times. Let N be the total frequency, i.e.,

$N = \sum_{i=1}^{n} f_i$

Alternatively, this can be written as follows:

Values	X_1	X_2	.....	X_n	Total Frequency
Frequency	f_1	f_2	.....	f_n	N

(a) *Direct Method:* The arithmetic mean of these observations using direct method is given by

$$X = \frac{\underbrace{X_1 + X_1 + ... + X_1}_{F_1 \text{ times}} + \underbrace{X_2 + ... + ... + X_2}_{F_2 \text{ times}} + ... + \underbrace{X_n + ... + X_n}_{F_3 \text{ times}}}{F_1 + F_2 + ... F_n}$$

Since $X_1 + X_1 + ... + X_1$ added F_1 times can also be written $F_1 X_1$. Similarly, by writing other observation in same manner, we have

$$\overline{X} = \frac{f_1X_1 + f_2X_2 + ... + f_nX_n}{f_1 + f_2 + ... + f_n} = \frac{\sum_{i=1}^{n} f_iX_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_iX_i}{N} \quad ...(1)$$

(b) *Shortcut Method:* As before, we take the deviations of observations from an arbitrary value A. The deviation of *i*th observation from A is $d_i = X_i - A$.

Multiplying both sides by f_i we have $f_i\, d_i = f_i(X_i - A)$

Taking sum over all the observations

$\Sigma f_i d_i = \Sigma f_i (X_i - A) = \Sigma f_i X_i - A\Sigma f_i = \Sigma f_i X_i - A \times N$

Dividing both sides by N we have

$$\frac{\Sigma f_i d_i}{N} = \frac{\Sigma f_i X_i}{N} - A = \overline{X} - A \quad \text{or} \quad \overline{X} = A + \frac{\Sigma f_i d_i}{N} = A + \overline{d}$$

3. ***When Data are in the form of a Grouped Frequency Distribution***

In a grouped frequency distribution, there are classes along with their respective frequencies. Let l_i be the lower limit and u_i be the upper limit of *i* th class. Further, let the number of classes be *n*, so that *i* = 1, 2, ... *n*. Also let f_i be the frequency of *i* th class. This distribution can written in tabular form, as shown.

Caution Here *u*1 may or may not be equal to *l*2, i.e., the upper limit of a class may or may not be equal to the lower limit of its following class.

It may be recalled here that, in a grouped frequency distribution, we only know the number of observations in a particular class interval and not their individual magnitudes. Therefore, to calculate mean, we have to make a fundamental assumption that the observations in a class are uniformly distributed. Under this assumption, the mid-value of

Notes

a class will be equal to the mean of observations in that class and hence can be taken as their representative. Therefore, if Xi is the mid-value of ith class with frequency fi, the above assumption implies that there are fi observations each with magnitude Xi (i = 1 to n). Thus, the arithmetic mean of a grouped frequency distribution can also be calculated by the use of the formula.

Class Intervals	Frequency (f)
$l_1 - u_1$ $l_2 - u_2$ .... $l_n - u_n$	f_1 f_2 .... f_n
Total Frequency	$\Sigma f_i = N$

Notes The accuracy of arithmetic mean calculated for a grouped frequency distribution depends upon the validity of the fundamental assumption. This assumption is rarely met in practice. Therefore, we can only get an approximate value of the arithmetic mean of a grouped frequency distribution.

Example: The following table gives the distribution of weekly wages of workers in a factory. Calculate the arithmetic mean of the distribution.

Weekly Wages	240-269	270-299	300-329	330-359	360-389	390-419	420-449
No. of Workers	7	19	27	15	12	12	8

Solution:

It may be noted here that the given class intervals are inclusive. However, for the computation of mean, they need not be converted into exclusive class intervals.

Class Intervals	Mid-values (X)	Frequency	$d = X - 344.5$	f_d
240 - 269	254.5	7	– 90	– 630
270 - 299	284.5	19	– 60	– 1140
300 - 329	314.5	27	– 30	– 810
330 - 359	344.5	15	– 0	0
360 - 389	374.5	12	30	360
390 - 419	404.5	12	60	720
420 - 449	434.5	8	90	720
	Total	**100**		**– 780**

$$\bar{X} = A + \frac{\sum fd}{N} = 344.5 - \frac{780}{100} = 336.7$$

Step Deviation Method or Coding Method

In a grouped frequency distribution, if all the classes are of equal width, say 'h', the successive mid-values of various classes will differ from each other by this width. This fact can be utilised for reducing the work of computations.

Notes

Let us define $u_i = \frac{X_i - A}{h}$. Multiplying both sides by f_i and taking sum over all the observations we have, $\sum_{i=1}^{n} f_i u_i = \frac{1}{h}\sum_{i=1}^{n} f_i (X_i - A)$

or $$h\sum_{i=1}^{n} f_i u_i = \sum_{i=1}^{n} f_i X_i - A\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_i X_i - A.N$$

Dividing both sides by N, we have

$$h \cdot \frac{\sum_{i=1}^{n} f_i u_i}{N} = \frac{\sum_{i=1}^{n} f_i X_i}{N} - A = \bar{X} - A$$

$$\therefore \quad \bar{X} = A + h \cdot \frac{\sum_{i=1}^{n} f_i u_i}{N} \qquad ...(2)$$

Using this relation we can simplify the computations of Example, as shown below.

$u = \frac{X - 344.5}{30}$	– 3	– 2	– 1	0	1	2	3	**Total**
f	7	19	27	15	12	12	8	**100**
fu	– 21	– 38	– 27	0	12	24	24	**– 26**

Using formula (2), we have

$$\bar{X} = 344.5 - \frac{30 \times 26}{100} = 336.7$$

Did u know? **What is Charlier's check of accuracy?**

When the arithmetic mean of a frequency distribution is calculated by shortcut or step-deviation method, the accuracy of the calculations can be checked by using the following formulae, given by Charlier.

For shortcut method

$$\Sigma f_i(d_i + 1) = \Sigma f_i d_i + \Sigma f_i$$

or $$\Sigma f_i d_i = \Sigma f_i(d_i + 1) - \Sigma f_i = \Sigma f_i(d_i + 1) - N$$

Similarly, for step-deviation method

$$\Sigma f_i(u_i + 1) = \Sigma f_i u_i + \Sigma f_i$$

or $$\Sigma f_i u_i = \Sigma f_i(u_i + 1) - \Sigma f_i = \Sigma f_i(u_i + 1) - N$$

8.2.2 Weighted Arithmetic Mean

In the computation of simple arithmetic mean, equal importance is given to all the items. But this may not be so in all situations. If all the items are not of equal importance, then simple arithmetic mean will not be a good representative of the given data. Hence, weighing of different items becomes necessary. The weights are assigned to different items depending upon their importance, i.e., more important items are assigned more weight. For example, to calculate

Notes

mean wage of the workers of a factory, it would be wrong to compute simple arithmetic mean if there are a few workers (say managers) with very high wages while majority of the workers are at low level of wages. The simple arithmetic mean, in such a situation, will give a higher value that cannot be regarded as representative wage for the group. In order that the mean wage gives a realistic picture of the distribution, the wages of managers should be given less importance in its computation. The mean calculated in this manner is called weighted arithmetic mean. The computation of weighted arithmetic is useful in many situations where different items are of unequal importance, e.g., the construction index numbers, computation of standardised death and birth rates, etc.

Formulae for Weighted Arithmetic Mean

Let X_1, X_2 ....., X_n be n values with their respective weights w_1, w_2 ....., w_n. Their weighted arithmetic mean denoted as is given by,

1. $\bar{X}_w = \frac{\sum w_i X_i}{\sum w_i}$ (Using direct method),

2. $\bar{X}_w = A + \frac{\sum w_i d_i}{\sum w_i}$ (where $di = Xi - A$) (Using shortcut method),

3. $\bar{X}_w = A + \frac{\sum w_i u_i}{\sum w_i} \times h \quad \left(\text{where } u_i = \frac{X_i - A}{h}\right)$ (Using step-deviation method)

Example: From the following results of two colleges A and B, find out which of the two is better:

Examination	College A		College B	
	Appeared	Passed	Appeared	Passed
M.Sc.	60	40	200	160
M.A.	100	60	240	200
B.Sc.	200	150	200	140
B.A.	120	75	160	100

Solution:

Performance of the two colleges can be compared by taking weighted arithmetic mean of the pass percentage in various classes. The calculation of weighted arithmetic mean is shown in the following table.

	College A				College B			
Class	Appeared w_A	Passed	Pass Percentage X_A	$w_A X_A$	Appeared w_B	Passed	Pass Percentage X_B	$w_B X_B$
M.Sc.	60	40	66.67	4000.2	200	160	80.00	16000.0
M.A.	100	60	60.00	6000.0	240	200	83.33	19999.2
B.Sc.	200	150	75.00	15000.0	200	140	70.00	14000.0
B.A.	120	75	62.50	7500.0	160	100	62.50	10000.0
Total	**480**	**325**		**32500.2**	**800**	**600**		**59999.2**

Notes

$$\bar{X}_w \text{ for College A } = \frac{\sum w_A X_A}{\sum w_A} = \frac{32500.2}{480} = 67.71\%$$

$$\bar{X}_w \text{ for College B } = \frac{\sum w_B X_B}{\sum w_B} = \frac{59999.2}{800} = 75\%$$

Since the weighted average of pass percentage is higher for college B, hence college B is better.

Notes If $\bar{X}$ denotes simple mean and $\bar{X}_w$ denotes the weighted mean of the same data, then

1. $\bar{X} = \bar{X}_w$, when equal weights are assigned to all the items.
2. $\bar{X} > \bar{X}_w$, when items of small magnitude are assigned greater weights and items of large magnitude are assigned lesser weights.
3. $\bar{X} < \bar{X}_w$, when items of small magnitude are assigned lesser weights and items of large magnitude are assigned greater weights.

Task Analyse the properties of Arithmetic Mean.

8.2.3 Median

Median of distribution is that value of the variate which divides it into two equal parts. In terms of frequency curve, the ordinate drawn at median divides the area under the curve into two equal parts. Median is a positional average because its value depends upon the position of an item and not on its magnitude.

Median can be determined under various situations like:

When Individual Observations are Given

The following steps are involved in the determination of median:

1. The given observations are arranged in either ascending or descending order of magnitude.
2. Given that there are n observations, the median is given by:

   (a) The size of $\left(\frac{n+1}{2}\right)$ th observations, when n is odd.

   (b) The mean of the sizes of $\frac{n}{2}$ th and $\left(\frac{n}{2}+1\right)$ th observations, when n is even.

Example: Find median of the following observations:

20, 15, 25, 28, 18, 16, 30.

Solution:

Writing the observations in ascending order, we get 15, 16, 18, 20, 25, 28, 30.

Notes

Since n = 7, i.e., odd, the median is the size of $\left(\frac{7+1}{2}\right)$, i.e., 4th observation.

Hence, median, denoted by Md = 20.

Notes The same value of Md will be obtained by arranging the observations in descending order of magnitude.

Task Find median of data: 245, 230, 265, 236, 220, 250.

When Ungrouped Frequency Distribution is Given

In this case, the data are already arranged in the order of magnitude. Here, cumulative frequency is computed and the median is determined in a manner similar to that of individual observations.

Example: Locate median of the following frequency distribution:

X	0	1	2	3	4	5	6	7
f	7	14	18	36	51	54	52	20

Solution:

X	0	1	2	3	4	5	6	7
f	7	14	18	36	51	54	52	20
c.f.	7	21	39	75	126	180	232	252

Here N = 252, i.e., even.

Now $\frac{N}{2} = \frac{252}{2} = 126$ and $\frac{N}{2} + 1 = 127.$

Therefore, Median is the mean of the size of 126th and 127th observation. From the table we note that 126th observation is 4 and 127th observation is 5.

$$M_d = \frac{4+5}{2} = 4.5$$

Alternative Method: Looking at the frequency distribution we note that there are 126 observations which are less than or equal to 4 and there are 252 – 75 = 177 observations which are greater than or equal to 4. Similarly, observation 5 also satisfies this criterion. Therefore, median = $\frac{4+5}{2} = 4.5.$

When Grouped Frequency Distribution is Given (Interpolation formula)

The determination of median, in this case, will be explained with the help of the following example.

Example: Suppose we wish to find the median of the following frequency distribution.

Class Intervals	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
Frequency	5	12	14	18	13	8

Notes

Solution:

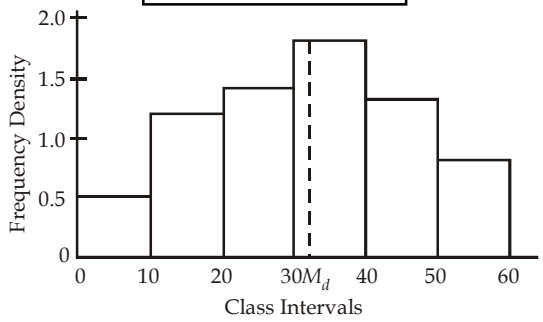
The median of a distribution is that value of the variate which divides the distribution into two equal parts. In case of a grouped frequency distribution, this implies that the ordinate drawn at the median divides the area under the histogram into two equal parts. Writing the given data in a tabular form, we have:

Class Intervals (1)	Frequency (f) (2)	'Less than' type c.f. (3)	Frequency Density (4)
0 - 10	5	5	0.5
10 - 20	12	17	1.2
20 - 30	14	31	1.4
30 - 40	18	49	1.8
40 - 50	13	62	1.3
50 - 60	8	70	0.8

$\left(\textit{Note :}\ \text{Frequency density in a class} = \dfrac{\text{Frequency of the class}}{\text{Width of the class}} = \dfrac{f}{h}\right)$

For the location of median, we make a histogram with heights of different rectangles equal to frequency density of the corresponding class. Such a histogram is shown below:

Figure 8.1: Histogram



Since the ordinate at median divides the total area under the histogram into two equal parts, therefore we have to find a point (like Md as shown in the Figure) on X-axis such that an ordinate (AMd) drawn at it divides the total area under the histogram into two equal parts.

We may note here that area under each rectangle is equal to the frequency of the corresponding class.

Since area = Length × Breadth = Frequency density × Width of class = $\dfrac{f}{h} \times h = f$.

Thus, the total area under the histogram is equal to total frequency N. In the given example N = 70, therefore $\dfrac{N}{2} = 35.$ We note that area of first three rectangles is 5 + 12 + 14 = 31 and the area of first four rectangles is 5 + 12 + 14 + 18 = 49. Thus, median lies in the fourth class interval which is also termed as median class. Let the point, in median class, at which median lies be denoted by Md. The position of this point should be such that the ordinate AMd (in the above histogram) divides the area of median rectangle so that there are only 35 – 31 = 4 observations to its left. From the histogram, we can also say that the position of Md should be such that

$$\frac{M_d - 30}{40 - 30} = \frac{4}{18} \qquad \ldots(1)$$

Notes

Thus, $$M_d = \frac{40}{18} + 30 = 32.2$$

Writing the above equation in general notations, we have

$$\frac{M_d - L_m}{h} = \frac{\frac{N}{2} - C}{f_m} \text{ or } M_d = L_m + \left(\frac{\frac{N}{2} - C}{f_m} \right) h \quad ...(2)$$

Where, *Lm* is lower limit, *h* is the width and *fm* is frequency of the median class and *C* is the cumulative frequency of classes preceding median class. Equation (2) gives the required formula for the computation of median.

Remarks:

1. Since the variable, in a grouped frequency distribution, is assumed to be continuous we always take exact value of including figures after decimals, when N is odd.
2. The above formula is also applicable when classes are of unequal width.
3. Median can be computed even if there are open end classes because here we need to know only the frequencies of classes preceding or following the median class.

Determination of Median When 'greater than' type Cumulative Frequencies are G

By looking at the histogram, we note that one has to find a point denoted by Md such that area to the right of the ordinate at Md is 35. The area of the last two rectangles is 13 + 8 = 21. Therefore, we have to get 35 – 21 = 14 units of area from the median rectangle towards right of the ordinate. Let *Um* be the upper limit of the median class. Then the formula for median in this case can be written as

$$\frac{U_m - M_d}{h} = \frac{\frac{N}{2} - C}{f_m}$$

or $$M_d = U_m - \frac{\frac{N}{2} - C}{f_m} \times h \quad ...(3)$$

Note that C denotes the 'greater than type' cumulative frequency of classes following the median class. Applying this formula to the above example, we get

$$M_d = 40 - \frac{(35 - 21)}{18} \times 10 = 32.2$$

Example: The following table gives the distribution of marks by 500 students in an examination. Obtain median of the given data.

Marks	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69	70 - 79
No. of Students	30	40	50	48	24	162	132	14

Solution:

Since the class intervals are inclusive, therefore, it is necessary to convert them into class boundaries.

Notes

Class Intervals	Class Boundries	Frequency	'Less than' type c.f.
0 - 9	– 0.5 - 9.5	30	30
10 - 19	9.5 - 19.5	40	70
20 - 29	19.5 - 29.5	50	120
30 - 39	29.5 - 39.5	48	168
40 - 49	39.5 - 49.5	24	192
50 - 59	49.5 - 59.5	162	354
60 - 69	59.5 - 69.5	132	486
70 - 79	69.5 - 79.5	14	500

Since $\frac{N}{2} = 250$, the median class is 49.5 – 59.5 and, therefore, $Lm = 49.5$, $h = 10$, $fm = 162$, $C = 192$.

Thus, $M_d = 49.5 + \frac{250 - 192}{162} \times 10 = 53.08$ marks.

Determination of Missing Frequencies

If the frequencies of some classes are missing, however, the median of the distribution is known, then these frequencies can be determined by the use of median formula.

Example: The following table gives the distribution of daily wages of 900 workers. However the frequencies of the classes 40-50 and 60-70 are missing. If the median of the distribution is ₹ 59.25, find the missing frequencies.

Wages (Rs.)	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80
No. of Workers	120	?	200	?	185

Solution:

Let f_1 and f_2 be the frequencies of the classes 40 - 50 and 60 - 70 respectively.

Class Intervals	Frequency	C.f. (less than)
30 - 40	120	120
40 - 50	f_1	$120 + f_1$
50 - 60	200	$320 + f_1$
60 - 70	f_2	$320 + f_1 + f_2$
70 - 80	185	900

Since median is given as 59.25, the median class is 50 - 60.

Therefore, we can write

$$59.25 = 50 + \frac{450 - (120 + f_1)}{200} \times 10 = 50 + \frac{330 - f_1}{20}$$

or $9.25 \times 20 = 330 - f_1$ or $f_1 = 330 - 185 = 145$

Further, $f_2 = 900 - (120 + 145 + 200 + 185) = 250$.

Notes

Graphical Location of Median

So far we have calculated median by the use of a formula. Alternatively, it can be determined graphically, as illustrated in the following example.

Example: The following table shows the daily sales of 230 footpath sellers of Chandni Chowk:

Sales (in Rs.)	0 - 500	500 - 1000	1000 - 1500	1500 - 2000	2000 - 2500	2500 - 3000	3000 - 3500	3500 - 4000
No. of Sellers	12	18	35	42	50	45	20	8

Locate the median of the above data using

1. Only the less than type ogive, and
2. Both, the less than and the greater than type ogives.

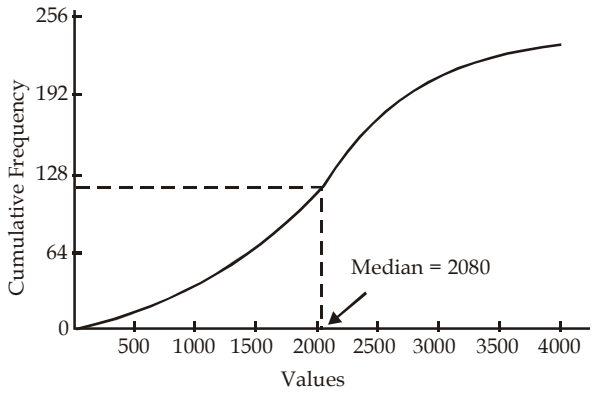
Solution:

To draw ogives, we need to have a cumulative frequency distribution.

Class Intervals	Frequency	Less than c.f.	More than c.f.
0 - 500	12	12	230
500 - 1000	18	30	218
1000 - 1500	35	65	200
1500 - 2000	42	107	165
2000 - 2500	50	157	123
2500 - 3000	45	202	73
3000 - 3500	20	222	28
3500 - 4000	8	230	8

1. ***Using the less than type ogive***

Figure 8.2

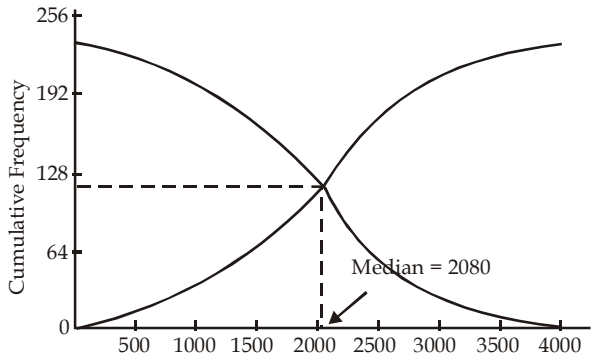


Notes

The value $\frac{N}{2} = 115$ is marked on the vertical axis and a horizontal line is drawn from this point to meet the ogive at point S. Drop a perpendicular from S. The point at which this meets X-axis is the median.

2. ***Using both types of ogives***

Figure 8.3



A perpendicular is dropped from the point of intersection of the two ogives. The point at which it intersects the X-axis gives median. It is obvious from Figure 8.2 and 8.3 that median = 2080.

Properties of Median

1. It is a positional average.
2. It can be shown that the sum of absolute deviations is minimum when taken from median. This property implies that median is centrally located.

8.2.4 Other Partition or Positional Measures

Median of a distribution divides it into two equal parts. It is also possible to divide it into more than two equal parts. The values that divide a distribution into more than two equal parts are commonly known as partition values or fractiles. Some important partition values are discussed in the following sections.

Quartiles

The values of a variable that divide a distribution into four equal parts are called quartiles. Since three values are needed to divide a distribution into four parts, there are three quartiles, viz. Q_1, Q_2 and Q_3, known as the first, second and the third quartile respectively.

For a discrete distribution, the first quartile (Q_1) is defined as that value of the variate such that at least 25% of the observations are less than or equal to it and at least 75% of the observations are greater than or equal to it.

For a continuous or grouped frequency distribution, Q_1 is that value of the variate such that the area under the histogram to the left of the ordinate at Q_1 is 25% and the area to its right is 75%. After locating the first quartile class, the formula for Q_1 is as follows:

Notes

$$Q_1 = L_{Q_1} + \frac{\left(\frac{N}{4} - C\right)}{f_{Q_1}} \times h$$

Here, L_{Q_1} is lower limit of the first quartile class, h is its width, f_{Q_1} is its frequency and C is cumulative frequency of classes preceding the first quartile class.

By definition, the second quartile is median of the distribution. The third quartile (Q_3) of a distribution can also be defined in a similar manner.

For a discrete distribution, Q_3 is that value of the variate such that at least 75% of the observations are less than or equal to it and at least 25% of the observations are greater than or equal to it.

For a grouped frequency distribution, Q_3 is that value of the variate such that area under the histogram to the left of the ordinate at Q_3 is 75% and the area to its right is 25%. The formula for computation of Q_3 can be written as

$Q_3 = L_{Q_3} + \frac{\left(\frac{3N}{4} - C\right)}{f_{Q_3}} \times h$, where the symbols have their usual meaning.

Deciles

Deciles divide a distribution into 10 equal parts and there are, in all, 9 deciles denoted as D_1, D_2, ... D_9 respectively.

For a discrete distribution, the ith decile D_i is that value of the variate such that at least (10i)% of the observation are less than or equal to it and at least (100 – 10i)% of the observations are greater than or equal to it (i = 1, 2, ... 9).

For a continuous or grouped frequency distribution, D_i is that value of the variate such that the area under the histogram to the left of the ordinate at D_i is (10i)% and the area to its right is (100 - 10i)%. The formula for the ith decile can be written as

$$D_i = L_{D_i} + \frac{\left(\frac{iN}{10} - C\right)}{f_{D_i}} \times h \quad (i = 1, 2, \ldots\ldots 9)$$

Percentiles

Percentiles divide a distribution into 100 equal parts and there are, in all, 99 percentiles denoted as P_1, P_2, ... P_{25}, ... P_{40}, ... P_{60}, ... P_{99}, respectively.

For a discrete distribution, the kth percentile P_k is that value of the variate such that at least k% of the observations are less than or equal to it and at least (100 – k)% of the observations are greater than or equal to it.

For a grouped frequency distribution, P_k is that value of the variate such that the area under the histogram to the left of the ordinate at P_k is k% and the area to its right is (100 – k)%. The formula for the kth percentile can be written as

$$P_k = L_{P_k} + \frac{\left(\frac{kN}{100} - C\right)}{f_{P_k}} \times h, \quad (k = 1, 2, \ldots\ldots 99)$$

Notes

Remarks:

1. We may note here that $P_{25} = Q_1$, $P_{50} = D_5 = Q_2 = M_d$, $P_{75} = Q_3$, $P_{10} = D_1$, $P_{20} = D_2$, etc.
2. In continuation of the above, the partition values are known as Quintiles (Octiles) if a distribution is divided in to 5 (8) equal parts.
3. The formulae for various partition values of a grouped frequency distribution, given so far, are based on 'less than' type cumulative frequencies. The corresponding formulae based on 'greater than' type cumulative frequencies can be written in a similar manner, as given below:

$$Q_1 = U_{Q_1} - \frac{\left(\frac{3N}{4} - C\right)}{f_{Q_1}} \times h \quad Q_3 = U_{Q_3} - \frac{\left(\frac{N}{4} - C\right)}{f_{Q_3}} \times h$$

$$D_i = U_{D_i} - \frac{\left[\left(N - \frac{iN}{10}\right) - C\right]}{f_{D_i}} \times h \quad P_k = U_{P_k} - \frac{\left[\left(N - \frac{kN}{100}\right) - C\right]}{f_{P_k}} \times h$$

Here U_{Q_1}, U_{Q_3}, U_{D_i}, U_{P_k} are the upper limits of the corresponding classes and C denotes the greater than type cumulative frequencies.

8.2.5 Mode

Mode is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed. In the words of Croxton and Cowden, "The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded the most typical of a series of values." Further, according to A.M. Tuttle, "Mode is the value which has the greatest frequency density in its immediate neighbourhood."

If the frequency distribution is regular, then mode is determined by the value corresponding to maximum frequency. There may be a situation where concentration of observations around a value having maximum frequency is less than the concentration of observations around some other value. In such a situation, mode cannot be determined by the use of maximum frequency criterion. Further, there may be concentration of observations around more than one value of the variable and, accordingly, the distribution is said to be bimodal or multi-modal depending upon whether it is around two or more than two values.

The concept of mode, as a measure of central tendency, is preferable to mean and median when it is desired to know the most typical value, e.g., the most common size of shoes, the most common size of a ready-made garment, the most common size of income, the most common size of pocket expenditure of a college student, the most common size of a family in a locality, the most common duration of cure of viral-fever, the most popular candidate in an election, etc.

Mode can be determined under following situations like:

When Data are either in the form of Individual Observations or in the form of Ungrouped Frequency Distribution

Given individual observations, these are first transformed into an ungrouped frequency distribution. The mode of an ungrouped frequency distribution can be determined in two ways, as given below:

1. By inspection or
2. By method of grouping.

Notes

1. ***By inspection:*** When a frequency distribution is fairly regular, then mode is often determined by inspection. It is that value of the variate for which frequency is maximum. By a fairly regular frequency distribution we mean that as the values of the variable increase the corresponding frequencies of these values first increase in a gradual manner and reach a peak at certain value and, finally, start declining gradually in, approximately, the same manner as in case of increase.

Example: Compute mode of the following data:

3, 4, 5, 10, 15, 3, 6, 7, 9, 12, 10, 16, 18,

20, 10, 9, 8, 19, 11, 14, 10, 13, 17, 9, 11

Solution:

Writing this in the form of a frequency distribution, we get

Values	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequency	2	1	1	1	1	1	3	4	2	1	1	1	1	1	1	1	1	1

$\therefore$ Mode = 10

Notes

1. If the frequency of each possible value of the variable is same, there is no mode.
2. If there are two values having maximum frequency, the distribution is said to be bimodal.

Example: Determine the mode of the following distribution:

X	10	11	12	13	14	15	16	17	18	19
f	8	15	20	100	98	95	90	75	50	30

Solution:

This distribution is not regular because there is sudden increase in frequency from 20 to 100. Therefore, mode cannot be located by inspection and hence the method of grouping is used. Various steps involved in this method are as follows:

1. Prepare a table consisting of 6 columns in addition to a column for various values of X.
2. In the first column, write the frequencies against various values of X as given in the question.
3. In second column, the sum of frequencies, starting from the top and grouped in twos, are written.
4. In third column, the sum of frequencies, starting from the second and grouped in twos, are written.
5. In fourth column, the sum of frequencies, starting from the top and grouped in threes are written.
6. In fifth column, the sum of frequencies, starting from the second and grouped in threes are written.
7. In the sixth column, the sum of frequencies, starting from the third and grouped in threes are written.

Notes

The highest frequency total in each of the six columns is identified and analysed to determine mode. We apply this method for determining mode of the above example.

X	*f* (1)	(2)	(3)	(4)	(5)	(6)
10	8	23	35	43	135	218
11	15	120	198	293	283	260
12	20	193	185	215	155	
13	(100)	165	125			
14	98	80				
15	95					
16	90					
17	75					
18	50					
19	30					

Analysis Table

V	*A*	*R*	*I*	*A*	*B*	*L*	*E*			
Columns	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
1				1						
2					1	1				
3				1	1					
4				1	1	1				
5					1	1	1			
6						1	1	1		
Total	0	0	0	3	4	4	2	1	0	0

Since the value 14 and 15 are both repeated maximum number of times in the analysis table, therefore, mode is ill defined. Mode in this case can be approximately located by the use of the following formula, which will be discussed later, in this unit.

Mode = 3 Median – 2 Mean

Calculation of Median and Mean

X	10	11	12	13	14	15	16	17	18	19	**Total**
f	8	15	20	100	98	95	90	75	50	30	**581**
C.f.	8	23	43	143	241	336	426	501	551	581	
fX	80	165	240	1300	1372	1425	1440	1275	900	570	**8767**

Median = Size of $\left(\frac{581+1}{2}\right)$th, i.e., 291st observation = 15. Mean = $\frac{8767}{581} = 15.09$

$\therefore$ Mode = 3 × 15 – 2 × 15.09 = 45 – 30.18 = 14.82

Notes If the most repeated values, in the above analysis table, were not adjacent, the distribution would have been bimodal, i.e., having two modes

Notes

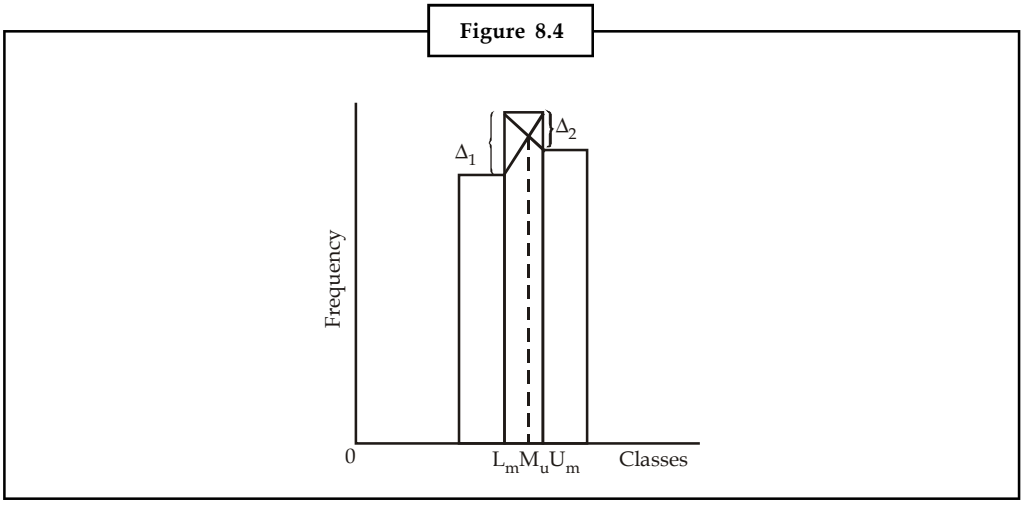
When Data are in the form of a Grouped Frequency Distribution

The following steps are involved in the computation of mode from a grouped frequency distribution.

1. ***Determination of modal class:*** It is the class in which mode of the distribution lies. If the distribution is regular, the modal class can be determined by inspection, otherwise, by method of grouping.

2. ***Exact location of mode in a modal class (interpolation formula):*** The exact location of mode, in a modal class, will depend upon the frequencies of the classes immediately preceding and following it. If these frequencies are equal, the mode would lie at the middle of the modal class interval. However, the position of mode would be to the left or to the right of the middle point depending upon whether the frequency of preceding class is greater or less than the frequency of the class following it. The exact location of mode can be done by the use of interpolation formula, developed below:

Figure 8.4



Let the modal class be denoted by $L_m - U_m$, where L_m and U_m denote its lower and the upper limits respectively. Further, let *fm* be its frequency and h its width. Also let f_1 and f_2 be the respective frequencies of the immediately preceding and following classes.

We assume that the width of all the class intervals of the distribution are equal. If these are not equal, make them so by regrouping under the assumption that frequencies in a class are uniformly distributed.

Make a histogram of the frequency distribution with height of each rectangle equal to the frequency of the corresponding class. Only three rectangles, out of the complete histogram, that are necessary for the purpose are shown in the above Figure.

Let $\Delta_1 = f_m - f_1$ and $\Delta_2 = f_m - f_2$. Then the mode, denoted by M_o, will divide the modal class interval in the ratio $\frac{\Delta_1}{\Delta_2}$.

To derive a formula for mode, the point M_o in the Figure, should be such that

$$\frac{M_o - L_m}{U_m - M_o} = \frac{\Delta_1}{\Delta_2} \text{ or } M_o\Delta_2 - L_m\Delta_2 = U_m\Delta_1 - M_o\Delta_1$$

$\Rightarrow$ $(\Delta_1 + \Delta_2)M_o = L_m\Delta_2 + U_m\Delta_1 = L_m\Delta_2 + (L_m + h)\Delta_1$ (where $U_m = L_m + h$)

$= (\Delta_1 + \Delta_2) L_m + \Delta_1 h$

Notes

Dividing both sides by $\Delta_1 + \Delta_2$, we have

$$M_o = L_m + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h \quad ...(1)$$

By slight adjustment, the above formula can also be written in terms of the upper limit (U_m) of the modal class.

$$M_o = U_m - h + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h = U_m - \left[1 - \frac{\Delta_1}{\Delta_1 + \Delta_2}\right] \times h$$

$$= U_m - \left[\frac{\Delta_2}{\Delta_1 + \Delta_2} \times h\right] \quad ...(2)$$

Replacing Δ_1 by $f_m - f_1$ and Δ_2 by $f_m - f_2$, the above equations can be written as

$$M_o = L_m + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h \quad ...(3)$$

and

$$M_o = U_m - \frac{f_m - f_2}{2f_m - f_1 - f_2} \times h \quad ...(4)$$

Notes The above formulae are applicable only to a unimodal frequency distribution.

8.2.6 Relation between Mean, Median and Mode

The relationship between the above measures of central tendency will be interpreted in terms of a continuous frequency curve.

If the number of observations of a frequency distribution are increased gradually, then accordingly, we need to have more number of classes, for approximately the same range of values of the variable, and simultaneously, the width of the corresponding classes would decrease. Consequently, the histogram of the frequency distribution will get transformed into a smooth frequency curve, as shown in Figure 8.5.

For a given distribution, the mean is the value of the variable which is the point of balance or centre of gravity of the distribution. The median is the value such that half of the observations are below it and remaining half are above it. In terms of the frequency curve, the total area under the curve is divided into two equal parts by the ordinate at median. Mode of a distribution is a value around which there is maximum concentration of observations and is given by the point at which peak of the curve occurs.

For a symmetrical distribution, all the three measures of central tendency are equal i.e. $\overline{X} = M_d = M_o$, as shown in Figure 8.6.

Notes

Figure 8.5

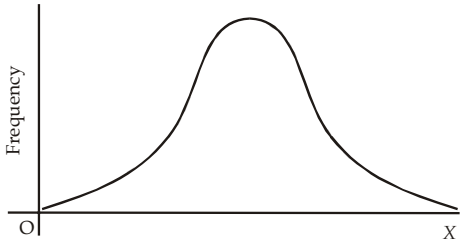
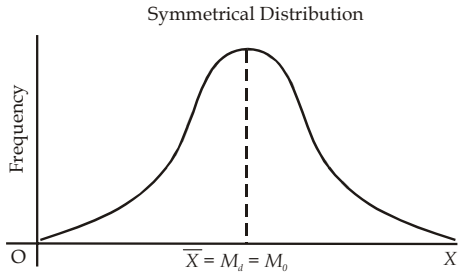
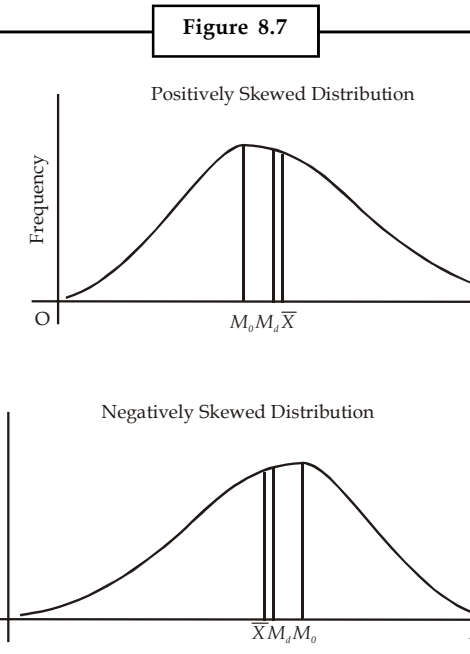


Figure 8.6



Imagine a situation in which the symmetrical distribution is made asymmetrical or positively (or negatively) skewed by adding some observations of very high (or very low) magnitudes, so that the right hand (or the left hand) tail of the frequency curve gets elongated. Consequently, the three measures will depart from each other. Since mean takes into account the magnitudes of observations, it would be highly affected. Further, since the total number of observations will also increase, the median would also be affected but to a lesser extent than mean. Finally, there would be no change in the position of mode. More specifically, we shall have $M_o < M_d < \overline{X}$, when skewness is positive and $\overline{X} < M_d < M_o$, when skewness is negative, as shown in Figure 8.7.

Figure 8.7



Empirical Relation between Mean, Median and Mode

Notes

Empirically, it has been observed that for a moderately skewed distribution, the difference between mean and mode is approximately three times the difference between mean and median, i.e., $\bar{X} - M_o = 3(\bar{X} - M_d)$.

This relation can be used to estimate the value of one of the measures when the values of the other two are known.

Self Assessment

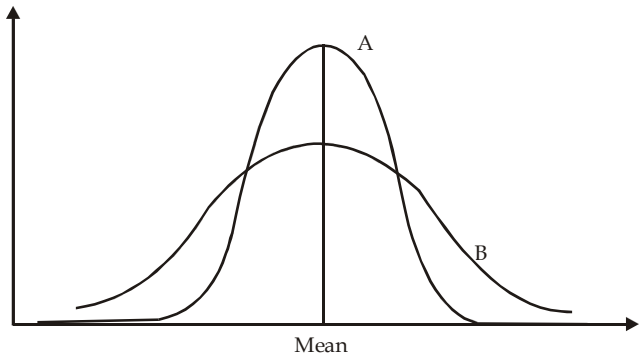
Fill in the blanks:

3. ..... is defined as the sum of observations divided by the number of observations.
4. ..... is used when the magnitude of individual observations is large.
5. Median and mode are also known as the ..... averages.
6. In a grouped frequency distribution, there are classes along with their respective .....
7. ..... check of accuracy is used when the arithmetic mean of a frequency distribution is calculated by shortcut or step-deviation method.
8. Median of distribution is that value of the variate which divides it into ..... parts.
9. The total area under a histogram is equal to total .....
10. ..... divide a distribution into 10 equal parts.
11. A distribution may have only ..... quartiles.
12. Mode is that value of the variate which occurs ..... number of times in a distribution.
13. It is ..... around which other items are most densely distributed.

8.3 Measures of Dispersion

Dispersion is the spread of the data in a distribution. A measure of dispersion:

Figure 8.8



Notes

Indicates the degrees of the scatteredness of the observations. Let curves A and B represent two frequency distributions. Observe that A and B have the same mean. But curve A has less variability than B.

If we measure only the mean of these two distributions, we will miss an important difference between A and B. To increase our understanding of the pattern of the data, we must also measure its dispersion.

Let us understand various measures of dispersion:

1. ***Range:*** It is the difference between the highest and lowest observed values.

   i.e. range = H – L, H = Highest, L = Lowest.

Notes

1. Range is the crudest measure of dispersion.
2. $\frac{H - L}{H + L}$ is called the coefficient of range.

2. ***Semi-inter Quartile Range (Quartile deviation):*** Semi-inter quartile range Q.

   Q is given by $Q = \frac{Q_3 - Q_1}{2}$

Notes

1. $\frac{Q_3 - Q_1}{Q_3 + Q_1}$ is called the coefficient of quartile deviation.
2. Quartile deviation is not a true measure of dispersion but only a distance of scale.

3. ***Mean Deviation (MD):*** If A is any average then mean deviation about A is given by:

   $MD(A) = \frac{\Sigma f_i \mid x_i - A \mid}{N}$

Notes

1. Mean deviation about mean $MD(\bar{x}) = \frac{\Sigma f_i \mid x_i - \bar{x} \mid}{N}$
2. Of all the mean deviations taken about different averages mean derivation about the median is the least.
3. $\frac{MD(A)}{A}$ is called the coefficient of mean deviation.

Notes

4. ***Variance and Standard Deviation:***

Variance (σ^2): A measure of the average squared distance between the mean and each term in the population.

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

Standard deviation (σ) is the positive square root of the variance:

$$\sigma = \sqrt{\frac{1}{N}\Sigma f_i(x_i - \bar{x})^2}$$

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

Notes Combined variance of two sets of data of N_1 and N_2 items with means x_1 and x_2 and standard deviations σ_1 and σ_2 respectively is obtained by:

$$\sigma^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_1d_2^2}{N_1 + N_2}$$

Where $d_1^2 = (x - x_1)^2$ and $d_2^2 = (x - x_2)^2$

and $x = \frac{N_1\overline{x_1} + N_2\overline{x_2}}{N_1 + N_2}$

Sample variance (σ^2) : Let $x_1, x_2, x_3, \ldots\ldots x_n$, represent a sample with mean x.

Then sample variance σ^2 is given by:

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

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

Notes $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2}{n-1} - \frac{n(\bar{x})^2}{n-1}}$ is called the sample standard deviation.

5. ***Coefficient of Variation (C.V):*** It is a relative measure of dispersion that enables us to compare two distributions. It relates the standard deviation and the mean by expressing the standard deviation as a percentage of the mean.

$$\text{C.V.} = \frac{\sigma}{x} \times 100$$

Notes

1. Coefficient of variation is independent of the unit of the observation.
2. This measure cannot be used when x is zero or close to zero.

Notes

Example: For the data 103, 50, 68, 110, 105, 108, 174, 103, 150, 200, 225, 350, 103 find the range, coefficient of range and coefficient of quartile deviation.

Solution:

$$\text{Range} = H - L = 350 - 50 = 300$$

Coefficient of range $\frac{H-L}{H+L} = \frac{300}{350+50} = 0.7$

To find Q_1 and Q_3 we arrange the data in ascending order:

50, 68, 103, 103, 103, 103, 105, 108, 110, 150, 174, 200, 225, 350,

$$\frac{n+1}{4} = \frac{14}{4} = 3.5,$$

$$\frac{3(n+1)}{4} = 10.5$$

$\therefore$ $Q_1 = 103 + 0.5\,(103 - 103) = 103$

$Q_3 = 174 + 0.5\,(200 - 174) = 187$

$$\text{Coefficient of QD} = \frac{84}{290} = 0.2896$$

Self Assessment

Fill in the blanks:

14. ..... is the spread of the data in a distribution.

15. ..... is a measure of the average squared distance between the mean and each term in the population.

8.4 Summary

- Descriptive statistics are used to describe the basic features of the data in a study.
- They provide simple summaries about the sample and the measures.
- Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.
- Descriptive Statistics are used to present quantitative descriptions in a manageable form. In a research study we may have lots of measures.
- When summarizing a quantity like length or weight or age, it is common to answer the first question with the arithmetic mean, the median, or, in case of a unimodal distribution, the mode.
- Sometimes, we choose specific values from the cumulative distribution function called quantiles.
- The most common measures of variability for quantitative data are the variance; its square root, the standard deviation; the range; interquartile range; and the average absolute deviation (average deviation).

Notes

8.5 Keywords

Average: It is a single value which can be taken as representative of the whole distribution.

Descriptive Statistics: Descriptive statistics are used to describe the basic features of the data in a study.

Dispersion: It is the spread of the data in a distribution.

Median: It is that value of the variate which divides it into two equal parts.

Mode: It is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed.

8.6 Review Questions

1. Show that if all observations of a series are added, subtracted, multiplied or divided by a constant b, the mean is also added, subtracted, multiplied or divided by the same constant.
2. Prove that the algebraic sum of deviations of a given set of observations from their mean is zero.
3. Prove that the sum of squared deviations is least when taken from the mean.
4. The heights of 15 students of a class were noted as shown below. Compute arithmetic mean by using (i) Direct Method and (ii) Short-Cut Method.

S. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ht. (cms)	160	167	174	168	166	171	162	182	186	175	178	167	177	162	163

5. Compute arithmetic mean of the following series:

Marks	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
No. of Students	12	18	27	20	17	6

6. Calculate mode from the following data:

Mid-points	1	2	3	4	5	6	7	8
Frequency	5	50	45	30	20	10	15	5

7. Calculate median and mode from the following data:

Size	10 - 20	10 - 30	10 - 40	10 - 50	10 - 60	10 - 70	10 - 80	10 - 90
No. of Students	4	16	56	97	124	137	146	150

8. Distinguish between an absolute measure and relative measure of dispersion. What are the advantages of using the latter?

Notes

9. Calculate mean deviation from median for the following data:

Wages per week	50 - 59	60 - 69	70 - 79	80 - 89	90 - 99	100 - 109	110 - 119
No. of Students	15	40	50	60	45	90	15

10. Calculate the coefficient of mean deviation from mean and median from the following data:

Marks	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 -80	80 -90
No. of Students	2	6	12	18	25	20	10	7

11. Calculate mean deviation from mode of the following data:

(a) 7, 4, 6, 4, 4, 5, 2, 4, 1, 7, 7, 6, 2, 3, 4, 2

(b)

Size	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
Frequency	6	20	44	26	3	1

12. Calculate the standard deviation of the following series:

Size	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
Frequency	10	8	15	8	4

13. Calculate the standard deviation from the following data:

Age less than (in years)	10	20	30	40	50	60	70	80
No. of Persons	15	30	53	75	100	110	115	125

14. 14.Find out standard deviation from the following data:

Mid Value	30	35	40	45	50	55	60	65	70	75	80
Frequency	1	2	4	7	9	13	17	12	7	6	3

Answers: Self Assessment

1. Summarisation
2. averages
3. Arithmetic Mean
4. Shortcut Method
5. Positional
6. Frequencies
7. Charlier's
8. Two equal
9. Frequency
10. Deciles
11. 3
12. Maximum
13. Mode
14. Dispersion
15. Variance

8.7 Further Readings

Notes

Books

Abrams, M.A, *Social Surveys and Social Action*, London: Heinemann, 1951.

Arthur, Maurice, *Philosophy of Scientific Investigation*, Baltimore: John Hopkins University Press, 1943.

R.S. Bhardwaj, *Business Statistics*, Excel Books, New Delhi, 2008.

S.N. Murthy and U. Bhojanna, *Business Research Methods*, Excel Books, 2007.

Notes

Unit 9: Correlation and Regression

CONTENTS

Objectives

After studying this unit, you will be able to:

- Explain the Concept of correlation
- Judge the Scope of correlation analysis
- Define the Rank Correlation
- Discuss the Regression analysis
- Describe the Simple Regression

Notes

Introduction

Once best estimates are chosen, both from a statistical and epidemiologic perspective, hypotheses about the estimated association between a single mean, proportion, or rate and a fixed value, typically standard or goal, or about the estimated association between two or more means, proportions, or rates can be tested.

The measures of association refer to a wide variety of coefficients that measure the strength of the relationship that has been described in several ways. The word 'association' in measures of association measures the strength of association in which there is at least one of the variables that is dichotomous in nature, generally nominal or ordinal. The measures of association define the strength of the linear relationship in terms of the degree of monotonicity. This degree of monotonicity used by the measures of association is based on the counting of various types of pairs in a relationship.

9.1 Correlation

Various experts have defined correlation in their own words and their definitions, broadly speaking, imply that correlation is the degree of association between two or more variables. Some important definitions of correlation are given below:

1. "If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated."

   — *L.R. Connor*

2. "Correlation is an analysis of covariation between two or more variables."

   — *A.M. Tuttle*

3. "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

   — *Croxton and Cowden*

4. "Correlation analysis attempts to determine the 'degree of relationship' between variables".

   — *Ya Lun Chou*

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

The Scope of Correlation Analysis

The existence of correlation between two (or more) variables only implies that these variables (i) either tend to increase or decrease together or (ii) an increase (or decrease) in one is accompanied by the corresponding decrease (or increase) in the other. The questions of the type, whether changes in a variable are due to changes in the other, i.e., whether a cause and effect type relationship exists between them, are not answered by the study of correlation analysis. If there is a correlation between two variables, it may be due to any of the following situations:

1. ***One of the variable may be affecting the other:*** A correlation coefficient calculated from the data on quantity demanded and corresponding price of tea would only reveal that the degree of association between them is very high. It will not give us any idea about whether price is affecting demand of tea or vice-versa. In order to know this, we need to

Notes

have some additional information apart from the study of correlation. For example if, on the basis of some additional information, we say that the price of tea affects its demand, then price will be the cause and quantity will be the effect. The causal variable is also termed as independent variable while the other variable is termed as dependent variable.

2. ***The two variables may act upon each other:*** Cause and effect relation exists in this case also but it may be very difficult to find out which of the two variables is independent.

Example: If we have data on price of wheat and its cost of production, the correlation between them may be very high because higher price of wheat may attract farmers to produce more wheat and more production of wheat may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of wheat.

For the purpose of determining a relationship between the two variables in such situations, we can take any one of them as independent variable.

3. ***The two variables may be acted upon by the outside influences:*** In this case we might get a high value of correlation between the two variables, however, apparently no cause and effect type relation seems to exist between them.

Example: The demands of the two commodities, say X and Y, may be positively correlated because the incomes of the consumers are rising. Coefficient of correlation obtained in such a situation is called a spurious or nonsense correlation.

4. ***A high value of the correlation coefficient may be obtained due to sheer coincidence (or pure chance):*** This is another situation of spurious correlation. Given the data on any two variables, one may obtain a high value of correlation coefficient when in fact they do not have any relationship.

Example: A high value of correlation coefficient may be obtained between the size of shoe and the income of persons of a locality.

9.1.1 Scatter Diagram

Let the bivariate data be denoted by (X_i, Y_i), where $i = 1, 2 \ldots\ldots n$. In order to have some idea about the extent of association between variables X and Y, each pair (X_i, Y_i), $i = 1, 2 \ldots\ldots n$, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

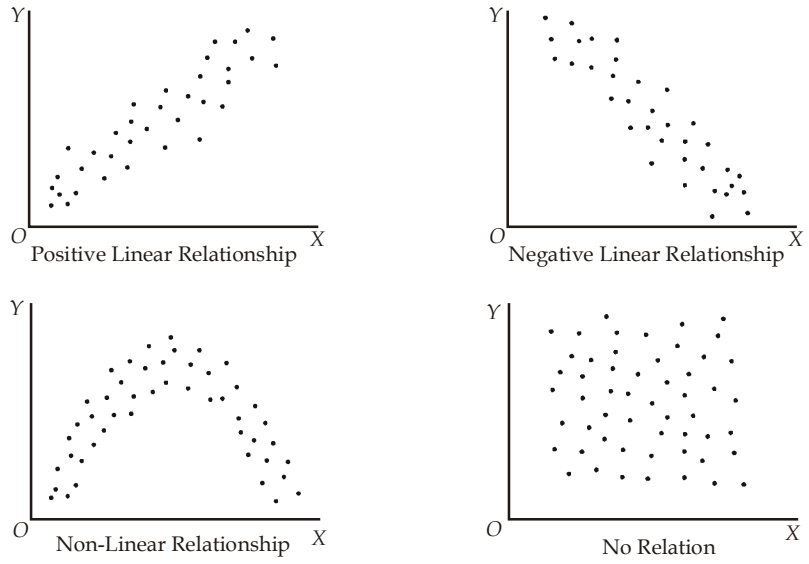
Each pair of values (X_i, Y_i) is denoted by a point on the graph. The set of such points may cluster around a straight line or a curve or may not show any tendency of association. Various possible situations are shown with the help of following diagrams:

Did u know? **What the sets of point in generally known?**

The sets of points in scatter diagram are known as dots of the diagram

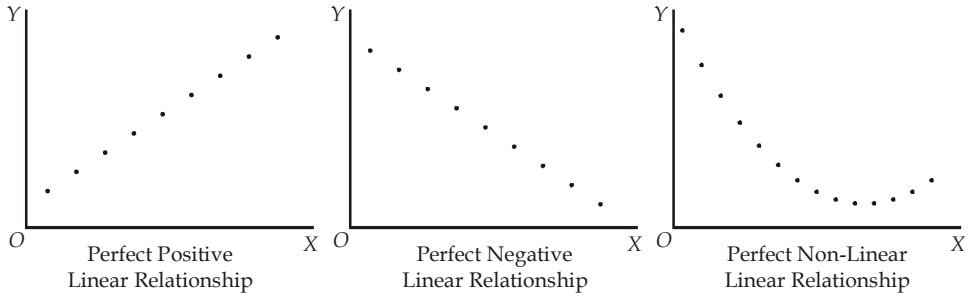
Notes

Figure 9.1



If all the points or dots lie exactly on a straight line or a curve, the association between the variables is said to be perfect. This is shown below:

Figure 9.2



A scatter diagram of the data helps in having a visual idea about the nature of association between two variables. If the points cluster along a straight line, the association between variables is linear. Further, if the points cluster along a curve, the corresponding association is non-linear or curvilinear. Finally, if the points neither cluster along a straight line nor along a curve, there is absence of any association between the variables.

It is also obvious from the above figure that when low (high) values of X are associated with low (high) value of Y, the association between them is said to be positive. Contrary to this, when low (high) values of X are associated with high (low) values of Y, the association between them is said to be negative.

Notes

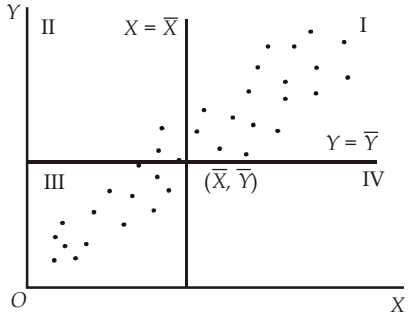
This unit deals only with linear association between the two variables X and Y. We shall measure the degree of linear association by the Karl Pearson's formula for the coefficient of linear correlation.

9.1.2 Karl Pearson's Coefficient of Linear Correlation

Let us assume, again, that we have data on two variables X and Y denoted by the pairs (X_i, Y_i), $i = 1, 2, \ldots\ldots n$. Further, let the scatter diagram of the data be as shown in Figure.

Let $\overline{X}$ and $\overline{Y}$ be the arithmetic means of X and Y respectively. Draw two lines $X = \overline{X}$ and $Y = \overline{Y}$ on the scatter diagram. These two lines, intersect at the point $(\overline{X}, \overline{Y})$ and are mutually perpendicular, divide the whole diagram into four parts, termed as I, II, III and IV quadrants, as shown.

Figure 9.3



As mentioned earlier, the correlation between X and Y will be positive if low (high) values of X are associated with low (high) values of Y. In terms of the above Figure, we can say that when values of X that are greater (less) than $\overline{X}$ are generally associated with values of Y that are greater (less) than $\overline{Y}$, the correlation between X and Y will be positive. This implies that there will be a general tendency of points to concentrate in I and III quadrants. Similarly, when correlation between X and Y is negative, the point of the scatter diagram will have a general tendency to concentrate in II and IV quadrants.

Further, if we consider deviations of values from their means, i.e., $(X_i - \overline{X})$ and $(Y_i - \overline{Y})$, we note that:

1. Both $(X_i - \overline{X})$ and $(Y_i - \overline{Y})$ will be positive for all points in quadrant I.
2. $(X_i - \overline{X})$ will be negative and $(Y_i - \overline{Y})$ will be positive for all points in quadrant II.
3. Both $(X_i - \overline{X})$ and $(Y_i - \overline{Y})$ will be negative for all points in quadrant III.
4. $(X_i - \overline{X})$ will be positive and $(Y_i - \overline{Y})$ will be negative for all points in quadrant IV.

It is obvious from the above that the product of deviations, i.e., $(X_i - \overline{X})(Y_i - \overline{Y})$ will be positive for points in quadrants I and III and negative for points in quadrants II and IV.

Notes Since, for positive correlation, the points will tend to concentrate more in I and III quadrants than in II and IV, the sum of positive products of deviations will outweigh the sum of negative products of deviations. Thus, $\Sigma(X_i - \overline{X})(Y_i - \overline{Y})$ will be positive for all the n observations.

Contd...

Notes

Similarly, when correlation is negative, the points will tend to concentrate more in II and IV quadrants than in I and III. Thus, the sum of negative products of deviations will outweigh the sum of positive products and hence $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$ will be negative for all the n observations.

Further, if there is no correlation, the sum of positive products of deviations will be equal to the sum of negative products of deviations such that $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$ will be equal to zero.

On the basis of the above, we can consider $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$ as an absolute measure of correlation. This measure, like other absolute measures of dispersion, skewness, etc., will depend upon (i) the number of observations and (ii) the units of measurements of the variables.

In order to avoid its dependence on the number of observations, we take its average, i.e., $\frac{1}{n}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$. This term is called covariance in statistics and is denoted as $Cov(X, Y)$.

To eliminate the effect of units of measurement of the variables, the covariance term is divided by the product of the standard deviation of X and the standard deviation of Y. The resulting expression is known as the Karl Pearson's coefficient of linear correlation or the product moment correlation coefficient or simply the coefficient of correlation, between X and Y.

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad ...(1)$$

or

$$r_{XY} = \frac{\frac{1}{n}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\frac{1}{n}\sum\left(X_i - \bar{X}\right)^2}\sqrt{\frac{1}{n}\sum\left(Y_i - \bar{Y}\right)^2}} \quad ...(2)$$

Cancelling $\frac{1}{n}$ from the numerator and the denominator, we get

$$r_{XY} = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum\left(X_i - \bar{X}\right)^2}\sqrt{\sum\left(Y_i - \bar{Y}\right)^2}} \quad ...(3)$$

Consider $\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \sum\left(X_i - \bar{X}\right)Y_i - \bar{Y}\sum\left(X_i - \bar{X}\right)$

$= \Sigma X_i Y_i - \bar{X}\Sigma Y_i$ (second term is zero)

$= \sum X_i Y_i - n\bar{X}\bar{Y}\left(\sum Y_i = n\bar{Y}\right)$

Similarly we can write $\sum\left(X_i - \bar{X}\right)^2 = \sum X_i^2 - n\bar{X}^2$

and $$\sum\left(Y_i - \bar{Y}\right)^2 = \sum Y_i^2 - n\bar{Y}^2$$

Substituting these values in equation (3), we have

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left[\sum X_i^2 - n\bar{X}^2\right]}\sqrt{\left[\sum Y_i^2 - n\bar{Y}^2\right]}} \quad ...(4)$$

Notes

$$r_{XY} = \frac{\sum X_i Y_i - n \cdot \frac{\sum X_i}{n} \times \frac{\sum Y_i}{n}}{\sqrt{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2}\sqrt{\sum Y_i^2 - n\left(\frac{\sum Y_i}{n}\right)^2}}$$

$$= \frac{\sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}}{\sqrt{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}\sqrt{\sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}}} \quad ...(5)$$

On multiplication of numerator and denominator by n, we can write

$$r_{XY} = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{n\sum X_i^2 - \left(\sum X_i\right)^2}\sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}} \quad ...(6)$$

Further, if we assume $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$, equation (2), given above, can be written as

or $$r_{XY} = \frac{\sum x_i y_i}{\sqrt{\frac{1}{n}\sum x_i^2}\sqrt{\frac{1}{n}\sum y_i^2}} \quad ...(7)$$

or $$r_{XY} = \frac{\sum X_i Y_i}{\sqrt{\sum X_i^2}\sqrt{\sum Y_i^2}} \quad ...(8)$$

or $$r_{XY} = \frac{1}{n}\frac{\sum x_i y_i}{\sigma_x \sigma_y} \quad ...(9)$$

Equations (5) or (6) are often used for the calculation of correlation from raw data, while the use of the remaining equations depends upon the forms in which the data are available. For example, if standard deviations of X and Y are given, equation (9) may be appropriate.

Example: Calculate the Karl Pearson's coefficient of correlation from the following pairs of values:

Values of Xi	12	9	8	10	11	13	7
Values of Yi	14	8	6	9	11	12	3

Solution:

The formula for Karl Pearson's coefficient of correlation is

$$\frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{n\sum X_i^2 - \left(\sum X_i\right)^2}\sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}}$$

Notes

The values of different terms, given in the formula, are calculated from the following table:

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
12	14	168	144	196
9	8	72	81	64
8	6	48	64	36
10	9	90	100	81
11	11	121	121	121
13	12	156	169	144
7	3	21	49	9
70	63	676	728	651

Here n = 7 (no. of pairs of observations)

$$r_{XY} = \frac{7 \times 676 - 70 \times 63}{\sqrt{7 \times 728 - (70)^2}\sqrt{7 \times 651 - (63)^2}} = 0.949$$

Example: Calculate the Karl Pearson's coefficient of correlation between X and Y from the following data:

No. of pairs of observations $n = 8$, $\Sigma\left(X_i - \bar{X}\right)^2 = 184$, $\Sigma\left(Y_i - \bar{Y}\right)^2 = 148$,

$\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = 164$, $\bar{X} = 11$ and $\bar{Y} = 10$

Solution:

Using the formula, $r_{XY} = \dfrac{\sum\left(X - \bar{X}_i\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum\left(X_i - \bar{X}\right)^2}\sqrt{\sum\left(Y_i - \bar{Y}\right)^2}}$, we get

$$r_{XY} = \frac{164}{\sqrt{184}\sqrt{148}} = 0.99$$

Example: Calculate the correlation between Reading (X) and Spelling (Y) for the 10 students whose scores are given below:

Student	Reading	Spelling
1	13	11
2	7	1
3	2	19
4	9	5
5	8	17
6	4	3
7	1	15
8	10	9
9	6	15
10	5	8

Notes

Solution:

Student	Reading (X)	Spelling (Y)	$X - \mu_x$	$Y - \mu_y$	$(X - \mu_x)(Y - \mu_y)$
1	3	11	– 2.5	0.7	– 1.75
2	7	1	1.5	– 9.3	– 13.95
3	2	19	– 3.5	8.7	– 30.45
4	9	5	3.5	– 5.3	– 18.55
5	8	17	2.5	6.7	16.75
6	4	3	– 1.5	– 7.3	10.95
7	1	15	– 4.5	4.7	– 21.15
8	10	9	4.5	– 1.3	– 5.85
9	6	15	0.5	4.7	2.35
10	5	8	– 0.5	– 2.3	1.15
Sum	55	103	0.0	0.0	– 60.5
Mean	5.5	10.3			
Standard Deviation	2.872	5.832			

Using the correlation formula;

$$r = \frac{\Sigma(X-\mu_x)(Y-\mu_y)}{N\sigma_x\sigma_y}$$

$$= \frac{-60.5}{(10)(2.872)(5.832)} = \frac{-60.5}{167.495} = -0.36$$

However, in real practice, we use the computational or raw score formula for the correlation coefficient:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Where:

(i) N is the number of subjects

(ii) ΣXY is the sum of each subject X score times the Y score,

(iii) ΣX is the sum of the X scores

(iv) ΣY is the sum of the Y scores

(v) ΣX^2 is the sum of the squared X scores

(vi) ΣY^2 is the sum of the squared Y scores,

Correlation between Reading and Spelling for the data given in example using Computational Formula:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Notes

$$= \frac{(10)(506)-(55)(103)}{\sqrt{(10)(385)-(55)^2}\sqrt{(10)(1401)-(103)^2}}$$

$$= \frac{(5060-5665}{\sqrt{3850-3205}\sqrt{14010-10609}} = \frac{-605}{\sqrt{825}\sqrt{3401}}$$

$$= \frac{-605}{(28.723)(58.318)} = \frac{-605}{1675.0679} = -0.36$$

Thus, the correlation is –.36, indicating that there is a small negative correlation between reading and spelling. The correlation coefficient is a number that can range from 1 (perfect negative correlation) through 0 (no correlation) to 1 (perfect positive correlation).

Task

1. The covariance between the length and weight of five items is 6 and their standard deviations are 2.45 and 2.61 respectively. Find the coefficient of correlation between length and weight.
2. The Karl Pearson's coefficient of correlation and covariance between two variables X and Y is –0.85 and –15 respectively. If variance of Y is 9, find the standard deviation of X.

9.1.3 Properties of Coefficient of Correlation

1. The coefficient of correlation is independent of the change of origin and scale of measurements.

   In order to prove this property, we change origin and scale of both the variables X and Y.

   Let $u_i = \frac{X_i - A}{h}$ and $v_i = \frac{Y_i - B}{k}$, where the constants A and B refer to change of origin and the constants h and k refer to change of scale. We can write

   $X_i = A + hu_i, \quad \therefore \quad \bar{X} = A + h\bar{u}$

   Thus, we have $X_i - \bar{X} = A + hu_i - A - h\bar{u} = h(u_i - \bar{u})$

   Similarly, $Y_i = B + kv_i, \quad \therefore \quad \bar{Y} = B + k\bar{v}$

   Thus, $Y_i - \bar{Y} = B + kv_i - B - k\bar{v} = k(v_i - \bar{v})$

   The formula for the coefficient of correlation between X and Y is

   $$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

   Substituting the values of $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$, we get

   $$r_{XY} = \frac{\sum h(u_i - \bar{u})k(v_i - \bar{v})}{\sqrt{\sum h^2(u_i - \bar{u})^2}\sqrt{\sum k^2(v_i - \bar{v})^2}} = \frac{\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum(u_i - \bar{u})^2}\sqrt{\sum(v_i - \bar{v})^2}}$$

Notes

$\therefore \quad r_{XY} = r_{uv}$

This shows that correlation between X and Y is equal to correlation between u and v, where u and v are the variables obtained by change of origin and scale of the variables X and Y respectively.

This property is very useful in the simplification of computations of correlation. On the basis of this property, we can write a short-cut formula for the computation of r_{XY}:

$$r_{XY} = \frac{n\sum u_i v_i - \left(\sum u_i\right)\left(\sum v_i\right)}{\sqrt{n\sum u_i^2 - \left(\sum u_i\right)^2}\sqrt{n\sum v_i^2 - \left(\sum v_i\right)^2}} \quad ...(10)$$

2. The coefficient of correlation lies between – 1 and + 1.

To prove this property, we define

$$x'_i = \frac{X_i - \bar{X}}{\sigma_X} \text{ and } y'_i = \frac{Y_i - \bar{Y}}{\sigma_Y}$$

$$\therefore \quad x'^2_i = \frac{\left(X_i - \bar{X}\right)^2}{\sigma_X^2} \text{ and } y'^2_i = \frac{\left(Y_i - \bar{Y}\right)^2}{\sigma_Y^2}$$

$$\text{or } \sum x'^2_i = \frac{\sum\left(X_i - \bar{X}\right)^2}{\sigma_X^2} \text{ and } \sum y'^2_i = \frac{\sum\left(Y_i - \bar{Y}\right)^2}{\sigma_Y^2}$$

From these summations we can write $\sum x'^2_i = \sum y'^2_i = n$

$$\text{Also, } r = \frac{\frac{1}{n}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sigma_X \sigma_Y} = \frac{1}{n}\cdot\sum\left(\frac{X_i - \bar{X}}{\sigma_X}\right)\left(\frac{Y_i - Y}{\sigma_Y}\right) = \frac{1}{n}\sum x'_i y'_i$$

Consider the sum $x_i' + y_i'$. The square of this sum is always a non-negative number, i.e., $(x_i' + y_i')^2 \geq 0$.

Taking sum over all the observations and dividing by n, we get

$$\frac{1}{n}\sum\left(x'_i + y'_i\right)^2 \geq 0 \qquad \text{or} \qquad \frac{1}{n}\sum\left(x'^2_i + y'^2_i + 2x'_i y'_i\right) \geq 0$$

$$\text{or} \qquad \frac{1}{n}\sum x'^2_i + \frac{1}{n}\sum y'^2_i + \frac{2}{n}\sum x'_i y'_i \geq 0$$

$$\text{or} \quad 1 + 1 + 2r \geq 0 \text{ or } 2 + 2r \geq 0 \text{ or } r \geq -1 \quad .... (11)$$

Further, consider the difference $x_i' - y_i'$. The square of this difference is also non-negative, i.e., $(x_i' - y_i')^2 \geq 0$.

Taking sum over all the observations and dividing by n, we get

$$\frac{1}{n}\sum\left(x'_i - y'_i\right)^2 \geq 0$$

$$\text{or} \quad \frac{1}{n}\sum\left(x'^2_i + y'^2_i - 2x'_i y'_i\right) \geq 0$$

Notes

or $\frac{1}{n}\sum x_i'^2+\frac{1}{n}\sum y_i'^2-\frac{2}{n}\sum x_i' y_i' \geq 0$

or $1 + 1 - 2r \geq 0$ or $2 - 2r \geq 0$ or $r \leq 1$ .... (12)

Combining the inequalities (11) and (12), we get $-1 \leq r \leq 1$. Hence r lies between – 1 and + 1.

3. If X and Y are independent they are uncorrelated, but the converse is not true.

If X and Y are independent, it implies that they do not reveal any tendency of simultaneous movement either in same or in opposite directions. The dots of the scatter diagram will be uniformly spread in all the four quadrants. Therefore, $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ or Cov(X, Y) will be equal to zero and hence, $r_{XY} = 0$. Thus, if X and Y are independent, they are uncorrelated.

The converse of this property implies that if $r_{XY} = 0$, then X and Y may not necessarily be independent. To prove this, we consider the following data:

X	1	2	3	4	5	6	7
Y	9	4	1	0	1	4	9

Here $\Sigma X_i = 28$, $\Sigma Y_i = 28$ and $\Sigma X_i Y_i = 112$.

$$\therefore \text{Cov}(X,Y) = \frac{1}{n}\left[\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}\right] = \frac{1}{7}\left[112 - \frac{28\times 28}{7}\right] = 0 \text{ Thus, } r_{XY} = 0$$

A close examination of the given data would reveal that although $r_{XY} = 0$, but X and Y are not independent. In fact they are related by the mathematical relation $Y = (X - 4)^2$.

Caution r_{XY} is only a measure of the degree of linear association between X and Y. If the association is non-linear, the computed value of r_{XY} is no longer a measure of the degree of association between the two variables.

9.1.4 Merits and Limitations of Coefficient of Correlation

The only merit of Karl Pearson's coefficient of correlation is that it is the most popular method for expressing the degree and direction of linear association between the two variables in terms of a pure number, independent of units of the variables. This measure, however, suffers from certain limitations, given below:

1. Coefficient of correlation r does not give any idea about the existence of cause and effect relationship between the variables. It is possible that a high value of r is obtained although none of them seem to be directly affecting the other. Hence, any interpretation of r should be done very carefully.
2. It is only a measure of the degree of linear relationship between two variables. If the relationship is not linear, the calculation of r does not have any meaning.
3. Its value is unduly affected by extreme items.

Notes

4. If the data are not uniformly spread in the relevant quadrants the value of r may give a misleading interpretation of the degree of relationship between the two variables. For example, if there are some values having concentration around a point in first quadrant and there is similar type of concentration in third quadrant, the value of r will be very high although there may be no linear relation between the variables.
5. As compared with other methods, to be discussed later in this unit, the computations of r are cumbersome and time consuming.

9.1.5 Probable Error of r

It is an old measure to test the significance of a particular value of r without the knowledge of test of hypothesis. Probable error of r, denoted by P.E.(r) is 0.6745 times its standard error. The value 0.6745 is obtained from the fact that in a normal distribution $\bar{r} \pm 0.6745 \times \text{S.E.}$ covers 50% of the total distribution.

According to Horace Secrist "The probable error of correlation coefficient is an amount which if added to and subtracted from the mean correlation coefficient, gives limits within which the chances are even that a coefficient of correlation from a series selected at random will fall."

Since standard error of r, i.e., $\text{S.E.}_r = \frac{1-r^2}{\sqrt{n}}$, $\therefore \text{P.E.}(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$

Uses of P.E.(r)

1. It can be used to specify the limits of population correlation coefficient r (rho) which are defined as r – P.E.(r) ≤ r ≤ r + P.E.(r), where r denotes correlation coefficient in population and r denotes correlation coefficient in sample.
2. It can be used to test the significance of an observed value of r without the knowledge of test of hypothesis. By convention, the rules are:
   (a) If $|r| < 6$ P.E.(r), then correlation is not significant and this may be treated as a situation of no correlation between the two variables.
   (b) If $|r| > 6$ P.E.(r), then correlation is significant and this implies presence of a strong correlation between the two variables.
   (c) If correlation coefficient is greater than 0.3 and probable error is relatively small, the correlation coefficient should be considered as significant.

Example: Find out correlation between age and playing habit from the following information and also its probable error.

Age	15	16	17	18	19	20
No. of Students	250	200	150	120	100	80
Regular Players	200	150	90	48	30	12

Solution:

Let X denote age, p the number of regular players and q the number of students. Playing habit, denoted by Y, is measured as a percentage of regular players in an age group, i.e., $Y = (p/q) \times 100$.

Notes

Table for Calculation of r

X	q	p	Y	u = X – 17	v = Y – 40	uv	u^2	v^2
15	250	200	80	– 2	40	– 80	4	1600
16	200	150	75	– 1	35	– 35	1	1225
17	150	90	60	0	20	0	0	400
18	120	48	40	1	0	0	1	0
19	100	30	30	2	– 10	– 20	4	100
20	80	12	15	3	– 25	– 75	9	625
Total				**3**	**60**	**– 210**	**19**	**3950**

$$r_{XY} = \frac{-6\times 210 - 3\times 60}{\sqrt{6\times 19 - 9}\sqrt{6\times 3950 - 3600}} = -0.99$$

Probable error of r, i.e., $P.E.(r) = 0.6745\times\frac{\left[1-(0.99)^2\right]}{\sqrt{6}} = 0.0055$

9.1.6 Correlation in a Bivariate Frequency Distribution

Let the two variables X and Y take respective values X_i, i = 1, 2, ... m and Y_j, j = 1, 2, ... n. These values, taken together, will make $m \times n$ pairs (X_i, Y_j). Let f_{ij} be the frequency of this pair. This frequency distribution can be presented in a tabular form as given below:

$Y \rightarrow$ $X \downarrow$	Y_1	Y_2	...	Y_j	...	Y_n	**Total**
X_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1n}	f_1
X_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2n}	f_2
⋮	⋮	⋮		⋮		⋮	⋮
X_i	f_{i1}	f_{i2}		f_{ij}		f_{in}	f_i
⋮	⋮	⋮		⋮		⋮	⋮
X_m	f_{m1}	f_{m2}	...	f_{mj}	...	f_{mn}	f_m
Total	f_1'	f_2'	...	f_j'	...	f_n'	N

Here $\Sigma\Sigma f_{ij} = \Sigma f_i' = \Sigma f_j' = N$ (the total frequency).

The formula for correlation can be written on the basis of the formula discussed earlier.

$$r_{XY} = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f_j'Y_j\right)}{\sqrt{N\sum f_iX_i^2 - \left(\sum f_iX_i\right)^2}\sqrt{N\sum f_j'Y_j^2 - \left(\sum f_j'Y_j\right)^2}}$$

When we make changes of origin and scale by making the transformations $u_i = \frac{X_i - A}{h}$ and $v_j = \frac{Y_j - B}{k}$, then we can write

Notes

$$r_{XY} = \frac{N\sum\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f_j'v_j\right)}{\sqrt{N\sum f_iu_i^2 - \left(\sum f_iu_i\right)^2}\sqrt{N\sum f_j'v_j^2 - \left(\sum f_j'v_j\right)^2}}$$

Example: Calculate Karl Pearson's coefficient of correlation from the following data:

Age (yrs.) → Marks ↓	**18**	**19**	**20**	**21**	**22**
20 - 25	3	2			
15 - 20		5	4		
10 - 15			7	10	
5 - 10				3	2
0 - 5					4

Solution:

Let X_i denote the mid-value of the class interval of marks. Various values of X_i can be written as 22.5, 17.5, 12.5, 7.5 and 2.5.

Further, let $u_i = (X_i - 12.5) \div 5$. Various values of u_i would be 2, 1, 0, – 1 and – 2.

Similarly, let Y_j denote age. Various values of Y_j are 18, 19, 20, 21 and 22.

Assuming $v_j = Y_j - 20$, various values of v_j would be – 2, – 1, 0, 1 and 2.

We shall use the values of u_i and v_j in the computation of *r*.

Table for Calculation of r

u_i \ v_j	–2	–1	0	1	2	f_i	f_iu_i	$f_iu_i^2$	$f_{ij}u_iv_j$
2	3 (–12)	2 (–4)	...	...	...	5	10	20	–16
1	...	5 (–5)	4 (0)	...	...	9	9	9	–5
0	...	...	7 (0)	10 (0)	...	17	0	0	0
–1	...	...	...	3 (–3)	2 (–4)	5	–5	5	–7
–2	...	...	...	...	4 (–16)	4	–8	16	–16
f_j'	3	7	11	13	6	40	6	50	–44
$f_j'v_j$	–6	–7	0	13	12	12			
$f_j'v_j^2$	12	7	0	13	24	56			

Substituting various values in the formula for *r*, we get

$$r = \frac{40\times(-44) - 6\times 12}{\sqrt{40\times 50 - 36}\sqrt{40\times 56 - 144}} = \frac{-1832}{\sqrt{1964}\sqrt{2096}} = -0.903$$

Notes

Example: Given the following data, compute the coefficient of correlation *r*, between *X* and *Y*.

Age (yrs.) → Marks ↓	30 -50	50 -70	70 -90	Total
0 - 5	10	6	2	18
5 - 10	3	5	4	12
10 - 15	4	7	9	20
Total	**17**	**18**	**15**	**50**

Solution:

Note: Instead of doing the computation work in a single table, it can be split into the following steps:

Taking mid-values of the class intervals, we have

Mid-values (X)	2.5	7.5	12.5
Mid-values (Y)	40	60	80

Let $u_i = \frac{X_i - 7.5}{5}$ and $v_i = \frac{Y_i - 60}{20}$

$\therefore$ various u values are : – 1 0 1

and various v values are : – 1 0 1

1. Calculation of $\Sigma f_{ij} u_i v_j$

u_i \ v_j	–1	0	1	Total
–1	10 (10)	6 (0)	2 (–2)	8
0	3 (0)	5 (0)	4 (0)	0
1	4 (–4)	7 (0)	9 (9)	5
Total	6	0	7	13

$\therefore$ $\Sigma f_{ij} u_i v_j = 13$

2. Calculation of $\Sigma f_i u_i$ and $\Sigma f_i u_i^2$

u_i	f_i	$f_i u_i$	$f_i u_i^2$
- 1	18	- 18	18
0	12	0	0
1	20	20	20
Total	50	2	38

3. Calculation of $\Sigma f_j' v_j$ and $\Sigma f_j' v_j^2$

v_j	f_j'	$f_j' v_j$	$f_j' v_j^2$
–1	17	– 17	17
0	18	0	0
1	15	15	15
Total	50	– 2	32

Substituting these values in the formula of r, we have

$$r = \frac{50 \times 13 - 2 \times (-2)}{\sqrt{50 \times 38 - 4}\sqrt{50 \times 32 - 4}} = \frac{654}{\sqrt{1896}\sqrt{1596}} = 0.376$$

Notes

9.1.7 Spearman's Rank Correlation

This is a crude method of computing correlation between two characteristics. In this method, various items are assigned ranks according to the two characteristics and a correlation is computed between these ranks. This method is often used in the following circumstances:

1. When the quantitative measurements of the characteristics are not possible, e.g., the results of a beauty contest where various individuals can only be ranked.
2. Even when the characteristics is measurable, it is desirable to avoid such measurements due to shortage of time, money, complexities of calculations due to large data, etc.
3. When the given data consist of some extreme observations, the value of Karl Pearson's coefficient is likely to be unduly affected. In such a situation the computation of the rank correlation is preferred because it will give less importance to the extreme observations.
4. It is used as a measure of the degree of association in situations where the nature of population, from which data are collected, is not known.

The coefficient of correlation obtained on the basis of ranks is called 'Spearman's Rank Correlation' or simply the 'Rank Correlation'. This correlation is denoted by ρ(rho).

Let X_i be the rank of ith individual according to the characteristics X and Y_i be its rank according to the characteristics Y. If there are n individuals, there would be n pairs of ranks (X_i, Y_i), $i = 1, 2, \ldots\ldots n$. We assume here that there are no ties, i.e., no two or more individuals are tied to a particular rank. Thus, X_i's and Y_i's are simply integers from 1 to n, appearing in any order.

The means of X and Y, i.e., $\overline{X} = \overline{Y} = \dfrac{1+2+\ldots\ldots+n}{n} = \dfrac{n(n+1)}{2n} = \dfrac{n+1}{2}$.

Also, $\sigma_x^2 =$ $\qquad \sigma_y^2 = \dfrac{1}{n}[1^2+2^2+\ldots+n^2] - \dfrac{(n+1)^2}{4} = \dfrac{1}{n}\left[\dfrac{n(n+1)(2n+1)}{6}\right] - \dfrac{(n+1)^2}{4} = \dfrac{n^2-1}{12}$

Let d_i be the difference in ranks of the ith individual, i.e.,

$$d_i = X_i - Y_i = \left(X_i - \overline{X}\right) - \left(Y_i - \overline{Y}\right) \left(\because \overline{X} = \overline{Y}\right)$$

Squaring both sides and taking sum over all the observations, we get

$$\Sigma d_i^2 = \sum\left[\left(X_i - \overline{X}\right) - \left(Y_i - \overline{Y}\right)\right]^2$$

$$= \sum\left(X_i - \overline{X}\right)^2 + \sum\left(Y_i - \overline{Y}\right)^2 - 2\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

Dividing both sides by n, we get

$$\frac{1}{n}\sum d_i^2 = \frac{1}{n}\sum\left(X_i - \overline{X}\right)^2 + \frac{1}{n}\sum\left(Y_i - \overline{Y}\right)^2 - \frac{2}{n}\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

$$= \sigma_X^2 + \sigma_Y^2 - 2Cov(X,Y) = 2\sigma_X^2 - 2Cov(X,Y) \left(\because \sigma_X^2 = \sigma_Y^2\right)$$

From this, we can write $1 - \rho = \dfrac{1}{n} \times \dfrac{\sum d_i^2}{2\sigma_X^2}$

or $\qquad \rho = 1 - \dfrac{1}{n} \times \dfrac{\sum d_i^2}{2\sigma_X^2} = 1 - \dfrac{1}{n} \times \dfrac{\sum d_i^2}{2} \times \dfrac{12}{n^2-1} = 1 - \dfrac{6\sum d_i^2}{n\left(n^2-1\right)}$

Notes

Notes This formula is not applicable in case of a bivariate frequency distribution.

Example: Following is the list of marks scored by eleven students in mathematics and English in their 12th standard examination.

Student	1	2	3	4	5	6	7	8	9	10
Maths	45	50	– 60	65	75	40	62	72	66	56
English	48	58	55	60	76	35	52	49	66	65

Solution:

Maths Score	Maths Rank	English Score	English Rank	d = Maths Rank – English rank	D^2
45	9	48	9	0	0
50	8	58	5	3	9
60	6	55	6	0	0
65	4	60	4	0	0
75	1	76	1	0	0
40	10	35	10	0	0
62	5	52	7	–2	4
72	2	49	8	–6	36
66	3	66	2	1	1
56	7	65	3	4	16

The sum of the squared difference in ranks (the sum of the entries in the D2 column) is given by:

$$0+9+0+0+0+0+0+4+36+1+16 = 66$$

Using the Spearman rank-correlation coefficient, we obtain:

$$r_s = 1 - \frac{6 \times 66}{10(10 \times 10 - 1)} = 0.56$$

The Spearman rank-correlation coefficient ranges from -1 to + 1. The estimate of 0.56 suggests a strong positive relationship between rank performance in Maths and English.

9.1.8 Case of Tied Ranks

In case of a tie, i.e., when two or more individuals have the same rank, each individual is assigned a rank equal to the mean of the ranks that would have been assigned to them in the event of there being slight differences in their values. To understand this, let us consider the series 20, 21, 21, 24, 25, 25, 25, 26, 27, 28. Here the value 21 is repeated two times and the value 25 is repeated three times. When we rank these values, rank 1 is given to 20. The values 21 and 21 could have been assigned ranks 2 and 3 if these were slightly different from each other. Thus, each value will be assigned a rank equal to mean of 2 and 3, i.e., 2.5. Further, the value 24 will be assigned a rank equal to 4 and each of the values 25 will be assigned a rank equal to 6, the mean of 5, 6 and 7 and so on.

Notes

Since the Spearman's formula is based upon the assumption of different ranks to different individuals, therefore, its correction becomes necessary in case of tied ranks. It should be noted that the means of the ranks will remain unaffected. Further, the changes in the variances are usually small and are neglected. However, it is necessary to correct the term Σd_i^2 and accordingly the correction factor $\frac{m(m^2-1)}{12}$, where m denotes the number of observations tied to a particular rank, is added to it for every tie. We note that there will be two correction factors, i.e., $\frac{2(4-1)}{12}$ and $\frac{3(9-1)}{12}$ in the above example.

9.1.9 Limits of Rank Correlation

A positive rank correlation implies that a high (low) rank of an individual according to one characteristic is accompanied by its high (low) rank according to the other. Similarly, a negative rank correlation implies that a high (low) rank of an individual according to one characteristic is accompanied by its low (high) rank according to the other. When r = +1, there is said to be perfect consistency in the assignment of ranks, i.e., every individual is assigned the same rank with regard to both the characteristics. Thus, we have $\Sigma d_i^2 = 0$ and hence, r = 1.

Similarly, when r = – 1, an individual that has been assigned 1st rank according to one characteristic must be assigned nth rank according to the other and an individual that has been assigned 2nd rank according to one characteristic must be assigned $(n - 1)$th rank according to the other, etc. Thus, the sum of ranks, assigned to every individual, is equal to $(n + 1)$, i.e., $X_i + Y_i = n + 1$ or $Y_i = (n + 1) - X_i$, for all $i = 1, 2, \ldots\ldots n$.

Further, $d_i = X_i - Y_i = X_i - (n + 1) + X_i = 2X_i - (n + 1)$

Squaring both sides, we have

$$d_i^2 = \left[2X_i - (n+1)\right]^2 = 4X_i^2 + (n+1)^2 - 4(n+1)X_i$$

Taking sum over all the observations, we have

$$\sum d_i^2 = 4\sum X_i^2 + n(n+1)^2 - 4(n+1)\sum X_i = \frac{4n(n+1)(2n+1)}{6} + n(n+1)^2 - \frac{4n(n+1)^2}{2}$$

$$= n(n+1)\left[\frac{2}{3}(2n+1) + (n+1) - 2(n+1)\right] = \frac{n(n+1)(n-1)}{3} = \frac{n(n^2-1)}{3}$$

Substituting this value in the formula for rank correlation we have

$$\rho = 1 - \frac{6n(n^2-1)}{3} \times \frac{1}{n(n^2-1)} = -1$$

Hence, the Spearman's coefficient of correlation lies between – 1 and + 1.

Example: The following table gives the marks obtained by 10 students in commerce and statistics. Calculate the rank correlation.

Marks in Statistics	35	90	70	40	95	45	60	85	80	50
Marks in Commerce	45	70	65	30	90	40	50	75	85	60

Solution:

Notes

Calculation Table

		Rank of Marks in			
Marks in Statistics	**Marks in Commerce**	**Statistics *X***	**Commerce *Y***	$d_i = \overline{X}_i - \overline{Y}_i$	$\mathbf{d_i^2}$
35	45	1	3	–2	4
90	70	9	7	2	4
70	65	6	6	0	0
40	30	2	1	1	1
95	90	10	10	0	0
45	40	3	2	1	1
60	50	5	4	1	1
85	75	8	8	0	0
80	85	7	9	–2	4
50	60	4	5	–1	1

From the above table, we have $\Sigma d_i^2 = 16$.

$\therefore$ Rank Correlation r = $1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 16}{10 \times 99} = 0.903$

9.1.10 Coefficient of Correlation by Concurrent Deviation Method

This is another simple method of obtaining a quick but crude idea of correlation between two variables. In this method, only direction of change in the concerned variables are noted by comparing a value from its preceding value. If the value is greater than its preceding value, it is indicated by a '+' sign; if less, it is indicated by a '–' sign and equal values are indicated by '=' sign. All the pairs having same signs, i.e., either both the deviations are positive or negative or have equal sign ('='), are known as concurrent deviations and are indicated by '+' sign in a separate column designated as 'concurrences'. The number of such concurrences is denoted by *C*. Similarly, the remaining pairs are marked by '–' sign in another column designated as 'disagreements'. The coefficient of correlation, denoted by *rC*' is given by the formula $r_C = \pm\sqrt{\pm\left(\frac{2C - D}{D}\right)}$, where *C* denotes the number of concurrences and *D* (= number of observations – 1) is the number of pairs of deviation.

Notes

1. The sign of *rC* is taken to be equal to the sign of $\left(\frac{2C - D}{D}\right)$.
2. When $\left(\frac{2C - D}{D}\right)$ is negative, we make it positive for the purpose of taking its square root. However, the computed value will have a negative sign.
3. The sign of *rC* will be positive when $\left(\frac{2C - D}{D}\right)$ is positive.
4. This method gives same weights to smaller as well as to the larger deviations.

Notes

Caution This method is suitable only for the study of short term fluctuations because it does not take into account the changes in magnitudes of the values.

Self Assessment

Fill in the blanks:

1. The coefficient of correlation obtained on the basis of ranks is called .....
2. The only merit of Karl Pearson's coefficient of correlation is that it is the most popular method for expressing the ..... and ..... of linear association.
3. The ..... of correlation coefficient is an amount which if added to and subtracted from the mean correlation coefficient, gives limits within which the chances are even that a coefficient of correlation from a series selected at random will fall.
4. The value of Karl Pearson's coefficient is unduly affected by ..... items.
5. ..... correlation is used as a measure of the degree of association in situations where the nature of population, from which data are collected, is not known.
6. A ..... rank correlation implies that a high (low) rank of an individual according to one characteristic is accompanied by its high (low) rank according to the other.
7. When two or more individuals have the same rank, each individual is assigned a rank equal to the ..... of the ranks that would have been assigned to them in the event of there being slight differences in their values.
8. If correlation coefficient is greater than ..... and probable error is relatively ....., the correlation coefficient should be considered as significant.
9. Coefficient of correlation r does not give any idea about the existence of ..... relationship between the variables.
10. If X and Y are independent they are .....
11. The coefficient of correlation lies between .....
12. The coefficient of correlation is independent of the change of ..... and ..... of measurements.
13. Even when the characteristics are measurable, it is desirable to ..... such measurements due to shortage of time, money, complexities of calculations due to large data, etc.

9.2 Multiple Correlation

The coefficient of multiple correlation in case of regression of x_i on x_j and x_k, denoted by $R_{i\cdot jk}$, is defined as a simple coefficient of correlation between x_i and x_{ic}.

Thus
$$R_{i\times jk} = \frac{Cov(x_i, x_{ic})}{\sqrt{Var(x_i)Var(x_{ic})}} = \frac{\sum x_i x_{ic}}{\sqrt{\sum x_i^2 \sum x_{ic}^2}} = \frac{\sum x_i (x_i - x_{i.jk})}{\sqrt{\sum x_i^2 \sum (x_i - x_{i.jk})^2}}$$

$$= \frac{\sum x_i^2 - \sum x_i x_{i\cdot jk}}{\sqrt{\sum x_i^2 \sum (x_i - x_{i\cdot jk}) x_i}} = \frac{\sum x_i^2 - \sum x_i x_{i\cdot jk}}{\sqrt{\sum x_i^2 \left(\sum x_i^2 - \sum x_i x_{i\cdot jk}\right)}}$$

(Using property III)

Notes

$$= \frac{nS_i^2 - nS_{i.jk}^2}{\sqrt{nS_i^2\left(nS_i^2 - nS_{i.jk}^2\right)}} = \frac{1}{S_i}\sqrt{S_i^2 - S_{i.jk}^2} \quad \text{.... (13)}$$

Square of $R_{i\times jk}$ is known as the coefficient of multiple determination.

$$R_{i.jk}^2 = \frac{1}{S_i^2}\left(S_i^2 - S_{i.jk}^2\right) = 1 - \frac{S_{i.jk}^2}{S_i^2} \quad \text{.... (14)}$$

It may be noted here that $\frac{S_{i.jk}^2}{S_i^2}$ is proportion of unexplained variation. Thus, we can also write $R_{i.jk}^2 = 1 - \frac{x_{i.jk}^2}{x_i^2}$.

Further, we can write $R_{i.jk}^2$ in terms of the simple correlation coefficients.

$$R_{i.jk}^2 = 1 - \frac{S_i^2\left(1 - r_{ij}^2 - r_{ik}^2 - r_{jk}^2 + 2r_{ij}r_{ik}r_{jk}\right)}{S_i^2\left(1 - r_{jk}^2\right)} = \frac{r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}r_{jk}}{1 - r_{jk}^2}$$

Notes If there are m variables, $R_{1\cdot 23....m}^2 = 1 - \frac{S_{1\cdot 23....m}^2}{S_1^2} = 1 - \frac{\sum x_{1\cdot 23....m}^2}{\sum x_1^2}$

Self Assessment

Fill in the blanks:

14. The coefficient of ……correlation in case of regression of *xi* on *xj* and *xk*, denoted by *Ri*×*jk*
15. The coefficient of multiple correlation in case of regression of *xi* on *xj* and *xk* is defined as a ……coefficient of correlation between *xi* and *xic*.

9.3 Partial Correlation

In case of three variables x_i, x_j and x_k, the partial correlation between x_i and x_j is defined as the simple correlation between them after eliminating the effect of x_k. This is denoted as $r_{ij\times k}$.

We note that $x_{i\times k} = x_i - b_{ik}x_k$ is that part of x_i which is left after the removal of linear effect of x_k on it. Similarly, $x_{j\times k} = x_j - b_{jk}x_k$ is that part of x_j which is left after the removal of linear effect of x_k on it. Equivalently, $r_{ij\times k}$ can also be regarded as correlation between $x_{i\times k}$ and $x_{j\times k}$. Thus, we can write

$$r_{ij\cdot k} = \frac{\sum x_{i\cdot k}x_{j\cdot k}}{\sqrt{\sum x_{i\cdot k}^2 x_{j\cdot k}^2}}.$$

Using property III of residual products, we can write

$$Sx_{i\times k}x_{j\times k} = Sx_{i\times k}x_j = S(x_i - b_{ik}x_k)x_j = Sx_ix_j - b_{ik}Sx_jx_k$$

$$= nS_iS_jr_{ij} - r_{ik}\frac{S_i}{S_k}nS_jS_kr_{jk} = nS_iS_j\left(r_{ij} - r_{ik}r_{jk}\right)$$

Notes

Further, using property III, we can write

$$\Sigma x_{i\times k}^2 = Sx_i x_{i\times k} = Sx_i(x_i - b_{ik}x_k) = Sx_i^2 - b_{ik}Sx_i x_k$$

$$= nS_i^2 - r_{ik}\frac{S_i}{S_k}nS_iS_kr_{ik} = nS_i^2\left(1-r_{ik}^2\right)$$

Similarly, $\Sigma x_{i\times k}^2 = nS_j^2\left(1-r_{jk}^2\right).$

Thus, we have $$r_{ij\times k} = \frac{nS_iS_j\left(r_{ij}-r_{ik}r_{jk}\right)}{\sqrt{nS_i^2\left(1-r_{ik}^2\right)nS_j^2\left(1-r_{jk}^2\right)}} = \frac{r_{ij}-r_{ik}r_{jk}}{\sqrt{\left(1-r_{ik}^2\right)\left(1-r_{jk}^2\right)}}$$

Self Assessment

Fill in the blanks:

16. In case of three variables *xi*, *xj* and *xk*, the partial correlation between *xi* and *xj* is defined as the simple correlation between them after eliminating the effect of……..
17. …… correlation is denoted as *rij*×*k*

9.4 Regression Analysis

If the coefficient of correlation calculated for bivariate data (X_i, Y_i), i = 1, 2, ... n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as regression equation in statistics. Since the coefficient of correlation is measure of the degree of linear association of the variables, we shall discuss only linear regression equation. This does not, however, imply the non-existence of non-linear regression equations.

The regression equations are useful for predicting the value of dependent variable for given value of the independent variable. As pointed out earlier, the nature of a regression equation is different from the nature of a mathematical equation, e.g., if $Y = 10 + 2X$ is a mathematical equation then it implies that Y is exactly equal to 20 when $X = 5$. However, if $Y = 10 + 2X$ is a regression equation, then $Y = 20$ is an average value of Y when $X = 5$.

The term regression was first introduced by Sir Francis Galton in 1877. In his study of the relationship between heights of fathers and sons, he found that tall fathers were likely to have tall sons and vice-versa. However, the mean height of sons of tall fathers was lower than the mean height of their fathers and the mean height of sons of short fathers was higher than the mean height of their fathers. In this way, a tendency of the human race to regress or to return to a normal height was observed. Sir Francis Galton referred this tendency of returning to the mean height of all men as regression in his research paper, "Regression towards mediocrity in hereditary stature". The term 'Regression', originated in this particular context, is now used in various fields of study, even though there may be no existence of any regressive tendency.

9.4.1 Simple Regression

For a bivariate data (X_i, Y_i), i = 1, 2, ... n, we can have either X or Y as independent variable. If X is independent variable then we can estimate the average values of Y for a given value of X. The relation used for such estimation is called regression of Y on X. If on the other hand Y is used for estimating the average values of X, the relation will be called regression of X on Y. For a

Notes

bivariate data, there will always be two lines of regression. It will be shown later that these two lines are different, i.e., one cannot be derived from the other by mere transfer of terms, because the derivation of each line is dependent on a different set of assumptions.

Line of Regression of Y on X

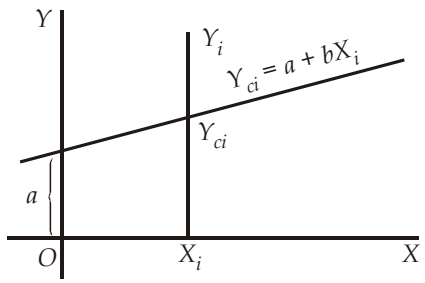
The general form of the line of regression of Y on X is $Y_{Ci} = a + bX_i$, where Y_{Ci} denotes the average or predicted or calculated value of Y for a given value of $X = X_i$. This line has two constants, a and b. The constant a is defined as the average value of Y when $X = 0$. Geometrically, it is the intercept of the line on Y-axis. Further, the constant b, gives the average rate of change of Y per unit change in X, is known as the regression coefficient.

The above line is known if the values of a and b are known. These values are estimated from the observed data (X_i, Y_i), $i = 1, 2, \ldots\ldots n$.

Notes It is important to distinguish between Y_{Ci} and Y_i. Where as Y_i is the observed value, Y_{Ci} is a value calculated from the regression equation.

Using the regression $YC_i = a + bX_i$, we can obtain $Y_{C1}, Y_{C2}, \ldots\ldots Y_{Cn}$ corresponding to the X values $X_1, X_2, \ldots\ldots X_n$ respectively. The difference between the observed and calculated value for a particular value of X say Xi is called error in estimation of the i th observation on the assumption of a particular line of regression. There will be similar type of errors for all the n observations. We denote by $e_i = Y_i - Y_{Ci}$ $(i = 1, 2, \ldots\ldots n)$, the error in estimation of the i th observation. As is obvious from Figure 9.4, ei will be positive if the observed point lies above the line and will be negative if the observed point lies below the line. Therefore, in order to obtain a Figure of total error, ei¢s are squared and added. Let S denote the sum of squares of these errors, i.e., $S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - Y_{Ci})^2$.

Figure 9.4



The regression line can, alternatively, be written as a deviation of Y_i from Y_{Ci} i.e. $Y_i - Y_{Ci} = e_i$ or $Y_i = Y_{Ci} + e_i$ or $Y_i = a + bX_i + e_i$. The component $a + bX_i$ is known as the deterministic component and e_i is random component.

The value of S will be different for different lines of regression. A different line of regression means a different pair of constants a and b. Thus, S is a function of a and b. We want to find such values of a and b so that S is minimum. This method of finding the values of a and b is known as the Method of Least Squares.

Rewrite the above equation as $\Sigma = \Sigma(Y_i - a - bX_i)^2$ $(\because\ Y_{Ci} = a + bX_i)$.

Notes

The necessary conditions for minima of S are

(i) $\frac{\partial S}{\partial a}=0$ and (ii) $\frac{\partial S}{\partial b}=0$, where $\frac{\partial S}{\partial a}$ and $\frac{\partial S}{\partial b}$ are the partial derivatives of S w.r.t. a and b respectively.

Now $$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n}\left(Y_i - a - bX_i\right)=0$$

or $$\sum_{i=1}^{n}\left(Y_i - a - bX_i\right) = \sum_{i=1}^{n}Y_i - na - b\sum_{i=1}^{n}X_i = 0$$

or $$\sum_{i=1}^{n}Y_i = na + b\sum_{i=1}^{n}X_i \quad \text{.... (1)}$$

Also, $$\frac{\partial S}{\partial b} = 2\sum_{i=1}^{n}\left(Y_i - a - bX_i\right)\left(-X_i\right)=0$$

or $$-2\sum_{i=1}^{n}\left(X_iY_i - aX_i - bX_i^2\right) = \sum_{i=1}^{n}\left(X_iY_i - aX_i - bX_i^2\right)=0$$

or $$\sum_{i=1}^{n}X_iY_i - a\sum_{i=1}^{n}X_i - b\sum_{i=1}^{n}X_i^2 = 0$$

or $$\sum_{i=1}^{n}X_iY_i = a\sum_{i=1}^{n}X_i + b\sum_{i=1}^{n}X_i^2 \quad \text{.... (2)}$$

Equations (1) and (2) are a system of two simultaneous equations in two unknowns a and b, which can be solved for the values of these unknowns. These equations are also known as normal equations for the estimation of a and b. Substituting these values of a and b in the regression equation $Y_{Ci} = a + b_{Xi}$, we get the estimated line of regression of Y on X.

Expressions for the Estimation of a and b.

Dividing both sides of the equation (1) by n, we have

$$\frac{\sum Y_i}{n}=\frac{na}{n}+\frac{b\sum X_i}{n} \text{ or } \bar{Y}=a+b\bar{X} \quad \text{.... (3)}$$

This shows that the line of regression $Y_{Ci} = a + bX_i$ passes through the point $\left(\bar{X}, \bar{Y}\right)$.

From equation (3), we have $a=\bar{Y}-b\bar{X}$ .... (4)

Substituting this value of a in equation (2), we have

$$\Sigma X_iY_i = \left(\bar{Y}-b\bar{X}\right)\Sigma X_i + b\Sigma X_i^2$$

$$= \bar{Y}\sum X_i - b\bar{X}\sum X_i + b\sum X_i^2 = n\bar{X}\bar{Y} - b.n\bar{X}^2 + b\sum X_i^2$$

or $$\sum X_iY_i - n\bar{X}\bar{Y} = b\left(\sum X_i^2 - n\bar{X}^2\right)$$

or $$b = \frac{\sum X_iY_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} \quad \text{.... (5)}$$

Notes

Also, $$\sum X_i Y_i - n\bar{X}\bar{Y} = \sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

and $$\sum X_i^2 - n\bar{X}^2 = \sum\left(X_i - \bar{X}\right)^2$$

$\therefore$ $$b = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_i - \bar{X}\right)^2} \quad \text{.... (6)}$$

or $$b = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{.... (7)}$$

where x_i and y_i are deviations of values from their arithmetic mean.

Dividing numerator and denominator of equation (6) by n we have

$$b = \frac{\frac{1}{n}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\frac{1}{n}\sum\left(X_i - \bar{X}\right)^2} = \frac{Cov(X,Y)}{\sigma_X^2} \quad \text{.... (8)}$$

The expression for b, which is convenient for use in computational work, can be written from equation (5) is given below:

$$b = \frac{\sum X_i Y_i - n\frac{\sum X_i}{n}\cdot\frac{\sum Y_i}{n}}{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2} = \frac{\sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

Multiplying numerator and denominator by n, we have

$$b = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum X_i^2 - \left(\sum X_i\right)^2} \quad \text{.... (9)}$$

To write the shortcut formula for b, we shall show that it is independent of change of origin but not of change of scale.

As in case of coefficient of correlation we define

$$u_i = \frac{X_i - A}{h}$$

and $$v_i = \frac{Y_i - B}{k}$$

or $$X_i = A + hu_i$$

and $$Y_i = B + kv_i$$

$\therefore$ $$\bar{X} = A + h\bar{u}$$

and $$\bar{Y} = B + k\bar{v}$$

also $$\left(X_i - \bar{X}\right) = h\left(u_i - \bar{u}\right)$$

and $$Y_i - \bar{Y} = k\left(v_i - \bar{v}\right)$$

Notes Substituting these values in equation (6), we have

$$b = \frac{hk\sum(u_i-\bar{u})(v_i-\bar{v})}{h^2\sum(u_i-\bar{u})^2} = \frac{k\sum(u_i-\bar{u})(v_i-\bar{v})}{h\sum(u_i-\bar{u})^2}$$

$$= \frac{k}{h}\left[\frac{n\sum u_i v_i - \left(\sum u_i\right)\left(\sum v_i\right)}{n\sum u_i^2 - \left(\sum u_i\right)^2}\right] \quad \text{.... (10)}$$

(***Note:*** if $h = k$ they will cancel each other)

Consider equation (8), $b = \frac{Cov(X,Y)}{\sigma_X^2}$

Writing $Cov(X, Y) = r \cdot \sigma_X\sigma_Y$, we have $b = \frac{r\cdot\sigma_X\sigma_Y}{\sigma_X^2} = r\cdot\frac{\sigma_Y}{\sigma_X}$

The line of regression of Y on X, i.e $Y_{Ci} = a + bX_i$ can also be written as

or $$Y_{Ci} = \bar{Y} - b\bar{X} + bX_i \quad \text{or} \quad Y_{Ci} - \bar{Y} = b\left(X_i - \bar{X}\right) \quad \text{.... (11)}$$

or $$\left(Y_{Ci} - \bar{Y}\right) = r\cdot\frac{\sigma_Y}{\sigma_X}\left(X_i - \bar{X}\right) \quad \text{.... (12)}$$

Line of Regression of X on Y

The general form of the line of regression of X on Y is $X_{Ci} = c + dY_i$, where X_{Ci} denotes the predicted or calculated or estimated value of X for a given value of $Y = Y_i$ and c and d are constants. d is known as the regression coefficient of regression of X on Y.

In this case, we have to calculate the value of c and d so that

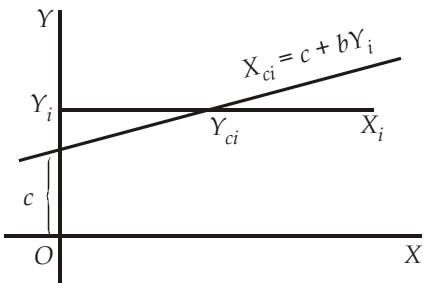
$S' = (X_i - X_{Ci})^2$ is minimised.

As in the previous section, the normal equations for the estimation of c and d are

$$\Sigma X_i = nc + d\Sigma Y_i \quad \text{.... (13)}$$

and $$\Sigma X_i Y_i = c\Sigma Y_i + d\Sigma Y_i^2 \quad \text{.... (14)}$$

Figure 9.5



Dividing both sides of equation (13) by n, we have $\bar{X} = c + d\bar{Y}$.

Notes

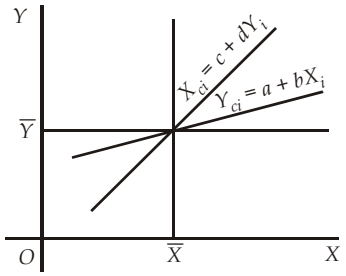
This shows that the line of regression also passes through the point $(\bar{X}, \bar{Y})$. Since both the lines of regression passes through the point $(\bar{X}, \bar{Y})$, therefore $(\bar{X}, \bar{Y})$ is their point of intersection as shown in Figure 9.6.

We can write $c = \bar{X} - d\bar{Y}$ .... (15)

As before, the various expressions for d can be directly written, as given below.

$$d = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum Y_i^2 - n\bar{Y}^2} \quad \text{.... (16)}$$

Figure 9.6



or $$d = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} \quad \text{.... (17)}$$

or $$d = \frac{\sum x_i y_i}{\sum y_i^2} \quad \text{.... (18)}$$

$$= \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum (Y_i - \bar{Y})^2} = \frac{Cov(X,Y)}{\sigma_Y^2} \quad \text{.... (19)}$$

Also $$d = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n\sum Y_i^2 - (\sum Y_i)^2} \quad \text{.... (20)}$$

This expression is useful for calculating the value of d. Another shortcut formula for the calculation of d is given by

$$d = \frac{h}{k}\left[\frac{n\sum u_i v_i - (\sum u_i)(\sum v_i)}{n\sum v_i^2 - (\sum v_i)^2}\right] \quad \text{.... (21)}$$

where $u_i = \frac{X_i - A}{h}$ and $v_i = \frac{Y_i - B}{k}$

Notes

Consider equation (19)

$$d = \frac{Cov(X,Y)}{\sigma_Y^2} = \frac{r\sigma_X\sigma_Y}{\sigma_Y^2} = r \cdot \frac{\sigma_X}{\sigma_Y} \quad \text{.... (22)}$$

Substituting the value of c from equation (15) into line of regression of X on Y we have

$$X_{Ci} = \bar{X} - d\bar{Y} + dY_i \text{ or } \left(X_{Ci} - \bar{X}\right) = d\left(Y_i - \bar{Y}\right) \quad \text{.... (23)}$$

or

$$\left(X_{Ci} - \bar{X}\right) = r \cdot \frac{\sigma_X}{\sigma_Y}\left(Y_i - \bar{Y}\right) \quad \text{.... (24)}$$

Remarks: It should be noted here that the two lines of regression are different because these have been obtained in entirely two different ways. In case of regression of Y on X, it is assumed that the values of X are given and the values of Y are estimated by minimising $S(Y_i - Y_{Ci})^2$ while in case of regression of X on Y, the values of Y are assumed to be given and the values of X are estimated by minimising $S(X_i - X_{Ci})^2$. Since these two lines have been estimated on the basis of different assumptions, they are not reversible, i.e., it is not possible to obtain one line from the other by mere transfer of terms. There is, however, one situation when these two lines will coincide. From the study of correlation we may recall that when $r = \pm 1$, there is perfect correlation between the variables and all the points lie on a straight line. Therefore, both the lines of regression coincide and hence they are also reversible in this case. By substituting $r = \pm 1$ in equation (12) or (24) it can be shown that the lines of regression in both the cases become

$$\left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right) = \pm\left(\frac{X_i - \bar{X}}{\sigma_X}\right)$$

Further when $r = 0$, equation (12) becomes $Y_{Ci} = \bar{Y}$ and equation (24) becomes $X_{Ci} = \bar{X}$. These are the equations of lines parallel to X-axis and Y-axis respectively. These lines also intersect at the point $(\bar{X}, \bar{Y})$ and are mutually perpendicular at this point, as shown in Figure.

9.4.2 Correlation Coefficient and the two Regression Coefficients

Since $b = r \cdot \frac{\sigma_Y}{\sigma_X}$ and $d = r \cdot \frac{\sigma_X}{\sigma_Y}$, we have

$b \cdot d = r\frac{\sigma_Y}{\sigma_X} \cdot r\frac{\sigma_X}{\sigma_Y} = r^2$ or $r = \sqrt{b \cdot d}$. This shows that correlation coefficient is the geometric mean of the two regression coefficients.

Remarks: The following points should be kept in mind about the coefficient of correlation and the regression coefficients:

1. Since $r = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$, $b = \frac{Cov(X,Y)}{\sigma_X^2}$ and $d = \frac{Cov(X,Y)}{\sigma_Y^2}$, therefore the sign of r, b and d will always be same and this will depend upon the sign of $Cov(X, Y)$.
2. Since $bd = r^2$ and $0 \le r^2 \le 1$, therefore either both b and d are less than unity or if one of them is greater than unity, the other must be less than unity such that $0 \le b \times d \le 1$ is always true.

Example: Obtain the two regression equations and find correlation coefficient between X and Y from the following data:

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

Solution:

Notes

Calculation Table

X	Y	XY	X^2	Y^2
10	6	60	100	36
9	3	27	81	9
7	2	14	49	4
8	4	32	64	16
11	5	55	121	25
45	**20**	**188**	**415**	**90**

(a) ***Regression of Y on X***

$$b = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2} = \frac{5\times 188 - 45\times 20}{5\times 415 - (45)^2} = 0.8$$

Also, $\bar{X} = \frac{45}{5} = 9$ and $\bar{Y} = \frac{20}{5} = 4$

Now $a = \bar{Y} - b\bar{X}$ = 4 - 0.8 × 9 = – 3.2

∴ Regression of Y on X is Y_C = – 3.2 + 0.8X

(b) ***Regression of X on Y***

$$d = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum Y^2 - \left(\sum Y\right)^2} = \frac{5\times 188 - 45\times 20}{5\times 90 - (20)^2} = 0.8$$

Also, $c = \bar{X} - d\bar{Y}$ = 9 – 0.8 × 4 = 5.8

∴ The regression of X on Y is X_C = 5.8 + 0.8Y

(c) ***Coefficient of correlation*** $r = \sqrt{b \cdot d} = \sqrt{0.8\times 0.8} = 0.8$

9.4.3 Non-parametric Regression

Nonparametric regression analysis traces the dependence of a response variable (y) on one or several predictors (xs) without specifying in advance the function that relates the response to the predictors:

$$E(y_i) = f(x_{1i},..., x_{pi})$$

where $E(y_i)$ is the mean of y for the ith of n observations. It is typically assumed that the conditional variance of y, $Var(y_i | x_{1i},..., x_{pi})$ is a constant, and that the conditional distribution of y is normal, although these assumptions can be relaxed.

There are many specific methods of non-parametric regression. Most, but not all, assume that the regression function is in some sense smooth.

The simplest use of non-parametric regression is in smoothing scatterplots. Here, there is a numerical response y and a single predictor x, and we seek to clarify visually the relationship between the two variables in a scatterplot. Three common methods of nonparametric regression are kernel estimation, local-polynomial regression (which is a generalization of kernel estimation), and smoothing splines.

Notes

Kernel Estimation

The kernel regression is a non-parametric technique in statistics to estimate the conditional expectation of a random variable. The objective is to find a non-linear relation between a pair of random variables X and Y.

In any non-parametric regression, the conditional expectation of a variable Y relative to a variable X may be written:

$$E(Y \mid X) = m(X)$$

where m is an unknown function.

Local Polynomial Regression

Figure 9.7: Local Polynomial Regression

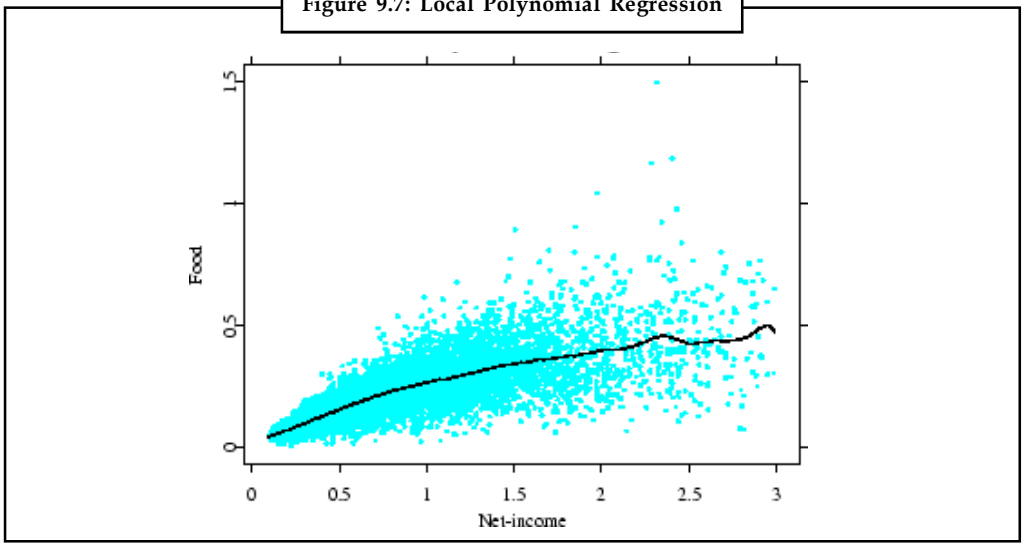


Figure above shows a local polynomial regression. A local polynomial regression is similar to kernel estimation, but the fitted values are produced by locally weighted regression rather than by locally weighted averaging. Most commonly, the order of the local polynomial is taken as k = 1, that is, a local linear fit. Local polynomial regression tends to be less biased than kernel regression, for example at the boundaries of data. More generally, the bias of the local-polynomial estimator declines and the variance increases with the order of the polynomial, but an odd-ordered local polynomial estimator has the same asymptotic variance as the preceding even-ordered estimator: Thus, the local-linear estimator (of order 1) is preferred to the kernel estimator (of order 0), and the local-cubic (order 3) estimator to the local-quadratic (order 2).

Smoothing Splines

The smoothing spline is a method of smoothing (fitting a smooth curve to a set of noisy observations) using a spline function.

Let (x_i, Y_i); $i = 1, \ldots, n$ be a sequence of observations, modeled by the relation $E(Yi) = (xi)$. The smoothing spline estimate of the function is defined to be the minimizer (over the class of twice differentiable functions) of

$$\sum_{i=1}^{n} (Y_i - \hat{\mu}(x_i))^2 + \lambda \int \hat{\mu}''(x)^2 dx$$

Notes

Remarks:

1. $l \geq 0$ is a smoothing parameter, controlling the trade-off between fidelity to the data and roughness of the function estimate.
2. The integral is evaluated over the range of the x_i.
3. As $\lambda \to 0$ (no smoothing), the smoothing spline converges to the interpolating spline.
4. As $\lambda \to \infty$ (infinite smoothing), the roughness penalty becomes paramount and the estimate converges to a linear least-squares estimate.
5. The roughness penalty based on the second derivative is the most common in modern statistics literature, although the method can easily be adapted to penalties based on other derivatives.
6. In early literature, with equally-spaced x_i, second or third-order differences were used in the penalty, rather than derivatives.
7. When the sum-of-squares term is replaced by a log-likelihood, the resulting estimate is termed penalized likelihood. The smoothing spline is the special case of *penalized likelihood* resulting from a Gaussian likelihood.

Multivariate Adaptive Regression Splines (MARS) is a form of regression analysis introduced by Jerome Friedman in 1991. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models non-linearities and interactions.

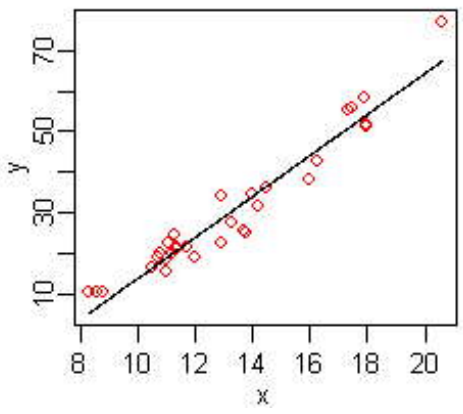
The term "MARS" is trademarked and licensed to Salford Systems.

This section introduces MARS using a few examples. We start with a set of data: a matrix of input variables x, and a vector of the observed responses y, with a response for each row in x. For example, the data could be:

x	y
10.5	16.4
10.7	18.8
10.8	19.7
...	...
20.6	77.0

Here there is only one independent variable, so the x matrix is just a single column. Given these measurements, we would like to build a model which predicts the expected y for a given x.

Figure 9.8



Notes

A Linear Model

A linear model for the above data is

$$\hat{y} = -37 + 5.1x$$

The hat on the indicates that is estimated from the data. The figure on the right shows a plot of this function: a line giving the predicted versus x, with the original values of y shown as red dots.

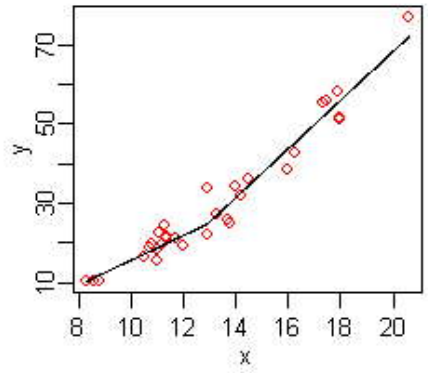
The data at the extremes of x indicates that the relationship between y and x may be non-linear (look at the red dots relative to the regression line at low and high values of x). We thus turn to MARS to automatically build a model taking into account non-linearities. MARS software constructs a model from the given x and y as follows:

$\hat{y}$ = 25

+ 6.1 max (0, x – 13)

– 3.1 max (0, 13 – x)

Figure 9.9



A Simple MARS Model of the Same Data

Figure 9.10 shows a plot of this function: the predicted versus x, with the original values of y once again shown as red dots. The predicted response is now a better fit to the original y values.

MARS has automatically produced a kink in the predicted y to take into account non-linearity. The kink is produced by hinge functions. The hinge functions are the expressions starting with max (where max(a, b) is a if $a > b$, else b). Hinge functions are described in more detail below.

In this simple example, we can easily see from the plot that the y has a non-linear relationship with x (and might perhaps guess that y varies with the square of x). However, in general there will be multiple independent variables, and the relationship between y and these variables will be unclear and not easily visible by plotting. We can use MARS to discover that non-linear relationship.

An example MARS expression with multiple variables is ozone = 5.2

+ 0.93 max(0, temp – 58)

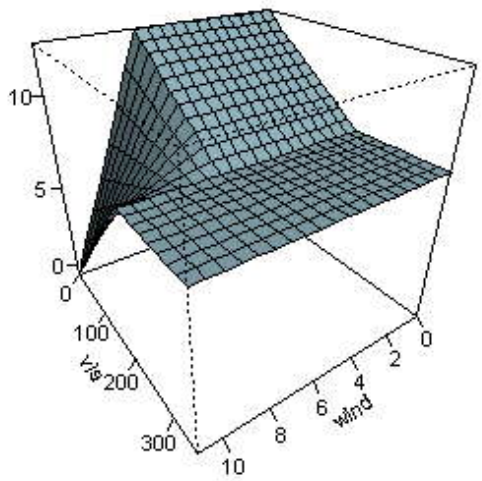
– 0.64 max(0, temp – 68)

Notes

– 0.046 max(0, 234 – ibt)

– 0.016 max(0, wind – 7) max(0, 200 – vis)

Figure 9.10



Variable Interaction in a MARS Model

This expression models air pollution (the ozone level) as a function of the temperature and a few other variables. Note that the last term in the formula (on the last line) incorporates an interaction between wind and vis.

The figure on the right plots the predicted ozone as wind and vis vary, with the other variables fixed at their median values. The figure shows that wind does not affect the ozone level unless visibility is low. We see that MARS can build quite flexible regression surfaces by combining hinge functions.

Self Assessment

Fill in the blanks:

18. The regression equations are useful for predicting the value of ..... variable for given value of the ..... variable.
19. …….is a form of regression analysis introduced by Jerome Friedman.
20. The ……regression is a non-parametric technique in statistics to estimate the conditional expectation of a random variable.

9.5 Summary

- Researchers sometimes put all the data together, as if they were one sample.
- There are two simple ways to approach these types of data.
- We can use the technique of correlation to test the statistical significance of the association.
- In other cases we use regression analysis to describe the relationship precisely by means of an equation that has predictive value.

Notes

- Straight-line (linear) relationships are particularly important because a straight line is a simple pattern that is quite common.
- The correlation measures the direction and strength of the linear relationship.
- The least-squares regression line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
- Non-parametric regression analysis traces the dependence of a response variable on one or several predictors without specifying in advance the function that relates the response to the predictors.

9.6 Keywords

Correlation: It is an analysis of covariation between two or more variables.

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

***Kernel Estimation*:** The kernel regression is a non-parametric technique in statistics to estimate the conditional expectation of a random variable.

Regression Equation: If the coefficient of correlation calculated for bivariate data (Xi, Yi), i = 1, 2, ... n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as regression equation in statistics.

***Smoothing Splines*:** It is a method of fitting a smooth curve to a set of noisy observations using a spline function.

9.7 Review Questions

1. Obtain the two lines of regression from the following data and estimate the blood pressure when age is 50 years. Can we also estimate the blood pressure of a person aged 20 years on the basis of this regression equation? Discuss.

Age (X) (in years)	*56*	*42*	*72*	*39*	*63*	*47*	*52*	*49*	*40*	*42*	*68*	*60*
Blood Pressure (Y)	*127*	*112*	*140*	*118*	*129*	*116*	*130*	*125*	*115*	*120*	*135*	*133*

2. Show that the coefficient of correlation, r, is independent of change of origin and scale.
3. Prove that the coefficient of correlation lies between – 1 and + 1.
4. "If two variables are independent the correlation between them is zero, but the converse is not always true". Explain the meaning of this statement.
5. What is Spearman's rank correlation? What are the advantages of the coefficient of rank correlation over Karl Pearson's coefficient of correlation?
6. Distinguish between correlation and regression. Discuss least square method of fitting regression.
7. What do you understand by linear regression? Why there are two lines of regression? Under what condition(s) can there be only one line?
8. What do you think as the reason behind the two lines of regression being different?
9. For a bivariate data, which variable can we have as independent? Why?
10. What can you conclude on the basis of the fact that the correlation between body weight and annual income were high and positive?

11. From the data given below, find out: **Notes**

    (i) Karl Pearson's coefficient of correlation.

    (ii) The two regression equations.

    (iii) The two regression coefficients.

    (iv) The most likely value of X when Y = 41.

    (v) The most likely value of Y when X = - 45.

X	52	60	58	39	41	53	47	34
Y	40	46	43	54	49	55	48	57

12. Find the multiple and partial correlation coefficients for the following data.

X	19	21	24	26	27	27	29	31	30	31
Y	24	28	29	39	30	31	34	35	36	37
Z	21	2-	26	30	27	32	31	36	33	38

13. Find the multiple correlation coefficient $R_{1.23}$, the partial correlation coefficient $r_{23.1}$ and the multiple regression equation of X_2 on X_3 X_1.

X1	55	59	63	68	56	73	82	76	64	74
Y1	58	60	53	52	61	70	76	77	63	80
Z1	63	55	51	56	59	74	74	81	61	84

14. Find the three multiple correlation coefficients for the following data.

X1	7	2	6	8	3	5	9	10	15	12
X2	1	5	8	4	9	3	6	7	13	10
X3	3	4	9	2	1	7	5	8	10	14

15. Find the regression equation of X_2 on X_3, X_1 for the following data.

X1	12	16	19	27	29	32	33	39	40	43
X2	7	9	11	13	16	19	24	31	33	37
X3	3	6	9	11	15	21	25	26	34	40

Answers: Self Assessment

1. 'Spearman's Rank Correlation
2. degree, direction
3. probable error
4. extreme
5. Rank
6. positive
7. mean
8. 0.3, small
9. cause and effect
10. uncorrelated
11. – 1 and + 1
12. origin, scale
13. avoid
14. multiple
15. simple
16. *xk*
17. Partial
18. dependent, independent
19. Multivariate Adaptive Regression Splines
20. kernel

Notes

9.8 Further Readings

Books

Abrams, M.A., *Social Surveys and Social Action*, London: Heinemann, 1951.

Arthur, Maurice, *Philosophy of Scientific Investigation*, Baltimore: John Hopkins University Press, 1943.

RS. Bhardwaj, *Business Statistics*, Excel Books, New Delhi, 2008.

S.N. Murthy and U. Bhojanna, *Business Research Methods*, Excel Books, 2007.

Unit 10: Time Series

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Recognize the time series analysis
- Identify the components of time series
- Explain the time series forecasting method
- Discuss the seasonal variations
- Describe the various methods to measure seasonal variations

Introduction

The future has always held a great fascination for mankind. Perhaps this is biologically determined. Man and the higher apes seem to have brains that are equipped to engage in actions for which a future reward is anticipated. In extreme situation reward is anticipated not in this life but in the next life.

Notes

There are two methodologies to anticipate future. They are called qualitative and quantitative. But both start with the same premise, that an understanding of the future is predicted on an understanding of the past and present environment. In this unit, we will mainly deal with quantitative methods. We will also distinguish between forecast and prediction. We use the word forecast when some logical method is used.

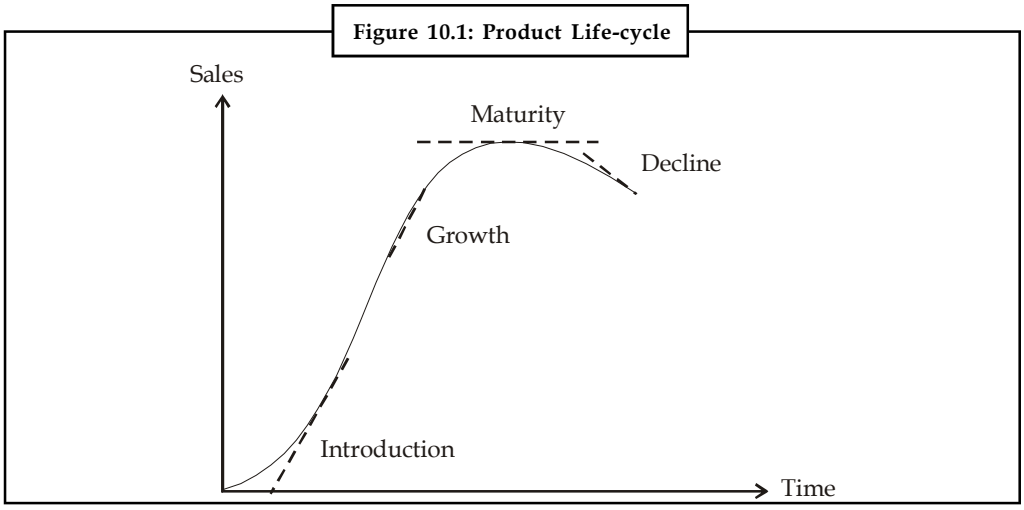
The quantitative decision maker always considers himself or herself accountable for a forecast—within reason. Let us look at the conceptual model first and then the mathematical model and algorithms in turn which are used for making forecast.

10.1 Time Series Analysis

Time has strange, fascinating and little understood properties. Virtually every process on earth is determined by a time variable. One of the most frequently encountered managerial decision situations involving forecasting is to measure the effect that time has on the sales of a product, the market price of a security, the output of individuals, work shifts, companies, industries, societies and so on. A fundamental conceptual model in all of these situations is the product life cycle concept which goes through four stages – introduction, growth, maturity and decline. Let us look at this concept in greater detail before we apply it.

Figure 10.1 depicts various stages in the life of a product. The sales performance of this product goes through the four stages—introduction, growth, maturity and decline. Data have been plotted and regression lines fitted to each of the four environments. Thus, when a sales forecast is made and the target horizon falls within the same stage, the linear fit yields valid results. If, however, the target horizon falls into a future stage, a linear forecast may be erroneous. In this case a curve should be fitted as shown. It is usually lightly speculative to select a forecasting horizon that spans more than two stages.

Figure 10.1: Product Life-cycle



Another point of interest is the behaviour of the sales variable over the short run. It fluctuates between a succession of peaks and troughs. How do these come about? In order to answer this question, the time series, must be decomposed. Then four independent motors for this behaviour become visible. First there is a long-term or secular trend (T) which is primarily noticeable within each stage of the cycle and over the entire cycle. Secondly cyclical variations (C) which are caused by an economy's business cycles affect product sales. Such cycles, whose origins are little understood, exist for all economies. Thirdly the product's sales may be influenced by the seasonality (S) of the item, and finally there may be the irregular (I) effects of inclement such as weather, strikes and so forth. In equation form the decomposed time series appears as TS = T + C + S + I.

Notes

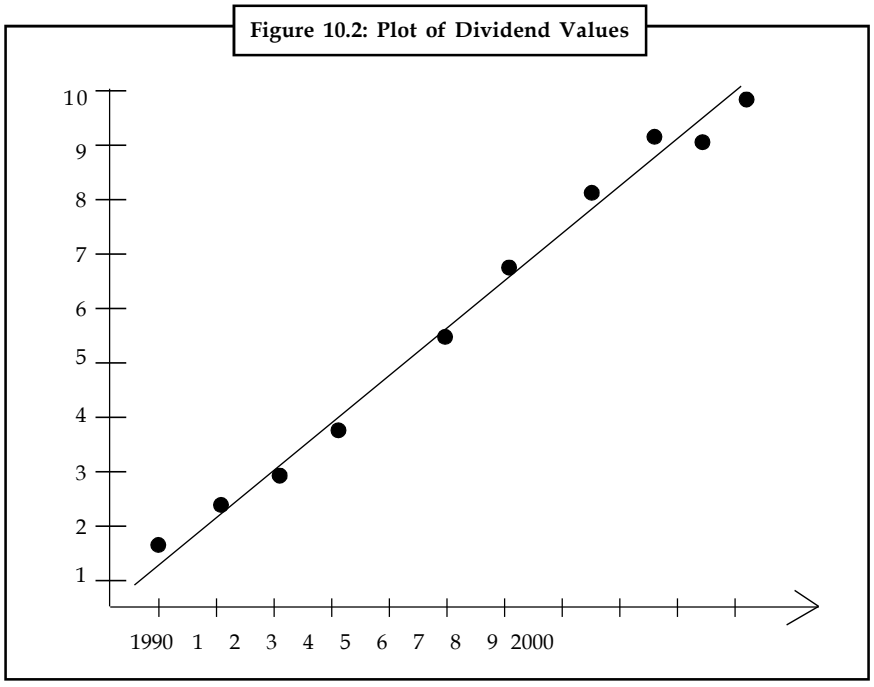
This creates a complex situation in time series analysis. Each factor must be quantified and its effect ascertained upon product sales. Let us see how this is done. The long-term trend effect T is reflected in the slope b of the regression equation. We already know how b is calculated even though minor modifications of the decision formulas will be encountered soon. The quantification of the cyclical component C is beyond the scope of this book. However, since business cycles always proceed from peak to trough to new peak and so on, their positive and negative effects upon a product's sales cancel out in the long-run. Hence in managerial, as opposed to economic, decision making, the sum effect of the business cycles may be set equal to zero. This eliminates the C factor from the equation. Seasonality, if present, is something that must be taken into consideration because it is a product-inherent variable and therefore it is under the immediate control of the decision maker. We will quantify the S component and keep it in the equation.

Finally, there are the irregular variations. Do we know in July whether the weather will be sunny and mild during the four weeks before Diwali? We don't, but we know that if this happens, Diwali sales will be severely impacted. Can we forecast such horrible weather conditions? Not really. We cannot forecast them because they cannot be quantified—a rather unpleasant characteristic they share with all other type of irregular variations like strikes, earthquakes, power failure, etc. Yet, something strange usually happens after such an irregular variation from "normal" has occurred. Whatever people did not do because of it, like not buying a product, they attempt to catch up with quickly. Therefore the factor effect may also be assumed to cancel out over time and it may be dropped from the equation which then appears to the manager as TS = T + S.

Linear Analysis

We will construct again the best fitting regression line by the method of least squares. It involves the dividend payments per share of the Smart, a well-known discount store chain, for the years 1990 through 1999. Suppose that a potential investor would like to know the dividend payment for 2001. The data are recorded in the work sheet (Table 10.1) that appears below. First, however, turn your attention to Figure 10.2 which shows the plot for this problem.

Figure 10.2: Plot of Dividend Values



Notes

Think for a moment about the qualitative nature of the time variable. It is expressed in years in this case but could be quarters, months, days, hours, minutes or any other time measurement unit. How does it differ from advertising expenditures, the independent variable that we examine in the preceding section? Is there a difference in the effect that a unit of each has on the dependent variable, or, ₹ 1 million in one case and 1 year in the other? Time, as you can readily see is constant. One year has the same effects as any other. This is not true for advertising expenditures, especially when you leave the linear environment and enter the nonlinear environments. Then there may be qualitative difference in the sales impact as advertising expenditures are increased or decreased by unit.

The worksheet is in Table 10.1 and calculations are as follows:

Table 10.1: Worksheet

YEAR	Code for an Even Series X	YEAR	Code for an Odd Series X	Divident payments in Rs Y	XY	X^2
1990	-9					
1991	-7	1991	-4	2.2	-8.8	16
1992	-5	1992	-3	2.4	-7.2	9
1993	-3	1993	-2	3.0	-6.0	4
1994	-1	1994	-1	5.0	-5.0	1
1995	1	1995	0	6.8	0	0
1996	3	1996	1	8.1	8.1	1
1997	5	1997	2	9.0	18.0	4
1998	7	1998	3	9.5	28.5	9
1999	9	1999	4	9.9	39.6	16
Total	**0**		**0**	**55.9**	**67.2**	**60**

Since time is constant in its effect, we may code the variable rather than to use the actual years or other time units x values. This code assigns a 1 to the first time period in the series and continues in unit distances to the nth period. Do not start with a zero as this may cause some computer programs to reject the input. The code is based on the fact that the unit periods are constant, and therefore their sum may be set equal to zero. See what effect this has on the normal equations for the straight line.

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

If $\Sigma x = 0$, the equations reduce to

$$\Sigma y = na$$

$$\Sigma xy = b\Sigma x^2$$

which allow the direct solution for a and b as follows

$$a = \frac{\Sigma y}{n}$$

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

Notes

This form simplifies the calculations substantially compared to the previous formulas. The code, however, that allows to set $\sum x = 0$ must incorporate the integrity of a unit distance series. Thus if the series is odd-numbered, the midpoint is set equal to zero and the code completed by negative and positive unit distances of x = 1 where each x unit stands for one year or other time period. If the series is even-numbered, let us say it ran from 1990 to 1999, the two midpoints (1994/1995) are set equal to -1 and +1, respectively. Since there is now a distance of x = 2 between + 1 (-1, 0, +1), the code continues by negative and positive units distance of x = 2 where each x unit stands for one-half year or other time period.

Then $a = \frac{5.59}{9} = 6.211$

$$b = \frac{\Sigma xy}{\Sigma x^2} = 1.12$$

and $Y_c = 6.211 + 1.12\,x$

origin 1995

x in 1 year units

The regression equation is plotted in Figure 10.2. Note that in the case of time series analysis, the origin of the code and the x units must be defined as part of the regression equation. In our problem the investor would like to obtain a dividend forecast for 2001. Since the origin is 1995 (x = 0) and x = 1 year units, the code for 2001 is x = 6. Therefore the forecast is yc = 6.211+1.12 (6) = ₹ 12.9. If the time series had been even numbered, let us say that dividend payments for 1990 had been included in the forecasting study, the definition under the regression equation would have read

origin 1994/95

x in 6 month units.

Thus, we know that for 1995, x=1; and since we must use x=2 units for each year, the code value for 2001 would be x=13. Once the yc value has been obtained, b is tested for significance and the 95% confidence interval constructed as previously shown.

Time series analysis is a long-term forecasting tool. Hence, it addresses itself to the trend component T in our time series equation TS = T + S. In the dividend forecast, b=1.120 was calculated which means that in the environment that is reflected in the set, smart increased the dividend payments on a average by ₹ 1.12 per year. Let us now turn out attention to the seasonal variation component that may be present in a time series. A product's seasonality is shown by the regularly recurring increases or decreases in sales or production that are caused by seasonal influences. In the case of some products, their seasonality is quite apparent. As an obvious example virtually all non-animal agricultural commodities may be cited. Seasonality of other products may be more difficult to detect. Take hogs in order to stay on the farm. Are they seasonal? They are lusty breeders and could not care less about seasonal influences. Yet, there is an induced season by the corn harvest. If corn is plentiful and cheap, farmers raise more hogs. This is known as the corn-hog cycle. Or take automobiles, Indian manufacturers are used to introduce major design or technological changes once every generation. This "season" has now been shortened somewhat. How about computers? There the season even has a special name. It is called a generation and prior to increased competitive pressures within the industry it used to be about seven years long. Our stock market investor knows that stock trades on the Stock Exchanges are seasonal. The daily season is V-shaped starting the trading with a relatively high volume which tapers off toward the lunch hour to pick up again in the afternoon. And so it goes with many other products, not ordinarily thought of as being seasonal.

Notes

Let us quantify this seasonality and illustrate how it may be used in a decision situation. There are, as is often the case, a number of decision tools that may be applied. The reader may be familiar with the term ratio-to-moving-average. It is a widely used method for constructing a seasonal index and programs are available in most larger computer libraries. Usually the method assumes a 12-period season like the twelve months of the year. There is a more efficient method which yields good statistical results. It is especially helpful in manual calculations of the seasonal index and when the number of seasonal periods is small like the four quarters of a year, the six hours of a stock exchange trading day or the five days of a work week. This method is known as simple average and will be used for illustration purposes.

To stay with the investment environment of this unit section, let us calculate a seasonal index for shares traded on the Stock Exchange from July 2 through July 7, 1999. This period includes the July 4 week-end. Volume of shares (DATA) for each trending day (SEASON) is given in thousands of shares per hour. The Individual steps of the analysis (OPERATIONS) are discussed in detail for each column of the worksheet below:

Table 10.2: Worksheet

(1)	Column (2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Hour (TS)	**Total Variation (T)**	**Trend Variation**	**Seasonal variation TS-T**	**Seasonal Index**	**7/2**	**7/3**	**7/6**	**7/7**	**Avg. for four days**
10-11	0.965	0	0.965	110.6	12.00	12.25	15.44	16.72	14.10
11-12	0.245	0.159	0.086	103.7	10.40	11.75	15.04	16.32	13.38
12-13	-0.885	0.318	-1.117	94.2	10.55	10.06	12.95	15.44	12.25
13-14	-1.555	0.477	-2.032	87.1	9.55	9.46	12.05	15.24	11.58
14-15	0.395	0.636	-0.241	101.1	11.02	11.55	14.82	16.73	13.53
15-16	0.835	0.795	0.040	103.3	11.58	12.25	15.38	16.69	13.97
Average			**-0.383**	**600**	**10.85**	**11.22**	**14.28**	**16.19**	**13.135**

As you inspect the data columns, you notice the V-shaped season for each trading day. You also notice in the total daily volume that there is a increase in shares traded. Hence, you can expect a positive slope of the regression line. The hourly mean number of shares is indicated also. This is the more important value because we are interested in quantifying a season by the hour for each trading day. Now turn to the operations. In last column the hourly trading activity for the four days has been summed. In this total all time series factors are assumed to be incorporated. You will recall that the positive or negative cyclical and irregular component effect is assumed to cancel out over time. Hence averaging the trading volume over a long term data set eliminates both components, yielding TS=T+S. You may ask, are four days a sufficiently long time span? The answer is NO. In a real study you would probably use 15 to 25 yearly averages for each trading hour. In an on-the-job application of this tool, you will have to know the specific time horizon in order to effectively eliminate cyclical and irregular variations. But by and large, what is a long or short time span depend upon situation.

In order to isolate the trend component (T) so that it may be subtracted from column (2) in the Table 10.2, yielding seasonal variation, the slope (b) of the regression line must be calculated. (Remember: b is T.) The necessary calculations are performed below using the mean hourly trading volume for each day. But since we are interested in an index by the hour, the calculated

Notes

daily b value must be apportioned to each hour. This is accomplished by a further division by six—the number of trading hours. The result is entered in column (3). Note that the origin of a time series is always zero. The origin of the time series is always the first period of the season. In our case this is the 10-11 trading hour. Therefore the first entry in column (3) is always zero to be followed by the equal (since this is a linear analysis) summed increment of the apportioned b-value.

Table 10.3: Worksheet for Trend Calculation

	Day	Code	Trading Volume Per Day		
	x	**X^1**	**y**	**x y**	**X^2**
	7/1	-3	10.85	-32.55	9
	7/5	-1	11.22	-11.22	1
	7/6	1	14.28	14.28	1
	7/7	3	16.19	48.57	9
Total			**52.54**	**19.08**	**20**

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

$$\frac{19.08}{20} = 0.954$$

and the apportioned b-value is

$$\frac{0.954}{6} = 0.159$$

It is not necessary to calculate the y-intercept (a) in this analysis unless of course, you wish to combine it with a long-term forecast of daily trading volume. Then, just to review the calculations, you would find:

$$a = \frac{\Sigma y}{n}$$

$$= \frac{52.54}{4}$$

$$= 13.135$$

and

$$y_c = 13.135 + 0.954\,x$$

origin 7/5 and 7/6

x in half trading day units.

In column (4) TS – T = S is performed. Column (4) is already a measure of seasonal variation. But in order to standardise the answer so that it may be compared with other stock exchange, for example, it is customary to convert the values in column (4) to a seasonal index. Every index has a base of 100 and the values above or below the base indicate percentages of above or below "normal" activity, hence the season. Since the base of column (5) is 100, the mean of the column should be 100 and the total 600 since there are 6 trading hours. In order to convert the obtained values of column (4) to index numbers, each of its entries is added to the total mean and then is divided by the column mean added to total mean and multiplied by 100 yielding the corresponding entry in column (5). It is customary to show index numbers with one significant digit.

Notes

Column (6) shows the seasonal effect of this decision variable—share trading on the Stock Exchange. Regardless of heavy or light daily volume, the first hour volume is the heaviest by far. It is 7.4% above what may be considered average trading volume for any given day. Keep in mind that a very limited data set was used in this analysis and while the season, reaching its low point between 1 and 2 p.m., is generally correctly depicted, individual index members may be exaggerated. What managerial action programs would result from an analyses such as this? Would traders go out for tea and samosas between 10-11? How about lunch between 1-2? When would brokers call clients with hot or lukewarm tips? Assuming that a decrease in volume means a decrease in prices in general during the trading day, when would a savvy trader buy? When would he sell? Think of some other intervening variables and you have yourself a nice little bull session in one of Dalal Street's watering holes. If, in addition, you make money for yourself or firm, then, you have got it.

Non-linear Analysis

Any number of different curves may be fitted to a data set. The most widely used program in computer libraries, known as CURFIT, offers a minimum of 5 curves plus the straight line. The curves may differ from program to program. So, which ones are the "best" ones? There is no answer. Every forecaster has to decide individually about his pet forecasting tools. We will discuss and apply three curves in this section. They appear to be promising decision tools especially in problem situations that in some way incorporate the life cycle concept and the range of such problems is vast, indeed.

If you take a look again at Figure 10.2, you see that three curves have been plotted. As we know from many empirical studies, achievement is usually normally distributed. Growth, on the other hand, seem to be exponentially distributed. The same holds true for decline. As the life cycle moves from growth to maturity, a parabolic trend may often be used as the forecasting tool. These are two of the curves that will be considered. The third one is related to the exponential curve. As you look at the growth stage and mentally extrapolate the trend, your eyes will run off the page. Now, we know—again from all sorts of empirical evidence—that trees don't grow into the high heavens. Even the most spectacular growth must come to an end. Therefore, when using the exponential forecast, care must be taken that the eventual ceiling or floor (in the case of a decline) are not overlooked. The modified exponential trend has the ceiling or floor build in. It is the third curve to be discussed.

One final piece of advice before we start fitting curves. If you can do it by straight line, do it. For obvious reasons, just look at Figure 10.2, any possible error—and there is always a built-in five percent chance—is worse when a curve is fitted. By extending the planning and forecasting horizon over a reasonable shorter period rather than spectacular but dangerous longer period, the straight line can serve as useful prediction tool.

The Parabola Fit

The parabola is defined by

$$y_c = a + bx + cx^2$$

Where a, b and c are constants a and b have been dealt. c can be treated as acceleration. The normal equations are (method of least square).

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Notes

Setting $\Sigma x = 0$ as previously explained, Sx3 will also be zero.

$$\Sigma y = na + c\Sigma x^2$$

$$\Sigma x^2 y = a\Sigma x^2 + c\Sigma x^4$$

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

There are direct formulas for a and c as well, but because of the possible compounding of arithmetic error in manual calculations, it is safer to solve a and c algebraically in this case.

To illustrate the parabolic trend let us forecast earnings per share in dollars for Storage Technology Corporation for the years 2000 and 2001. Storage Technology manufactures computer data storage equipment, printers, DVD-ROMS and telecommunication products. The company was founded in 1969 and after going through a period of explosive growth seems to be moving into the maturity stage. Data, code and calculations are shown below in the usual worksheet format.

Year	**Code**	**Earnings Per Share**				
	x	**y**	**xy**	**x^2**	**x^2y**	**x^4**
1993	-3	0.39	-1.17	9	3.51	81
1994	-2	0.54	-1.08	4	2.16	16
1995	-1	1.13	-1.13	1	1.13	1
1996	0	1.58	0	0	0	1
1997	1	1.72	1.72	1	1.72	1
1998	2	2.50	5.00	4	10.00	16
1999	3	1.84	5.52	9	16.56	81
Total	**0**	**9.70**	**8.86**	**28**	**35.08**	**196**

Then $b = \frac{8.86}{28}$

$= 0.3164$

and solving simultaneously

$$9.70 = 7a + 28c \times 4$$

$$\Rightarrow 38.8 = 28a + 112c$$

$$35.08 = 28a + 196c$$

$$3.72 = 84c$$

$$c = 0.0443$$

$$a = 1.2085$$

Therefore

$$y_c = 1.2085 + 0.3164\,x - 0.0443x^2$$

origin 1996

x in 1 year units

Notes

and specifically,

$$y_{2000} = 1.2085 + 0.3164(5) - 0.0443(5)^2$$

$$= ₹ 1.68,$$

$$y_{2001} = 1.2085 + 0.3164(6) - 0.0443(6)^2$$

$$= ₹ 1.51$$

Caution Remember that the data set is small. Quarterly earnings per share figures for the period may have been better because of the larger sample size. The significance test and construction of the confidence interval is performed as previously shown. Furthermore, as soon as new earnings per share figures become available, the regression line should be recalculated, because there is always the chance that there may be a change in the environment.

Self Assessment

Fill in the blanks:

1. Time series analysis is a ……..forecasting tool.
2. In time series analysis, Each factor must be …….and its effect ascertained upon product sales.
3. A product's seasonality is shown by the regularly recurring increases or decreases in sales or production that are caused by …….

10.2 Components of a Time Series

An observed value of a time series, Yt, is the net effect of many types of influences such as changes in population, techniques of production, seasons, level of business activity, tastes and habits, incidence of fire floods, etc. It may be noted here that different types of variables may be affected by different types of factors, e.g., factors affecting the agricultural output may be entirely different from the factors affecting industrial output. However, for the purpose of time series analysis, various factors are classified into the following three general categories applicable to any type of variable:

1. Secular Trend or Simply Trend
2. Periodic or Oscillatory Variations
   (a) Seasonal Variations
   (b) Cyclical Variations
3. Random or Irregular Variations

10.2.1 Secular Trend

Secular trend or simply trend is the general tendency of the data to increase or decrease or stagnate over a long period of time. Most of the business and economic time series would reveal a tendency to increase or to decrease over a number of years. For example, data regarding industrial production, agricultural production, population, bank deposits, deficit financing, etc.,

Notes

show that, in general, these magnitudes have been rising over a fairly long period. As opposed to this, a time series may also reveal a declining trend, e.g., in the case of substitution of one commodity by another, the demand of the substituted commodity would reveal a declining trend such as the demand for cotton clothes, demand for coarse grains like bajra, jowar, etc. With the improved medical facilities, the death rate is likely to show a declining trend, etc. The change in trend, in either case, is attributable to the fundamental forces such as changes in population, technology, composition of production, etc.

According to A.E. Waugh, secular trend is, "that irreversible movement which continues, in general, in the same direction for a considerable period of time". There are two parts of this definition: (i) movement in same direction, which implies that if the values are increasing (or decreasing) in successive periods, the tendency continues; and (ii) a considerable period of time.

Notes There is no specific period which can be called as a long period. Long periods are different for different situations. For example, in cases of population or output trends, the long period could be 10 years while it could be a month for the daily demand trend of vegetables. It should, however, be noted that longer is the period the more significant would be the trend. Further, it is not necessary that the increase or decrease of values must continue in the same direction for the entire period. The data may first show a rising (or falling) trend and subsequently a falling (or rising) trend.

Objectives of Measuring Trend

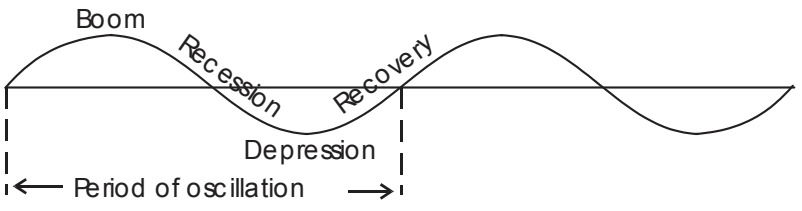
There are four main objectives of measuring trend of a time series data:

1. To study past growth or decline of the series. On ignoring the short term fluctuations, trend describes the basic growth or decline tendency of the data.
2. Assuming that the same behaviour would continue in future also, the trend curve can be projected into future for forecasting.
3. In order to analyse the influence of other factors, the trend may first be measured and then eliminated from the observed values.
4. Trend values of two or more time series can be used for their comparison.

10.2.2 Periodic Variations

These variations, also known as oscillatory movements, repeat themselves after a regular interval of time. This time interval is known as the period of oscillation. These oscillations are shown in the following Figure:

Figure 10.3



Notes

The oscillatory movements are termed as Seasonal Variations if their period of oscillation is equal to one year, and as Cyclical Variations if the period is greater than one year.

A time series, where the time interval between successive observations is less than or equal to one year, may have the effects of both the seasonal and cyclical variations. However, the seasonal variations are absent if the time interval between successive observations is greater than one year.

Although the periodic variations are more or less regular, they may not necessarily be uniformly periodic, i.e., the pattern of their variations in different periods may or may not be identical in respect of time period and size of periodic variations. For example, if a cycle is completed in five years then its following cycle may take greater or less than five years for its completion.

1. ***Causes of Seasonal Variations:*** The main causes of seasonal variations are: (a) Climatic Conditions and (b) Customs and Traditions

   (a) *Climatic Conditions:* The changes in climatic conditions affect the value of time series variable and the resulting changes are known as seasonal variations. For example, the sale of woolen garments is generally at its peak in the month of November because of the beginning of winter season. Similarly, timely rainfall may increase agricultural output, prices of agricultural commodities are lowest during their harvesting season, etc., reflect the effect of climatic conditions on the value of time series variable.

   (b) *Customs and Traditions:* The customs and traditions of the people also give rise to the seasonal variations in time series.

Example: The sale of garments and ornaments may be highest during the marriage season, sale of sweets during Diwali, etc., are variations that are the results of customs and traditions of the people.

It should be noted here that both of the causes, mentioned above, occur regularly and are often repeated after a gap of less than or equal to one year.

Objectives of Measuring Seasonal Variations: The main objectives of measuring seasonal variations are:

(a) To analyse the past seasonal variations.

(b) To predict the value of a seasonal variation which could be helpful in short-term planning.

(c) To eliminate the effect of seasonal variations from the data.

2. ***Causes of Cyclical Variations:*** Cyclical variations are revealed by most of the economic and business time series and, therefore, are also termed as trade (or business) cycles. Any trade cycle has four phases which are respectively known as boom, recession, depression and recovery phases. Various phases repeat themselves regularly one after another in the given sequence. The time interval between two identical phases is known as the period of cyclical variations. The period is always greater than one year. Normally, the period of cyclical variations lies between 3 to 10 years.

   Objectives of Measuring Cyclical Variations: The main objectives of measuring cyclical variations are:

   (a) To analyse the behaviour of cyclical variations in the past.

Notes

(b) To predict the effect of cyclical variations so as to provide guidelines for future business policies.

10.2.3 Random or Irregular Variations

As the name suggests, these variations do not reveal any regular pattern of movements. These variations are caused by random factors such as strikes, floods, fire, war, famines, etc. Random variations is that component of a time series which cannot be explained in terms of any of the components discussed so far. This component is obtained as a residue after the elimination of trend, seasonal and cyclical components and hence is often termed as residual component.

Random variations are usually short-term variations but sometimes their effect may be so intense that the value of trend may get permanently affected.

Self Assessment

Fill in the blanks:

4. ..... is the general tendency of the data to increase or decrease or stagnate over a long period of time.
5. The oscillatory movements are termed as …….if their period of oscillation is equal to one year.
6. Random variations are usually ……variations

10.3 Time Series Forecasting Method

Time series forecasting methods are based on analysis of historical data. They make the assumption that part patterns in data can be used to forecast future data points. In this unit, we will discuss two methods: (a) Moving Average Method and (b) Exponential Smoothing.

10.3.1 Method of Moving Average

This method is based on the principle that the total effect of periodic variations at different points of time in its cycle gets completely neutralised, i.e., SSt = 0 in one year and SCt = 0 in the period of cyclical variations.

In the method of moving average, successive arithmetic averages are computed from overlapping groups of successive values of a time series. Each group includes all the observations in a given time interval, termed as the period of moving average. The next group is obtained by replacing the oldest value by the next value in the series. The averages of such groups are known as the moving averages.

The moving average of a group is always shown at the centre of its period. The process of computing moving averages smoothens out the fluctuations in the time series data. It can be shown that if the trend is linear and the oscillatory variations are regular, the moving average with period equal to the period of oscillatory variations would completely eliminate them. Further, the effect of random variations would get minimised because the average of a number of observations must lie between the smallest and the largest observation. It should be noted here that the larger is the period of moving average the more would be the reduction in the effect of random component but more information is lost at the two ends of data.

When the trend is non-linear, the moving averages would give biased rather than the actual trend values.

Notes

Let Y_1, Y_2, ... Yn be the n values of a time series for successive time periods 1, 2, ... n respectively. The calculation of 3-period and 4-period moving averages are shown in the following tables:

Time Period	*Values of Y*	*3 - period M.A.*	*Time Period*	*Values of Y*	*4 - period M.A.*	*Centered Values*
1	Y_1	...	1	Y_1	...	...
2	Y_2	$\frac{Y_1+Y_2+Y_3}{3}$	2	Y_2	...	...
3	Y_3	$\frac{Y_2+Y_3+Y_4}{3}$	3	Y_3	$\frac{Y_1+Y_2+Y_3+Y_4}{4}=A_1$	$\frac{A_1+A_2}{2}$
4	Y_4	$\frac{Y_3+Y_4+Y_5}{3}$	4	Y_4	$\frac{Y_2+Y_3+Y_4+Y_5}{4}=A_2$	$\frac{A_2+A_3}{2}$
5	Y_5		5	Y_5	$\frac{Y_3+Y_4+Y_5+Y_6}{4}=A_3$	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	Y_n	...	*n*	Y_n	...	...

It should be noted that, in case of 3-period moving average, it is not possible to get the moving averages for the first and the last periods. Similarly, the larger is the period of moving average the more information will be lost at the ends of a time series.

When the period of moving average is even, the computed average will correspond to the middle of the two middle most periods. These values should be centred by taking arithmetic mean of the two successive averages. The computation of moving average in such a case is also illustrated in the above table.

Year	:	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production ('000' tonnes)	:	26	27	28	30	29	27	30	31	32	31

Example: Determine the trend values of the following data by using 3-year moving average. Also find short-term fluctuations for various years, assuming additive model. Plot the original and the trend values on the same graph.

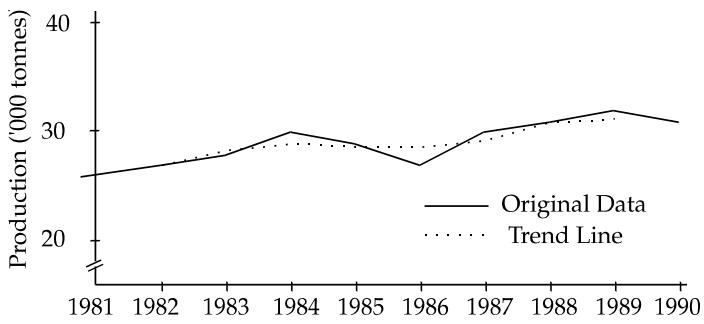
Solution:

Calculation of Trend and Short-term Fluctuations

Years	*Production (Y)*	*3-Year Moving Total*	*3-Year M.A. or Trend values (T)*	*Short-term fluctuations (Y – T)*
1981	26	...	...	...
1982	27	81	27.00	0.00
1983	28	85	28.33	−0.33
1984	30	87	29.00	1.00
1985	29	86	28.67	0.33
1986	27	86	28.67	−1.67
1987	30	88	29.33	0.67
1988	31	93	31.00	0.00
1989	32	94	31.33	0.67
1990	31	...	...	...

Notes

Graphical presentation of Y and T values



Example: Assuming a four-yearly cycle, find the trend values for the following data by the method of moving average.

Year	1979	1980	1981	1982	1983	1984	1985
Sales (in Rs '000)	74	100	97	87	90	115	126
Year	1986	1987	1988	1989	1990	1991	1992
Sales (in Rs '000)	108	100	125	118	113	122	126

Solution:

Calculation of Trend Values

Years	Scale (Y)	4 - Year Moving Total	Centered Total	4 - Year Moving Average (T)
1979	74	...	...	...
1980	100	358	732	91.50
1981	97	374	763	95.38
1982	87	389	807	100.88
1983	90	418	857	107.13
1984	115	439	888	111.00
1985	126	449	908	113.50
1986	108	459	910	113.75
1987	100	451	907	113.38
1988	125	456	934	116.75
1989	118	478	957	119.63
1990	113	479	...	...
1991	122	...		
1992	126			

10.3.2 Exponential Smoothing

Exponential smoothing gives greater weight to demand in more recent periods, and less weight to demand in earlier periods average: At = a Dt + (1 – a) At–1 = a Dt + (1 – a) Ft

forecast for period t + 1: Ft + 1 = At

where:

At –1 = "series average" calculated by the exponential smoothing model to period t – 1

a = smoothing parameter between 0 and 1 the larger the smoothing parameter, the greater the weight given to the most recent demand

Notes

Double Exponential Smoothing (Trend-Adjusted Exponential Smoothing)

When a trend exists, the forecasting technique must consider the trend as well as the series average ignoring the trend will cause the forecast to always be below (with an increasing trend) or above (with a decreasing trend) actual demand double exponential smoothing smooths (averages) both the series average and the trend

forecast for period t + 1: Ft + 1 = At + Tt

average: At = aDt + (1 – a) (At – 1 + Tt – 1) = aDt + (1 – a) Ft

average trend: Tt = B CTt + (1 – B) Tt - 1

current trend: CTt = At – At – 1

forecast for p periods into the future: Ft + p = At + pTt

where:

At = exponentially smoothed average of the series in period t

Tt = exponentially smoothed average of the trend in period t

CTt = current estimate of the trend in period t

a = smoothing parameter between 0 and 1 for smoothing the averages

B = smoothing parameter between 0 and 1 for smoothing the trend

Self Assessment

Fill in the blanks:

7. Time series forecasting methods are based on analysis of ……data.
8. ……..gives greater weight to demand in more recent periods.
9. ……..method is based on the principle that the total effect of periodic variations at different points of time in its cycle gets completely neutralised.

10.4 The Mean Absolute Deviation (MAD)

The mean absolute deviation is the sum of the absolute values of the deviations from the mean.

Procedure

1. Find the mean of the data
2. Subtract the mean from each data value to get the deviation from the mean
3. Take the absolute value of each deviation from the mean
4. Total the absolute values of the deviations from the mean
5. Divide the total by the sample size.

Typically the point from which the deviation is measured is the value of either the median or the mean of the data set.

$$|D| = |x_i - m(X)|$$

Notes

where

|D| is the absolute deviation,

x_i is the data element

and m(X) is the chosen measure of central tendency of the data set–sometimes the mean (), but most often the median.

The average absolute deviation or simply average deviation of a data set is the average of the absolute deviations and is a summary statistic of statistical dispersion or variability. It is also called the mean absolute deviation, but this is easily confused with the median absolute deviation.

The average absolute deviation of a set $\{x_1, x_2, ..., x_n\}$ is

$$\frac{\Sigma |x - \bar{x}|}{n}$$

Did u know? **What is median absolute deviation?**

The median absolute deviation is a measure of statistical dispersion. It is a more robust estimator of scale than the sample variance or standard deviation; it also exists for some distributions which may not have a mean or variance.

The MAD is a robust statistic, being more resilient to outliers in a data set than the standard deviation. In the standard deviation, the distances from the mean are squared, so on average, large deviations are weighted more heavily, and thus outliers can heavily influence it. In the MAD, the magnitude of the distances of a small number of outliers is irrelevant.

The MAD can be used to estimate the scale parameter of distributions for which the variance and standard deviation do not exist, such as the Cauchy distribution.

Self Assessment

Fill in the blanks:

10. Typically the point from which the deviation is measured is the value of either the ……..or the …….....of the data set.
11. The Mean Absolute Deviation is a robust statistic, being more resilient to outliers in a data set than the……...
12. The Mean Absolute Deviation can be used to estimate the scale parameter of distributions for which the ……and standard deviation do not exist.

10.5 Mean Squared Error (MSE)

MSE is the sum of the squared forecast errors for each of the observations divided by the number of observations. It is an alternative to the mean absolute deviation, except that more weight is placed on larger errors. While MSE is popular among statisticians, it is unreliable and difficult to interpret. The mean squared error of an estimator is the expected value of the square of the "error." The error is the amount by which the estimator differs from the quantity to be estimated. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.

MSE of an estimator is one of many ways to quantify the amount by which an estimator differs from the true value of the quantity being estimated. As a loss function, MSE is called squared error loss. MSE measures the average of the square of the "error." The error is the amount by

Notes

which the estimator differs from the quantity to be estimated. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.

The MSE is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator and its bias. For an unbiased estimator, the MSE is the variance. Like the variance, MSE has the same unit of measurement as the square of the quantity being estimated. In an analogy to standard deviation, taking the square root of MSE yields the root mean squared error or RMSE, which has the same units as the quantity being estimated; for an unbiased estimator, the RMSE is the square root of the variance, known as the standard error.

Mean squared error of an estimator b of true parameter vector B is:

$$MSE(b) = E[(b - B)^2]$$

which is also

$$MSE(b) = var(b) + (bias(b))(bias(b'))$$

Among unbiased estimators, the minimal MSE is equivalent to minimizing the variance, and is obtained by the MVUE. However, a biased estimator may have lower MSE. In statistical modelling, the MSE is defined as the difference between the actual observations and the response predicted by the model and is used to determine whether the model does not fit the data or whether the model can be simplified by removing terms. Like variance, mean squared error has the disadvantage of heavily weighting outliers. This is a result of the squaring of each term, which effectively weights large errors more heavily than small ones. This property, undesirable in many applications, has led researchers to use alternatives such as the mean absolute error, or those based on the median.

Did u know? **What is key criterion in selecting estimators?**

Minimizing MSE is a key criterion in selection estimators.

Self Assessment

Fill in the blanks:

13. …….is the sum of the squared forecast errors for each of the observations divided by the number of observations.
14. Mean Squared Error of an estimator is one of many ways to quantify the amount by which an estimator differs from the ……..of the quantity being estimated.
15. The ……is the amount by which the estimator differs from the quantity to be estimated.

10.6 Seasonal Variations

If the time series data are in terms of annual figures, the seasonal variations are absent. These variations are likely to be present in data recorded on quarterly or monthly or weekly or daily or hourly basis. As discussed earlier, the seasonal variations are of periodic nature with period equal to one year. These variations reflect the annual repetitive pattern of the economic or business activity of any society. The main objectives of measuring seasonal variations are:

1. To understand their pattern.
2. To use them for short-term forecasting or planning.

Notes

3. To compare the pattern of seasonal variations of two or more time series in a given period or of the same series in different periods.

4. To eliminate the seasonal variations from the data. This process is known as deseasonalisation of data.

The measurement of seasonal variation is done by isolating them from other components of a time series. There are four methods commonly used for the measurement of seasonal variations. These method are:

1. Method of Simple Averages

2. Ratio to Trend Method

3. Ratio to Moving Average Method

4. Method of Link Relatives

Notes In the discussion of the above methods, we shall often assume a multiplicative model. However, with suitable modifications, these methods are also applicable to the problems based on additive model.

10.6.1 Method of Simple Averages

This method is used when the time series variable consists of only the seasonal and random components. The effect of taking average of data corresponding to the same period (say 1st quarter of each year) is to eliminate the effect of random component and thus, the resulting averages consist of only seasonal component. These averages are then converted into seasonal indices, as explained in the following examples.

Example: Assuming that trend and cyclical variations are absent, compute the seasonal index for each month of the following data of sales (in ₹ '000) of a company.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1987	46	45	44	46	45	47	46	43	40	40	41	45
1988	45	44	43	46	46	45	47	42	43	42	43	44
1989	42	41	40	44	45	45	46	43	41	40	42	45

Solution:

Calculation Table

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1987	46	45	44	46	45	47	46	43	40	40	41	45
1988	45	44	43	46	46	45	47	42	43	42	43	44
1989	42	41	40	44	45	45	46	43	41	40	42	45
Total	133	130	127	136	136	137	139	128	124	122	126	134
A_i	44.3	43.3	42.3	45.3	45.3	45.7	46.3	42.7	41.3	40.7	42.0	44.7
S.I.	101.4	99.1	96.8	103.7	103.7	104.6	105.9	97.7	94.5	93.1	96.1	102.3

In the above table, A_i denotes the average and S.I. the seasonal index for a particular month of various years. To calculate the seasonal index, we compute grand average given by G = $\frac{\Sigma A_i}{12}$ = $\frac{524}{12}$ = 437. Then the seasonal index for a particular month is given by S.I = $\frac{A_i}{G} \times 100$.

Notes

Further, ΣS.I. = 1198.9 ≠ 1200. Thus, we have to adjust these values such that their total is 1200. This can be done by multiplying each figure by $\frac{1200}{1198.9}$. The resulting figures are the adjusted seasonal indices, as given below:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
101.5	99.2	96.9	103.8	103.8	104.7	106.0	97.8	94.6	93.2	96.2	102.3

Remarks: The totals equal to 1200, in case of monthly indices and 400, in case of quarterly indices, indicate that the ups and downs in the time series, due to seasons, neutralise themselves within that year. It is because of this that the annual data are free from seasonal component.

Task Compute the seasonal index from the following data by the method of simple averages.

Year	Quarter	Y	Year	Quarter	Y	Year	Quarter	Y
1980	*I*	106	1982	*I*	90	1984	*I*	80
	II	124		*II*	112		*II*	104
	III	104		*III*	101		*III*	95
	IV	90		*IV*	85		*IV*	83
1981	*I*	84	1983	*I*	76	1985	*I*	104
	II	114		*II*	94		*II*	112
	III	107		*III*	91		*III*	102
	IV	88		*IV*	76		*IV*	84

Merits and Demerits

This is a simple method of measuring seasonal variations which is based on the unrealistic assumption that the trend and cyclical variations are absent from the data. However, we shall see later that this method, being a part of the other methods of measuring seasonal variations, is very useful.

10.6.2 Ratio to Trend Method

This method is used when cyclical variations are absent from the data, i.e., the time series variable Y consists of trend, seasonal and random components.

Using symbols, we can write Y = T.S.R.

Various steps in the computation of seasonal indices are:

1. Obtain the trend values for each month or quarter, etc., by the method of least squares.
2. Divide the original values by the corresponding trend values. This would eliminate trend values from the data. To get figures in percentages, the quotients are multiplied by 100.

   Thus, we have $\frac{Y}{T} \times 100 = \frac{T.S.R}{T} \times 100 = \text{S.R.100}$

3. Finally, the random component is eliminated by the method of simple averages.

Notes

Example: Assuming that the trend is linear, calculate seasonal indices by the ratio to moving average method from the following data:

Quarterly output of coal in 4 years
(in thousand tonnes)

Year	*I*	*II*	*III*	*IV*
1982	65	58	56	61
1983	68	63	63	67
1984	70	59	56	52
1985	60	55	51	58

Solution:

By adding the values of all the quarters of a year, we can obtain annual output for each of the four years. Fit a linear trend to the data and obtain trend values for each quarter.

From the above table, $a = \frac{962}{4} = 240.5$ and $b = \frac{-72}{20} = -3.6$

Thus, the trend line is Y = 240.5 – 3.6X, Origin : 1st January 1984, unit of X : 6 months.

The quarterly trend equation is given by

or $Y = \frac{240.5}{4} - \frac{3.6}{8}X$ or Y = 60.13 – 0.45X, Origin : 1st January 1984, unit of X : 1 quarter (i.e., 3 months).

Shifting origin to 15th Feb. 1984, we get

$Y = 60.13 - 0.45(X + \frac{1}{2}) = 59.9 - 0.45X$, origin I-quarter, unit of X = 1 quarter.

The table of quarterly values is given by

Year	*I*	*II*	*III*	*IV*
1982	63.50	63.05	62.60	62.15
1983	61.70	61.25	60.80	60.35
1984	59.90	59.45	59.00	58.55
1985	58.10	57.65	57.20	56.75

The table of Ratio to Trend Values, i.e., $\frac{Y}{T} \times 100$

Years	*I*	*II*	*III*	*IV*
1982	102.36	91.99	89.46	98.15
1983	110.21	102.86	103.62	111.02
1984	116.86	99.24	94.92	88.81
1985	103.27	95.40	89.16	102.20
Total	432.70	389.49	377.16	400.18
Average	108.18	97.37	94.29	100.05
S.I.	108.20	97.40	94.32	100.08

Note: Grand Average, $G = \frac{399.89}{4} = 99.97$

Notes

Example: Find seasonal variations by the ratio to trend method, from the following data:

Year	*I - Qr*	*II - Qr*	*III - Qr*	*IV - Qr*
1975	30	40	36	34
1976	34	52	50	44
1977	40	58	54	48
1978	54	76	68	62
1979	80	92	86	82

Solution:

First we fit a linear trend to the annual totals.

Years	*Annual Totals (Y)*	*X*	*XY*	X^2
1975	140	- 2	- 280	4
1976	180	- 1	- 180	1
1977	200	0	0	0
1978	260	1	260	1
1979	340	2	680	4
Total	1120	0	480	10

Now a = $\frac{1120}{5}$ = 224 and b = $\frac{480}{10}$ = 48

∴ The trend equation is Y = 224 + 48X, origin : 1st July 1977, unit of X = 1 year.

The quarterly trend equation is Y = $\frac{224}{4}+\frac{48}{16}$X = 56 + 3X, origin : 1st July 1977, unit of X = 1 quarter.

Shifting the origin to III quarter of 1977, we get

Y = 56 + 3(X + $\frac{1}{2}$) = 57.5 + 3X

Table of Quarterly Trend Values

Year	*I*	*II*	*III*	*IV*
1975	27.5	30.5	33.5	36.5
1976	39.5	42.5	45.5	48.5
1977	51.5	54.5	57.5	60.5
1978	63.5	66.5	69.5	72.5
1979	75.5	78.5	81.5	84.5

Ratio to Trend Values

Year	*I*	*II*	*III*	*IV*
1975	109.1	131.1	107.5	93.2
1976	86.1	122.4	109.9	90.7
1977	77.7	106.4	93.9	79.3
1978	85.0	114.3	97.8	85.5
1979	106.0	117.2	105.5	97.0
Total	463.9	591.4	514.6	445.7
A_i	92.78	118.28	102.92	89.14
S.I.	92.10	117.35	102.11	88.44

Note that the Grand Average G = $\frac{403.12}{4}$ = 100.78. Also check that the sum of indices is 400.

Remarks: If instead of multiplicative model we have an additive model, then Y = T + S + R or S + R = Y – T. Thus, the trend values are to be subtracted from the Y values. Random component is then eliminated by the method of simple averages.

Merits and Demerits

Notes

It is an objective method of measuring seasonal variations. However, it is very complicated and doesn't work if cyclical variations are present.

10.6.3 Ratio to Moving Average Method

The ratio to moving average is the most commonly used method of measuring seasonal variations. This method assumes the presence of all the four components of a time series. Various steps in the computation of seasonal indices are as follows:

1. Compute the moving averages with period equal to the period of seasonal variations. This would eliminate the seasonal component and minimise the effect of random component. The resulting moving averages would consist of trend, cyclical and random components.
2. The original values, for each quarter (or month) are divided by the respective moving average figures and the ratio is expressed as a percentage, i.e., $\frac{Y}{M.A} = \frac{TCSR}{TCR'} = SR''$, where R′ and R′′ denote the changed random components.
3. Finally, the random component R′′ is eliminated by the method of simple averages.

Example: Given the following quarterly sale figures, in thousand of rupees, for the year 1986-1989, find the specific seasonal indices by the method of moving averages.

Year	*I*	*II*	*III*	*IV*
1986	34	33	34	37
1987	37	35	37	39
1988	39	37	38	40
1989	42	41	42	44

Solution:

Calculation of Ratio to Moving Averages

Year / Quarter	Sales	4 – Period Moving Total	Centred Total	4 Period M	$\frac{Y}{M} \times 100$
1986 I	34		…	…	…
II	33 →		…	…	…
III	34 →	138 →	279	34.9	97.4
IV	37 →	141 →	284	35.5	104.2
1987 I	37 →	143 →	289	36.1	102.5
II	35 →	146 →	294	36.8	95.1
III	37 →	148 →	298	37.3	99.2
IV	39 →	150 →	302	37.8	103.2
1988 I	39 →	152 →	305	38.1	102.4
II	37 →	153 →	307	38.4	96.4
III	38 →	154 →	311	38.9	97.7
IV	40 →	157 →	318	39.8	100.5
1989 I	42 →	161 →	326	40.8	102.9
II	41 →	165 →	334	41.8	98.1
III	42	169	…	…	…
IV	44		…	…	…

Notes

Calculation of Seasonal Indices

Year	*I*	*II*	*III*	*IV*
1986	-	-	97.4	104.2
1987	102.5	95.1	99.2	103.2
1988	102.4	96.4	97.7	100.5
1989	102.9	98.1	-	-
Total	307.8	289.6	294.3	307.9
A_i	102.6	96.5	98.1	102.6
S.I.	102.7	96.5	98.1	102.7

Note that the Grand Average G =** = 99.95. Also check that the sum of indices is 400.

Merits and Demerits

This method assumes that all the four components of a time series are present and, therefore, widely used for measuring seasonal variations. However, the seasonal variations are not completely eliminated if the cycles of these variations are not of regular nature. Further, some information is always lost at the ends of the time series.

10.6.4 Link Relatives Method

This method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern. As discussed in earlier unit, the link relatives are percentages of the current period (quarter or month) as compared with previous period. With the computation of link relatives and their average, the effect of cyclical and random component is minimised. Further, the trend gets eliminated in the process of adjustment of chained relatives. The following steps are involved in the computation of seasonal indices by this method:

1. Compute the link relative (L.R.) of each period by dividing the figure of that period with the figure of previous period. For example, link relative of 3rd quarter

$$= \frac{\text{figure of 3rd quarter}}{\text{figure of 2nd quarter}} \times 100$$

2. Obtain the average of link relatives of a given quarter (or month) of various years. A.M. or M_d can be used for this purpose. Theoretically, the later is preferable because the former gives undue importance to extreme items.

3. These averages are converted into chained relatives by assuming the chained relative of the first quarter (or month) equal to 100. The chained relative (C.R.) for the current period (quarter or month)

$$= \frac{\text{C.R. of the previous period} \times \text{L.R. of the current period}}{100}$$

4. Compute the C.R. of first quarter (or month) on the basis of the last quarter (or month). This is given by

$$= \frac{\text{C.R. of last quarter (or month)} \times \text{average L.R. of 1st quarter (or month)}}{100}$$

This value, in general, be different from 100 due to long term trend in the data. The chained relatives, obtained above, are to be adjusted for the effect of this trend. The adjustment factor is

Notes

$d = \frac{1}{4}$ [New C.R. for 1st month – 100] for quarterly data

and d = $\frac{1}{12}$ [New C.R. for 1st month – 100] for monthly data.

On the assumption that the trend is linear, d, 2d, 3d, etc., is respectively subtracted from the 2nd, 3rd, 4th, etc., quarter (or month).

5. Express the adjusted chained relatives as a percentage of their average to obtain seasonal indices.

6. Make sure that the sum of these indices is 400 for quarterly data and 1200 for monthly data.

Example: Determine the seasonal indices from the following data by the method of link relatives:

Year	1*st Qr*	2*nd Qr*	3*rd Qr*	4*th Qr*
1985	26	19	15	10
1986	36	29	23	22
1987	40	25	20	15
1988	46	26	20	18
1989	42	28	24	21

Solution:

Calculation Table

Year	*I*	*II*	*III*	*IV*
1985	-	73.1	78.9	66.7
1986	360.0	80.5	79.3	95.7
1987	181.8	62.5	80.0	75.0
1988	306.7	56.5	76.9	90.0
1989	233.3	66.7	85.7	87.5
Total	1081.8	339.3	400.8	414.9
Mean	270.5	67.9	80.2	83.0
C.R.	100.0	67.9	54.5	45.2
C.R.(adjusted)	100.0	62.3	43.3	28.4
S.I.	170.9	106.5	74.0	48.6

The chained relative (C.R.) of the 1st quarter on the basis of C.R. of the 4th quarter = $\frac{270.5 \times 45.2}{100} = 122.3$

The trend adjustment factor d = $\frac{1}{4}(122.3 - 100) = 5.6$

Thus, the adjusted C.R. of 1st quarter = 100

and for 2nd = 67.9 – 5.6 = 62.3

for 3rd = 54.5 – 2 × 5.6 = 43.3

for 4th = 45.2 – 3 × 5.6 = 28.4

The grand average of adjusted C.R., G = $\frac{100 + 62.3 + 43.3 + 28.4}{4} = 58.5$

Notes

$$\text{The seasonal index of a quarter} = \frac{\text{Adjusted C.R.} \times 100}{G}$$

Merits and Demerits

This method is less complicated than the ratio to moving average and the ratio to trend methods. However, this method is based upon the assumption of a linear trend which may not always hold true.

Remarks: Looking at the merits and demerits of various methods of measuring seasonal variations, we find that the ratio to moving average method is most general and, therefore, most popular method of measuring seasonal variations.

Self Assessment

Fill in the blanks:

16. If the time series data are in terms of annual figures, the ……are absent.
17. ……..method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern.
18. …….method is used when cyclical variations are absent from the data

10.7 Summary

- Many types of data are collected over time.
- Stock prices, sales volumes, interest rates, and quality measurements are typical examples.
- Because of the sequential nature of the data, special statistical techniques that account for the dynamic nature of the data are required.
- A time series is a sequence of data points, measured typically at successive times, spaced at time intervals.
- Time series analysis comprises methods that attempt to understand such time series, often either to understand the underlying context of the data points, or to make forecasts.
- Time series forecasting is the use of a model to forecast future events based on known past events: to forecast future data points before they are measured.
- seasonal variations are likely to be present in data recorded on quarterly or monthly or weekly or daily or hourly basis.
- There are four methods commonly used for the measurement of seasonal variations. They are: Method of Simple Averages, Ratio to Trend Method, Ratio to Moving Average Method and Method of Link Relatives

10.8 Keywords

Mean Squared Error: It is the sum of the squared forecast errors for each of the observations divided by the number of observations.

Period of Oscillation: The time interval between the variations is known as the period of oscillation.

Notes

Periodic Variations: The variations that repeat themselves after a regular interval of time.

Random Variations: The variations that do not reveal any regular pattern of movements.

Secular Trend: It is the general tendency of the data to increase or decrease or stagnate over a long period of time.

10.9 Review Questions

1. Smart Discount Stores: There are 2117 Smart stores in the India (the chain is building up). It is one of India's most interesting discounters tracing its origins back to 1980's and the opening of the first Smart store. At present Smart has reached an "upgrading" phase like so many discounters before.

   Given the data below, perform the indicated analyses.

Year	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
Earnings Per Share	19.0	17.5	20.7	28.4	27.4	23.9	21.1	16.1	8.5	11.1
Dividends Per Share	9.9	9.5	9.0	8.1	6.8	5.0	3.0	2.4	2.2	1.9
Pre-tax Margin	2.1	2.0	3.1	4.9	5.4	5.7	5.8	5.8	3.3	5.3

   (a) To what extent does the Board of directors regard dividend payments as a function of earnings? Test whether there is a significant relationship between the variables. Use a parametric analysis.

   (b) Find the linear forecasting equation that would allow you to predict dividend payments based on earnings and test the significance of the slope.

   (c) Is there a significant difference in pre-tax margin when comparing the periods 1995-1999 and 1990-1994. Perform a non-parametric analysis. Explain the managerial implications of your findings.

2. Big and Small Apples Employment figures in thousands for Neo-Classical City and suburbs are given below. Perform the required analyses:

   (a) Using linear forecasts, predict the year in which employment will be the same for the two locations.

   (b) Construct the NCC confidence interval for that year.

   (c) Correlate the employment figures for the two areas using both parametric and non-parametric methods and test the significance of the correlation coefficients.

   (d) Fit a modified exponential trend to SUB data and discuss the results in terms of your findings in (a) above.

   (e) Are NCC employment figures uniformly distributed over the period 1994 through 2000?

YEAR	1994	1995	1996	1997	1998	1999	2000
NYC	64.1	60.2	59.2	59.0	57.6	54.4	50.9
SUB	20.7	21.4	22.1	23.8	24.5	26.3	26.5

Notes

3. What are the normal equations of parabola fit?

4. Determine the seasonal indices from the following data by the method of link relatives:

Year	I Qtr	II Qtr	III Qtr	IV Qtr
2000	42	44	40	38
2001	36	38	36	34
2002	48	50	48	40
2003	38	42	40	38

5. Find seasonal variations by the ratio to trend method, from the following data:

Year	I Qtr	II Qtr	III Qtr	IV Qtr
2000	40	44	42	40
2001	36	40	38	36
2002	48	52	46	42
2003	38	42	40	38

6. Assuming that trend and cyclical variations are absent, compute the seasonal index for each month of the following data of sales (in ₹ '000) of a company:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	38	36	36	38	34	34	32	30	32	34	32	32
2002	34	32	32	32	30	34	30	32	30	34	30	34
2003	32	30	30	34	32	32	34	34	32	36	34	36

7. Why does minimizing the MSE remain a key criterion in selection estimators?

8. How would you estimate the scale parameter of distributions for which the variance and standard deviation do not exist?

9. What will be the effect on the moving averages if the trend is non-linear?

10. "Secular trend is that irreversible movement which continues, in general, in the same direction for a considerable period of time." Comment.

11. Fit a trend line to the following data by method of semi average and forecast the sales for 2007.

Year	Sales of the company in thousands of units
2000	105
2001	106
2002	116
2003	120
2004	114
2005	128
2006	134

12. Using the 3 yearly moving averages, determine the trend values and also the short-term error.

Notes

Year	Production of steel in 000 Tons
2003	30
2004	31
2005	32
2006	34
2007	33
2008	29

13. Seasonal index of the sale of woolen garments in a store is as follows:

Quarter	Seasonal Index
January to March	90
April to June	80
July to September	71
October to December	120

If the total sales of woolen garments in the 1st Quarter is worth ₹ 45,000 how much worth of woolen garments should be kept in stock to meet the demand for remaining quarters, anticipating winter season.

14. An agricultural cooperative society wants to measure variation in its member farmer's wheat harvest over a period of 8 years. The volume harvested actually each year is given as below.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Actual Production in (000) quintals	7.5	7.8	8.2	8.2	8.4	8.5	8.7	9.1

Answers: Self Assessment

1. long-term
2. quantified
3. seasonal influences
4. secular trend
5. Seasonal Variations
6. short-term
7. historical
8. Exponential smoothing
9. Moving average
10. median, mean
11. standard deviation
12. variance
13. Mean Squared Error
14. true value
15. error
16. seasonal variations
17. Link relatives
18. Ratio to trend

10.10 Further Readings

Books

Allan & Blumon, *Elementary Statistics: A Step by Step Approach*, McGraw-Hill College, June 2003.

Notes

David & Moae, *Introduction to the Practice of Statistics*, W.H. Freeman & Co., February 2005.

James T. McClave Terry Sincich, William Mendenhall, *Statistics*, Prentice Hall, February 2005.

Mario F. Triola, *Elementary Statistics*, Addison-Wesley, January 2006.

Mark L. Berenson, David M. Revine, Tineothy C. Krehbiel, *Basic Business Statistics: Concepts & Applications*, Prentice Hall, May 2005.

Unit 11: Index Numbers

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Define the conception of index numbers
- Discuss the uses of index numbers
- Describe the construction of index numbers
- Recognize the thought of consumer price index number
- Identify the problems in the construction of index numbers

Introduction

An index number is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations. Suppose that we want to compare the average price level of different items of food in 1992 with what it was in 1990. Let the different items of food be wheat, rice, milk, eggs, ghee, sugar, pulses, etc. If the prices of all these items change in the same ratio and in the same direction; assume that prices of all the items have increased by 10% in 1992 as compared with their prices in 1990; then there will be no difficulty in finding out the average change in price level for the group as a whole. Obviously,

Notes

the average price level of all the items taken as a group will also be 10% higher in 1992 as compared with prices of 1990. However, in real situations, neither the prices of all the items change in the same ratio nor in the same direction, i.e., the prices of some commodities may change to a greater extent as compared to prices of other commodities. Moreover, the price of some commodities may rise while that of others may fall. For such situations, the index numbers are very useful device for measuring the average change in prices or any other characteristics like quantity, value, etc. for the group as a whole.

11.1 Definitions and Characteristics of Index Numbers

Some important definitions of index numbers are given below:

1. "An index number is a device for comparing the general level of magnitude of a group of distinct, but related, variables in two or more situations."

   — *Karmel and Polasek*

2. "An index number is a special type of average that provides a measurement of relative changes from time to time or from place to place."

   — *Wessell, Wilett and Simone*

3. "Index number shows by its variation the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice."

   — *Edgeworth*

4. "An index number is a single ratio (usually in percentage) which measures the combined (i.e., averaged) change of several variables between two different times, places or situations."

   — *Tuttle*

On the basis of the above definitions, the following characteristics of index numbers are worth mentioning:

1. ***Index numbers are specialised averages:*** As we know that an average of data is its representative summary figure. In a similar way, an index number is also an average, often a weighted average, computed for a group. It is called a specialised average because the figures, that are averaged, are not necessarily expressed in homogeneous units.

2. ***Index numbers measure the changes for a group which are not capable of being directly measured:*** The examples of such magnitudes are: Price level of a group of items, level of business activity in a market, level of industrial or agricultural output in an economy, etc.

3. ***Index numbers are expressed in terms of percentages:*** The changes in magnitude of a group are expressed in terms of percentages which are independent of the units of measurement. This facilitates the comparison of two or more index numbers in different situations.

Self Assessment

Fill in the blanks:

1. Index numbers are called a specialised average because the figures, that are averaged, are not necessarily expressed in ……..units.

2. Index number is often recognized a …….average, computed for a group.

11.2 Uses of Index Numbers

Notes

The main uses of index numbers are:

1. ***To measure and compare changes:*** The basic purpose of the construction of an index number is to measure the level of activity of phenomena like price level, cost of living, level of agricultural production, level of business activity, etc. It is because of this reason that sometimes index numbers are termed as barometers of economic activity. It may be mentioned here that a barometer is an instrument which is used to measure atmospheric pressure in physics.

   The level of an activity can be expressed in terms of index numbers at different points of time or for different places at a particular point of time. These index numbers can be easily compared to determine the trend of the level of an activity over a period of time or with reference to different places.

2. ***To help in providing guidelines for framing suitable policies:*** Index numbers are indispensable tools for the management of any government or non-government organisation.

Example: The increase in cost of living index is helpful in deciding the amount of additional dearness allowance that should be paid to the workers to compensate them for the rise in prices. In addition to this, index numbers can be used in planning and formulation of various government and business policies.

3. ***Price index numbers are used in deflating:*** This is a very important use of price index numbers. These index numbers can be used to adjust monetary figures of various periods for changes in prices.

Example: The figure of national income of a country is computed on the basis of the prices of the year in question. Such figures, for various years often known as national income at current prices, do not reveal the real change in the level of production of goods and services. In order to know the real change in national income, these figures must be adjusted for price changes in various years. Such adjustments are possible only by the use of price index numbers and the process of adjustment, in a situation of rising prices, is known as deflating.

4. ***To measure purchasing power of money:*** We know that there is inverse relation between the purchasing power of money and the general price level measured in terms of a price index number. Thus, reciprocal of the relevant price index can be taken as a measure of the purchasing power of money.

Self Assessment

Fill in the Blanks:

3. Index numbers are termed as …….of economic activity.

4. The …….can be expressed in terms of index numbers at different points of time or for different places at a particular point of time.

11.3 Construction of Index Numbers

To illustrate the construction of an index number, we reconsider various items of food mentioned earlier. Let the prices of different items in the two years, 1990 and 1992, be as given below:

Notes

Item	Price in 1990 (in Rs/unit)	Price in 1992 (in Rs/unit)
1. Wheat	300/quintal	360/quintal
2. Rice	12/kg.	15/kg.
3. Milk	7/litre	8/litre
4. Eggs	11/dozen	12/dozen
5. Ghee	80/kg.	88/kg.
6. Sugar	9/kg.	10/kg.
7. Pulses	14/kg.	16/kg.

The comparison of price of an item, say wheat, in 1992 with its price in 1990 can be done in two ways, explained below:

1. By taking the difference of prices in the two years, i.e., 360 - 300 = 60, one can say that the price of wheat has gone up by ₹ 60/quintal in 1992 as compared with its price in 1990.

2. By taking the ratio of the two prices, i.e., $\frac{360}{300}$ = 1.20, one can say that if the price of wheat in 1990 is taken to be 1, then it has become 1.20 in 1992. A more convenient way of comparing the two prices is to express the price ratio in terms of percentage, i.e., $\frac{360}{300} \times 100 = 120$, known as Price Relative of the item. In our example, price relative of wheat is 120 which can be interpreted as the price of wheat in 1992 when its price in 1990 is taken as 100. Further, the figure 120 indicates that price of wheat has gone up by 120 – 100 = 20% in 1992 as compared with its price in 1990.

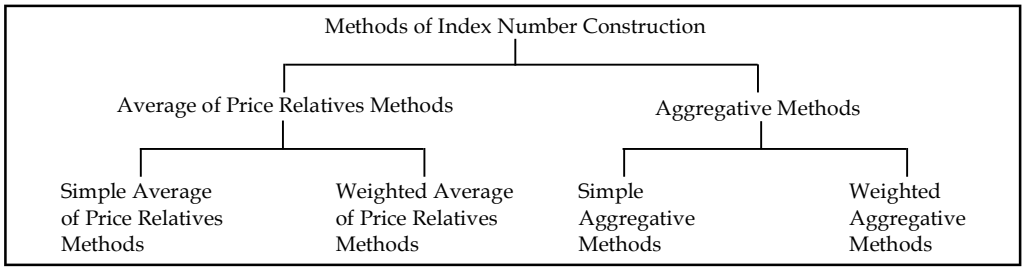
The first way of expressing the price change is inconvenient because the change in price depends upon the units in which it is quoted. This problem is taken care of in the second method, where price change is expressed in terms of percentage. An additional advantage of this method is that various price changes, expressed in percentage, are comparable. Further, it is very easy to grasp the 20% increase in price rather than the increase expressed as ₹ 60/quintal.

For the construction of index number, we have to obtain the average price change for the group in 1992, usually termed as the Current Year, as compared with the price of 1990, usually called the Base Year. This comparison can be done in two ways:

1. By taking suitable average of price relatives of different items. The methods of index number construction based on this procedure are termed as Average of Price Relative Methods.

2. By taking ratio of the averages of the prices of different items in each year. These methods are popularly known as Aggregative Methods.

Since the average in each of the above methods can be simple or weighted, these can further be divided as simple or weighted. Various methods of index number construction can be classified as shown below:



In addition to this, a particular method would depend upon the type of average used. Although, geometric mean is more suitable for averaging ratios, arithmetic mean is often preferred because of its simplicity with regard to computations and interpretation.

Notes

Notes Before writing various formulae of index numbers, it is necessary to introduce certain notations and terminology for convenience.

Base Year: The year from which comparisons are made is called the base year. It is commonly denoted by writing '0' as a subscript of the variable.

Current Year: The year under consideration for which the comparisons are to be computed is called the current year. It is commonly denoted by writing '1' as a subscript of the variable.

Let there be n items in a group which are numbered from 1 to n. Let p_{0i} denote the price of the i th item in base year and p_{1i} denote its price in current year, where i = 1, 2, ... n. In a similar way q_{0i} and q_{1i} will denote the quantities of the i th item in base and current years respectively.

Using these notations, we can write an expression for price relative of the ith item as

$P_i = \frac{p_{1i}}{p_{0i}} \times 100$ and quantity relative of the i th item as $Q_i = \frac{q_{1i}}{q_{0i}} \times 100$.

Further, P_{01} will be used to denote the price index number of period '1' as compared with the prices of period '0'. Similarly, Q_{01} and V_{01} would denote the quantity and the value index numbers respectively of period '1' as compared with period '0'.

Self Assessment

Fill in the blanks:

5. The year from which comparisons are made is called the…….

6. ……is commonly denoted by writing '1' as a subscript of the variable

11.4 Price Index Numbers

Simple Average of Price Relatives

1. When arithmetic mean of price relatives is used

   The index number formula is given by $P_{01} = \frac{\sum P_i}{n}$ or $P_{01} = \frac{\sum \frac{p_{1i}}{p_{0i}} \times 100}{n}$ Omitting the subscript i, the above formula can also be written as $P_{01} = \frac{\sum \frac{p_1}{p_0} \times 100}{n}$

2. When geometric mean of price relatives is used

   The index number formula is given by

Notes

$$P_{01} = \left(P_1 \times P_2 \times ..... P_n)^{\frac{1}{n}} \right) = \left(\prod_{i=1}^{n} P_i \right)^{\frac{1}{n}} = \text{Antilog} \left[\frac{\sum \log P_i}{n} \right]$$

(Π is used to denote the product of terms.)

Example: Given below are the prices of 5 items in 1985 and 1990. Compute the simple price index number of 1990 taking 1985 as base year. Use (a) arithmetic mean and (b) geometric mean.

Item	Price in1985 (Rs/unit)	Price in1990 (Rs/unit)
1	15	20
2	8	7
3	200	300
4	60	110
5	100	130

Solution:

Calculation Table

Item	Price in 1985 (P_{0i})	Price in 1990 (P_{0i})	Price Relative $P_i = \frac{p_{1i}}{p_{0i}} \times 100$	log P_i
1	15	20	133.33	2.1249
2	8	7	87.50	1.9420
3	200	300	150.00	2.1761
4	60	110	183.33	2.2632
5	100	130	130.00	2.1139
Total			684.16	10.6201

∴ Index number, using A.M., is $P_{01} = \frac{684.16}{5} = 136.83$ and Index number, using G.M., is

$$P_{01} = \text{Antilog}\left[\frac{10.6201}{5} \right] = 133.06$$

Weighted Average of Price Relatives

In the method of simple average of price relatives, all the items are assumed to be of equal importance in the group. However, in most of the real life situations, different items of a group have different degree of importance. In order to take this into account, weighing of different items, in proportion to their degree of importance, becomes necessary.

Let wi be the weight assigned to the i th item (i = 1, 2, ... n). Thus, the index number, given by the weighted arithmetic mean of price relatives, is $P_{01} = \frac{\sum P_i w_i}{\sum w_i}$.

Notes

Similarly, the index number, given by the weighted geometric mean of price relatives can be written as follows:

$$P_{01} = \left[P_1^{w_1} . P_2^{w_2} \ldots\ldots P_n^{w_n} \right]^{\frac{1}{\sum w_i}} = \left[\prod P_i^{w_i} \right]^{\frac{1}{\sum w_i}} \text{ or } P_{01} = \text{Antilog} \left[\frac{\sum w_i \log P_i}{\sum w_i} \right]$$

Nature of Weights

While taking weighted average of price relatives, the values are often taken as weights. These weights can be the values of base year quantities valued at base year prices, i.e., $p_{0i}q_{0i}$, or the values of current year quantities valued at current year prices, i.e., $p_{1i}q_{1i}$, or the values of current year quantities valued at base year prices, i.e., $p_{0i}q_{1i}$, etc., or any other value.

Example: Construct an index number for 1989 taking 1981 as base for the following data, by using

1. Weighted arithmetic mean of price relatives and
2. Weighted geometric mean of price relatives.

Commodities	Prices in 1981	Prices in 1989	Weights
A	60	100	30
B	20	20	20
C	40	60	24
D	100	120	30
E	120	80	10

Solution:

Calculation Table

Item	Price in 1985 (P_{0i})	Price in 1990 (P_{0i})	Price Relative $P_i = \frac{p_{1i}}{p_{0i}} \times 100$	$\log P_i$
1	15	20	133.33	2.1249
2	8	7	87.50	1.9420
3	200	300	150.00	2.1761
4	60	110	183.33	2.2632
5	100	130	130.00	2.1139
Total			684.16	10.6201

$\therefore$ Index number using A.M. is $P_{01} = \frac{14866.8}{114} = 130.41$ and index number using G.M. is

$$P_{01} = \text{Antilog}\left[\frac{239.498}{114} \right] = 126.15$$

Notes

Task Taking 1983 as base year, calculate an index number of prices for 1990, for the following data given in appropriate units, using:

1. Weighted arithmetic mean of price relatives by taking weights as the values of current year quantities at base year prices, and
2. Weighted geometric mean of price relatives by taking weights as the values of base year quantities at base year prices.

Commodity	1983		1990	
	Price	*Quantity*	*Price*	*Quantity*
A	82	63	160	56
B	80	75	182	53
C	105	92	185	64
D	102	25	177	13
E	102	63	175	54
F	190	61	140	60

Simple Aggregative Method

In this method, the simple arithmetic mean of the prices of all the items of the group for the current as well as for the base year are computed separately. The ratio of current year average to base year average multiplied by 100 gives the required index number.

Using notations, the arithmetic mean of prices of n items in current year is given by $\frac{\sum p_{1i}}{n}$ and the arithmetic mean of prices in base year is given by $\frac{\sum p_{0i}}{n}$

$\therefore$ Simple aggregative price index $P_{01} = \frac{\frac{\sum p_{1i}}{n}}{\frac{\sum p_{0i}}{n}} \times 100 = \frac{\sum p_{1i}}{\sum p_{0i}} \times 100$

Omitting the subscript i, the above index number can also be written as $P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$

Example: The following table gives the prices of six items in the years 1980 and 1981. Use simple aggregative method to find index of 1981 with 1980 as base.

Item	Price in 1980 (₹)	Price in 1981 (₹)
A	40	50
B	60	60
C	20	30
D	50	70
E	80	90
F	100	100

Notes

Solution:

Let p_0 be the price in 1980 and p_1 be the price in 1981. Thus, we have

$\Sigma p_0 = 350$ and $\Sigma p_1 = 400$

$$\therefore P_{01} = \frac{400}{350} \times 100 = 114.29$$

Weighted Aggregative Method

This index number is defined as the ratio of the weighted arithmetic means of current to base year prices multiplied by 100.

Using the notations, defined earlier, the weighted arithmetic mean of current year prices can be written as = $\frac{\sum p_{1i} w_i}{\sum w_i}$

Similarly, the weighted arithmetic mean of base year prices $= \frac{\sum p_{0i} w_i}{\sum w_i}$

$$\therefore \text{ Price Index Number, } P_{01} = \frac{\frac{\sum p_{1i} w_i}{\sum w_i}}{\frac{\sum p_{0i} w_i}{\sum w_i}} \times 100 = \frac{\sum p_{1i} w_i}{\sum p_{0i} w_i} \times 100$$

Omitting the subscript, we can also write $P_{01} = \frac{\sum p_1 w}{\sum p_0 w} \times 100$

Nature of Weights

In case of weighted aggregative price index numbers, quantities are often taken as weights. These quantities can be the quantities purchased in base year or in current year or an average of base year and current year quantities or any other quantities. Depending upon the choice of weights, some of the popular formulae for weighted index numbers can be written as follows:

1. ***Laspeyres' Index:*** Laspeyres' price index number uses base year quantities as weights. Thus, we can write

$$P_{01}^{La} = \frac{\sum p_{1i} q_{0i}}{\sum p_{0i} q_{0i}} \times 100 \text{ or } P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

2. ***Paasche's Index:*** This index number uses current year quantities as weights. Thus, we can write

$$P_{01}^{Pa} = \frac{\sum p_{1i} q_{1i}}{\sum p_{0i} q_{1i}} \times 100 \text{ or } P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

3. ***Fisher's Ideal Index:*** As will be discussed later that the Laspeyres's Index has an upward bias and the Paasche's Index has a downward bias. In view of this, Fisher suggested that an ideal index should be the geometric mean of Laspeyres' and Paasche's indices. Thus, the Fisher's formula can be written as follows:

Notes

$$P_{01}^{F} = \sqrt{P_{01}^{La} \times P_{01}^{Pa}} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

If we write L = $\frac{\sum p_1 q_0}{\sum p_0 q_0}$ and P = $\frac{\sum p_1 q_1}{\sum p_0 q_1}$, the Fisher's Ideal Index can also be written as

$P_{01} = \sqrt{L \times P} \times 100$.

4. ***Dorbish and Bowley's Index:*** This index number is constructed by taking the arithmetic mean of the Laspeyres's and Paasche's indices.

$$P_{01}^{DB} = \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 + \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100\right] = \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100 = \frac{1}{2}[L \times P] \times 100$$

5. ***Marshall and Edgeworth's Index:*** This index number uses arithmetic mean of base and current year quantities.

$$P_{01}^{ME} = \frac{\sum p_1\left(\frac{q_0 + q_1}{2}\right)}{\sum p_0\left(\frac{q_0 + q_1}{2}\right)} \times 100 = \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} \times 100 = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

6. ***Walsh's Index:*** Geometric mean of base and current year quantities are used as weights in this index number.

$$P_{01}^{Wa} = \frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100$$

7. ***Kelly's Fixed Weights Aggregative Index:*** The weights, in this index number, are quantities which may not necessarily relate to base or current year. The weights, once decided, remain fixed for all periods. The main advantage of this index over Laspeyres's index is that weights do not change with change of base year. Using symbols, the Kelly's Index can be written as

$$P_{01}^{Ke} = \frac{\sum p_1 q}{\sum p_0 q} \times 100$$

Example: Calculate the weighted aggregative price index for 1990 from the following data :

Item	Price in 1971	Price in 1990	Weights
A	8	9.5	5
B	12	12.5	1
C	6.5	9	3
D	4	4.5	6
E	6	7	4
F	2	4	3

Solution:

Notes

Calculation Table

Item	Price in 1971 (p_0)	Price in 1990 (p_1)	Weights (w)	p_0w	p_1w
A	8	9.5	5	40.0	47.5
B	12	12.5	1	12.0	12.5
C	6.5	9	3	19.5	27.0
D	4	4.5	6	24.0	27.0
E	6	7	4	24.0	28.0
F	2	4	3	6.0	12.0
Total				125.5	154.0

$\therefore$ Price Index (1971 = 100) $P_{01} = \frac{154.0}{125.5} \times 100 = 122.71$

The term within bracket, i.e., 1971 = 100, indicates that base year is 1971.

11.4.1 Use of Price Index Numbers in Deflating

This is perhaps the most important application of price index numbers. Deflating implies making adjustments for price changes. A rise of price level implies a fall in the value of money. Therefore, in a situation of rising prices, the workers who are getting a fixed sum in the form of wages are in fact getting less real wages. Similarly, in a situation of falling prices, the real wages of the workers are greater than their money wages. Thus, to determine the real wages, the money wages of the workers are to be adjusted for price changes by using relevant price index number.

The following formula is used for conversion of money wages into real wages.

$$\text{Real Wage} = \frac{\text{Money Wage}}{\text{Consumer Price Index}} \times 100 \quad \text{.... (1)}$$

Another application of the process of deflating to find the value of output at constant prices so as to facilitate the comparison of real changes in output. It may be pointed out here that the output of a given year is often valued at the current year prices. Since prices in various years are often different, the comparison of output at current year prices has no relevance.

The output at constant prices is obtained using the following formula.

$$\text{Output at Constant Prices} = \frac{\text{Output at Current Prices}}{\text{Price Index}} \times 100 \quad \text{.... (2)}$$

Example: The following table gives the average monthly wages of a worker along with the respective consumer price index numbers for ten years.

Years	:	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Average monthly wages (₹)	:	500	525	560	600	630	635	700	740	800	900
Consumer Price Index	:	100	110	120	125	135	160	185	200	210	240

Compute his real average monthly wages in various years.

Notes

Solution:

Computation of Real Wages

Years	Average Monthly wage	Consumer Price Index	Real average monthly wage
1980	500	100	$\frac{500}{100} \times 100 = 500.00$
1981	525	110	$\frac{525}{110} \times 100 = 477.27$
1982	560	120	$\frac{560}{120} \times 100 = 466.67$
1983	600	125	$\frac{600}{125} \times 100 = 480.00$
1984	630	135	$\frac{630}{135} \times 100 = 466.67$
1985	635	160	$\frac{635}{160} \times 100 = 396.88$
1986	700	185	$\frac{700}{185} \times 100 = 378.38$
1987	740	200	$\frac{740}{200} \times 100 = 370.00$
1988	800	210	$\frac{800}{210} \times 100 = 380.95$
1989	900	240	$\frac{900}{240} \times 100 = 375.00$

Purchasing Power of Money

When prices in general are rising, the real value of a rupee is declining. If, e.g., the price index in 1992 with base 1990 is 120, the real value of a rupee in 1992 as compared with its value in 1990. This implies that a rupee in 1992 is worth only 83 paise of 1990.

From the above we note that the purchasing power of a rupee in current year is equal to the reciprocal of the price index multiplied by 100. Thus, we can write

Purchasing Power of a Rupee or $\text{Constant Rupee} = \frac{\text{Current Rupee} \times 100}{\text{Price Index}} = \frac{100}{\text{Price Index}}$

Note that the Current Rupee is always equal to unity.

We can also write $\text{Price Index} = \frac{100}{\text{Constant Rupee}}$

Example: Given the following information on the Gross Domestic Product (in ₹ crores) at the constant (1980 - 81) prices and at current prices for five years. Calculate the series of price index numbers and of quantity index numbers for each of the five years with 1980 - 81 as base year.

	G.D.P. at constant (1980-81) Prices	G.D.P. at current Prices
1980-81	200	200
1981-82	150	240
1982-83	125	350
1983-84	120	360
1984-85	160	400

Solution:

Notes

Calculation of Price and Quantity Index Numbers

Year	GDP at constant Prices	GDP at current Prices	Quantity Index Number Series	Price Index* Number Series
1980-81	200	200	100	$\frac{200}{200}\times100=100$
1981-82	150	240	$\frac{150}{200}\times100=75$	$\frac{240}{150}\times100=160$
1982-83	125	350	$\frac{125}{200}\times100=62.5$	$\frac{350}{125}\times100=280$
1983-84	120	360	$\frac{120}{200}\times100=60$	$\frac{360}{120}\times100=300$
1984-85	160	400	$\frac{160}{200}\times100=80$	$\frac{400}{160}\times100=250$

$$\text{Price Index}=\frac{\text{Output at current Prices}}{\text{Output at constant Prices}}\times100$$

Did u know? **What is the usage of deflating?**

The concept of deflating can be used to determine the purchasing power or real value of a rupee.

Self Assessment

Fill in the blanks:

7. In case of weighted aggregative price index numbers, quantities are often taken as……..
8. Deflating implies making adjustments for ……changes.

11.5 Quantity Index Numbers

A quantity index number measures the change in quantities in current year as compared with a base year. The formulae for quantity index numbers can be directly written from price index numbers simply by interchanging the role of price and quantity. Similar to a price relative, we can define a quantity relative as

$$Q=\frac{q_1}{q_0}\times100$$

Various formulae for quantity index numbers are as given below:

1. Simple aggregative index $Q_{01}=\frac{\sum q_1}{\sum q_0}\times100$
2. Simple average of quantity relatives

   (a) Taking A.M. $Q_{01}=\frac{\frac{\sum q_1}{\sum q_0}\times100}{n}=\frac{\sum Q}{n}$

Notes

(b) Taking G.M. $Q_{01} = \text{Antilog}\left[\frac{\sum \log Q}{n}\right]$

3. Weighted aggregative index

(a) $Q_{01}^{La} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$ (base year prices are taken as weights)

(b) $Q_{01}^{Pa} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$ (current year prices are taken as weights)

(c) $Q_{01}^{Fi} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$ Other aggregative formulae can also be written in a similar way.

4. Weighted average of quantity relatives

(a) Taking A.M. $Q_{01} = \frac{\sum Qw}{\sum w}$

(b) Taking G.M. $Q_{01} = \text{Antilog}\left[\frac{\sum w \log Q}{\sum w}\right]$

Like weighted average of price relatives, values are taken as weights.

Example: Using Fisher's formula, the quantity index number from the following data:

Article	1974		1976	
	Price (Rs)	Value (Rs)	Price (Rs)	Value (Rs)
A	5	50	4	48
B	8	48	7	49
C	6	18	5	20

Solution:

Calculation Table

Article	1974			1976			p_0q_1	p_1q_0
	p_0	V_0	$q_0 = \frac{V_0}{p_0}$	p_1	V_1	$q_1 = \frac{V_1}{p_1}$		
A	5	50	10	4	48	12	60	40
B	8	48	6	7	49	7	56	42
C	6	18	3	5	20	4	24	15
Total	$\sum p_0q_0 = 116$			$\sum p_1q_1 = 117$			140	97

$$Q_{01}^{Fi} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100 = \sqrt{\frac{140}{116} \times \frac{117}{97}} = 120.65$$

Self Assessment

Notes

Fill in the blanks:

9. ……..measures the change in quantities in current year as compared with a base year.

10. The formulae for quantity index numbers can be directly written from …….simply by interchanging the role of price and quantity.

11.6 Consumer Price Index Number

The consumer price or the retail price is the price at which the ultimate consumer purchases his goods and services from the retailer. According to the Labour Bureau, *"with the help of Consumer Price Index Number, it is intended to show over time the average change in prices paid by the consumers belonging to the population group proposed to be covered by the index for a fixed list of goods and services consumed by them".*

Formerly, this index was also known as the cost of living index. However, since this index measures changes in cost of living due to changes in retail prices only and not due to changes in living standards, etc., the name was changed to consumer price index or retail price index.

11.6.1 Construction of Consumer Price Index

The following steps are involved in the construction of a consumer price index:

1. ***Scope and Coverage:*** The scope of consumer price index, proposed to be constructed, must be very clearly defined. This implies the identification of the class of people for whom the index will be constructed such as industrial workers, agricultural workers, urban wage earners, etc. Further, it is also necessary to define the coverage of the class of people, i.e., the definition of geographical location of their stay such as a city or two or more villages, etc. The selected class of people should form a homogeneous group so that weights of various commodities are same for all the people.

2. ***Selection of Base Period:*** A normal period having comparative economic stability should be selected as a base period in order that the consumption pattern used in the construction of the index remain practically stable over a fairly long period.

3. ***Conducting Family Budget Enquiry:*** A family budget gives the details of expenditure incurred by the family on various items in a given period. In order to estimate the consumption pattern, a sample survey of family budgets of the group of people, for whom the index is to be constructed, is conducted and from this an average family budget is prepared. The goods and services that are to be included in the construction of the index are selected from this average family budget. Efforts should be made to include as many commodities as possible. Generally the commodities are divided into five broad groups: (i) Food, (ii) Clothing, (iii) Fuel and Lighting, (iv) House Rent and (v) Miscellaneous.

   If necessary, these groups may further be divided into sub-groups. Percentage expenditure of a group is taken as its weight.

4. ***Obtaining Price Quotations:*** The next step in the construction of consumer price index is to obtain the retail price quotations of various items that are selected. The price quotations should be obtained from those markets from which the group of people, for whom the index number is being constructed, normally make purchases. The quality of various goods and services used by the group of people should also be kept in mind while obtaining price quotations.

Notes

5. ***Computation of the Index Number:*** After the collection of necessary data, the consumer price index can be computed by using either of the following formulae.

   (a) ***Aggregate Expenditure Method:*** Base year quantities are taken as weights in the aggregate expenditure method. The formula for the consumer price index is given by $p_{01}^{CP} = \frac{\sum p_1 p_0}{\sum p_0 p_0} \times 100$ which is the Laspeyres's formula.

   (b) ***Family Budget Method*:** This method is also known as weighted average of price relatives method and accordingly values are taken as weights. The formula for the consumer price index is given by $P_{01}^{CP} = \frac{\sum Pw}{\sum w}$, where $p = \frac{p_1}{p_0} \times 100$

Example: From the information given below, construct the consumer price index number of 1985 by (i) Aggregate Expenditure Method, and (ii) Family Budget Method.

Commodities	Quantities (q_0)	Price in 1980 (p_0)	Price in 1985 (p_1)
A	2	75	125
B	25	12	16
C	10	12	16
D	5	10	15
E	25	4.5	5
F	40	10	12
G	1	25	40

Solution:

Calculation of Consumer Price Index

Com.	p_0q_0	p_1q_0	$P = \frac{p_1}{p_0} \times 100$	$w = p_0q_0$	Pw
A	150	250	166.67	150	25000.5
B	300	400	133.33	300	39999.0
C	120	160	133.33	120	15999.6
D	50	75	150.00	50	7500.0
E	112.5	125	111.11	112.5	12499.9
F	400	480	120.00	400	48000.0
G	25	40	160.00	25	4000.0
Total	1157.5	1530		1157.5	152999.0

1. Index by agg. exp. method $\frac{1530}{1157.5} \times 100 = 132.18$

2. Index by F.B. method $\frac{152999}{1157.5} = 132.18$

11.6.2 Uses of Consumer Price Index

1. A consumer price index is used to determine the real wages from money wages and the purchasing power of money.

Notes

2. It is also used to determine the dearness allowance to compensate the workers for the rise in prices.
3. It can be used in the formulation of various economic policies of the government.
4. It may be useful in the analysis of markets of certain goods or services.

Example: A particular series of consumer price index covers five groups of items. Between 1975 and 1980 the index rose from 180 to 225. Over the same period the price index numbers of various groups changed as follows:

Food from 198 to 252; clothing from 185 to 205; fuel and lighting from 175 to 195; miscellaneous from 138 to 212; house rent remained unchanged at 150.

Given that the weights of clothing , house rent and fuel and lighting are equal, determine the weights for individual groups of items.

Solution:

Let w_1% be the weight of food, w_2% be the weight of miscellaneous group and w% be the weight of each of the remaining three groups. Therefore we can write $w_1 + w_2 + 3w = 100$ or $w_2 = 100 - w_1 - 3w$.

The given data can be written in the form of table as given below:

Groups	Weights	Index in 1975 (I_1)	Index in 1980 (I_2)
Food	w_1	198	252
Clothing	w	185	205
Fuel & Lighting	w	175	195
House Rent	w	150	150
Miscellaneous	$100 - w_1 - 3w$	138	212
Total	100		

On the basis of above, the consumer price index of 1975 is

$$\frac{1.98w_1 + (185 + 175 + 150)w + 138(100 - w_1 - 3w)}{100} = 180 \text{ (given)}$$

$$\text{or } 60w_1 + 96w = 4200 \quad \text{.... (1)}$$

Further, the consumer price index of 1980 is

$$= \frac{252w_1 + (205 + 195 + 150)w + 212(100 - w_1 - 3w)}{100} = 225 \text{ (given)}$$

$$\text{or } 40w_1 - 86w = 1300 \quad \text{.... (2)}$$

Solving equations (1) and (2) simultaneously, we get w = 10 and $w_1 = 54$

Substituting these values in expression for w_2, we get

$w_2 = 100 - w_1 - 3w = 100 - 54 - 30 = 16$

Self Assessment

Fill in the blanks:

11. Formerly, Consumer Price index was also known as the ……index.
12. A consumer price index is used to determine the …….from money wages.

Notes

11.7 Problems in the Construction of Index Numbers

The following are some general problems that are faced in the construction of any index number:

1. ***Definition of the purpose:*** Since it is possible to construct index numbers for a number of purposes and one cannot have an all purpose index, therefore, it is very essential to define the specific purpose of its construction. For example, if we are interested in the construction of a price index number, we must have knowledge about the purpose to be served by it, i.e., what is to be measured by it; like the cost of living of workers or the change in wholesale prices, etc. In the absence of this information, it may be difficult to carry out various steps in the construction of an index number. The questions like what are items to be included, from which of the markets the price quotations are to be obtained, what will be the weights of different items, etc., cannot be answered unless the purpose of the index number construction is known. Further, an index number can be of sensitive or general nature. In case of sensitive index, only those items are included whose variables (like prices in case of price index) fluctuate very often; while efforts are made to include as many items as possible when the index is of general nature. It may be pointed out that the index numbers are specialised tools and as such are more useful and efficient when properly used. The first step in this direction is a specific definition of the purpose of its construction.

2. ***Selection of the base period:*** Every index number is constructed with reference to a base period. There are two important points that must be kept in mind while selecting the base period of an index number.

   (a) The base period should correspond to a period of relative economic and political stability, i.e., it should be a normal or representative period in some way. In certain situations where identification of such a period is not possible, the average of certain periods can also be taken as base.

   (b) The comparison of current period with a remote base doesn't have much relevance. In the words of Morris Hamburg, "It is desirable that the base period be not too far away in time from the present. The further away we move from the base period the dimmer are our recollections of economic conditions prevailing at that time. Consequently, comparisons with these remote base periods tend to lose significance and become rather tenuous in meaning".

   Another problem with a remote base period can be that certain items that were in use in the base period are no longer in use while certain new items are in use in current period. In such a situation the two item bundles are no longer homogeneous and comparable. This problem is less likely to occur when fairly recent period is chosen as base.

Caution The base period should not be too distant from the current period.

3. ***Selection of number and type of items:*** An index number of a particular group of items is in fact based on a sample of items taken from it. It is neither possible nor necessary to include all the items of the group in the construction of an index number. The number of items to be included depends largely upon the purpose of the index number.

   There are no hard and fast rules that can be laid down with regard to the selection of the number of items, however, it must be remembered that more is the number of items the more representative will be the index number and more cumbersome will be the task of computations. Therefore, it is necessary to have some sort of balance between having a representative index and the work of computation involved in its construction.

Notes

The following points should be kept in mind in selecting the type of items:

(a) The items should be representative of the tastes, habits and customs of the people for whom the index is to be constructed.

(b) The selected items should be of stable quality. The standardised items should be given preference.

(c) As far as possible, the non-tangible items like personal services, goodwill, etc., should be excluded because it is difficult to ascertain their value.

4. ***Collection of data:*** The next important step in the construction of an index number is the collection of data. For example, for the construction of price index, price quotations are to be obtained. Since the prices of commodities may vary from one market to another and in certain cases from one shop to another, it is necessary to select those markets which are representative in the sense that the group under consideration generally make purchases from these markets. The next logical step is to select an agency through which price quotations are to be obtained. The selected agency should be highly reliable and if necessary the accuracy of price quotations reported by it may also be checked by appointing some other agency or agencies. Furthermore, care should always be taken to obtain price quotations for the same quality of items.

   Similar type of considerations are necessary for the collection of data for the construction of index numbers such as quantity index, value index, unemployment index, etc.

5. ***Selection of a suitable average:*** Since the index numbers are also averages, any of the five averages, viz. arithmetic mean, median, mode, geometric mean and harmonic mean can be used in its construction. However, since in most of the situations we have to average ratios of the values in current period to that in base period, geometric mean is the most suitable average in the construction of index numbers. The main difficulty of using the geometric mean is the complexities of its computations and hence, the use of arithmetic mean is more popular in spite of its being less suitable.

6. ***Selection of suitable weights:*** According to John I. Griffin, "Weighing is designed to give component series an importance in proper relation to their real significance." The basic purpose of weighing is to enable each item to have an influence, on the index number, in proportion to its importance in the group. It is, therefore, necessary to design a system of weighing such that true importance of the items is reflected by it. The system of weighing may be either arbitrary or rational. Arbitrary or chance weighing implies that the statistician is free to assign weights to different items as he thinks fit or reasonable. Rational or logical weighing, on the other hand, implies that some criterion has been fixed for assigning weights. Two types of weights are commonly used in the construction of a price index number: (i) physical quantities and (ii) money values. These weights can be quantities (or values) produced or consumed or sold in base or current or in any other period.

   Another problem, to be tackled, with regard to system of weights is whether weights should be fixed or fluctuating. When relative importance of various items change in different periods, it is desirable to have fluctuating system of weights to get better results.

Self Assessment

Fill in the blanks:

13. Every index number is constructed with reference to a ……..period.

14. The basic purpose of ……..is to enable each item to have an influence, on the index number, in proportion to its importance in the group.

Notes

11.8 Limitations of Index Numbers

Despite the fact that index numbers are very useful for the measurement of relative changes, these suffer from the following limitations:

1. The computation of an index number is based on the data obtained from a sample, which may not be a true representative of the universe.
2. The composition of the bundle of commodities may be for different years. This cannot be taken into account by the fixed base method. Although this difficulty can be overcome by the use of chain base index numbers, but their calculations are quite cumbersome.
3. An index number doesn't take into account the quality of the items. Since a superior item generally has a higher price and the increase in index may be due to an improvement in the quality of the items and not due to rise of prices.
4. Index numbers are specialised averages and as such these also suffer from all the limitations of an average.
5. An index number can be computed by using a number of formulae and different formulae will give different results. Unless a proper method is used, the results are likely to be inaccurate and misleading.
6. By the choice of a wrong base period or weighing system, the results of the index number can be manipulated and, thus, are likely to be misused.

Self Assessment

Fill in the blanks:

15. An index number doesn't take into account the ……..of the items.
16. Index number computed by using a number of formulae will give …….results

11.9 Summary

- An index number is a device for comparing the general level of magnitude of a group of distinct, but related, variables in two or more situations
- Simple Average of Price Relatives Index

$$P_{01} = \frac{\sum \frac{p_1}{p_0}}{n} \quad \text{(using A.M.)}$$

$$P_{01} = Antilog\left[\frac{\sum \log \frac{p_1}{p_0} \times 100}{n}\right] \quad \text{(using G.M.)}$$

- Simple Aggregative Index $P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$
- Weighted Average of Price relatives Index

$$P_{01} = \frac{\sum Pw}{\sum w} \quad \text{(using weighted A.M.)}$$

Notes

$$P_{01} = Antilog\left[\frac{\sum w \log P}{\sum w}\right]$$ (using weighted G.M.)

Here $P = \frac{p_1}{p_0} \times 100$ and w denotes values (weights)

- Weighted Aggregative Index Numbers

(a) Laspeyres's Index $P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

(b) Paasche's Index $P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$

(c) Fisher's Ideal Index $P_{01}^{Fi} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$

(d) Dorbish and Bowley's Index $P_{01}^{DB} = \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100$

(e) Marshall and Edgeworth Index $P_{01}^{ME} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_0} \times 100$

(f) Walsh's Index $P_{01}^{Wa} = \frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100$

(g) Kelly's Index $P_{01}^{Ke} = \frac{\sum p_1 q}{\sum p_0 q} \times 100$

- Real Wage = $\frac{Money\ Wage}{C.P.I.} \times 100$

- Output at Constant Prices = $\frac{\text{Output at Current Prices}}{\text{Price Index}} \times 100$

- Purchasing Power of Money = $\frac{1}{\text{Price Index}} \times 100$

11.10 Keywords

Base Year: The year from which comparisons are made is called the base year. It is commonly denoted by writing '0' as a subscript of the variable.

Consumer Price: It is the price at which the ultimate consumer purchases his goods and services from the retailer.

Current Year: The year under consideration for which the comparisons are to be computed is called the current year. It is commonly denoted by writing '1' as a subscript of the variable.

Index Number: An index number is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations.

Notes

Quantity Index Number: Index number that measures the change in quantities in current year as compared with a base year.

11.11 Review Questions

1. Construct Laspeyres's, Paasche's and Fisher's indices from the following data :

	1986		1987	
Item	Price (Rs)	Expenditure (Rs)	Price (Rs)	Expenditure (Rs)
1	10	60	15	75
2	12	120	15	150
3	18	90	27	81
4	8	40	12	48

2. From the following data, prove that Fisher's Ideal Index satisfies both the time reversal and the factor reversal tests.

	Base Year		Current Year	
Commodity	Price	Quantity	Price	Quantity
A	6	50	10	60
B	2	100	2	120
C	4	60	6	60

3. Examine various steps and problems involved in the construction of an index number.

4. Distinguish between average type and aggregative type of index numbers. Discuss the nature of weights used in each case.

5. Given the following data:

Year	Average weekly take-home wages	Consumer price index
	(₹)	(₹)
1968	109.50	112.8
1969	112.20	118.2
1970	116.40	127.4
1971	125.08	138.2
1972	135.40	143.5
1973	138.10	149.8

   (a) What was the real weekly wage for each year?

   (b) In which year did the employees had the greatest buying power?

   (c) What percentage increase in the average weekly wages for the year 1973 is required to provide the same buying power that the employees enjoyed in the year in which they had the highest real wages?

6. Construct Consumer Price Index for the year 1981 with 1971 as the base year.

Items	:	Food	Rent	Clothes	Fuel	Others
Percentage Expenses	:	35%	15%	20%	10%	20%
Value Index (1971)	:	150	50	100	20	60
Value Index (1981)	:	174	60	125	25	90

Notes

7. Compute consumer price index number from the following data by aggregate expenditure.

Commodity	Quantities consumed in base year	Units in which prices are quoted	Prices in base year	Prices in current year
Wheat	400 kgs	/quintal	350	400
Rice	2 quintals	/quintal	580	700
Gram	100 kgs	/quintal	740	950
Pulses	2 quintals	/quintal	980	1200
Ghee	50 kgs	/ kg ·	70	85
Sugar	50 kgs	/ kg ·	8	11
Fire wood	5 quintals	/quintal	50	60
House Rent	1 house	/house	1600	1800

8. A textile worker in the city of Ahmedabad earns ₹ 750 per month. The cost of living index for January 1986 is given as 160. Using the following data find out the amounts he spends on (i) Food and (ii) Rent.

9. "In the construction of index numbers the advantages of geometric mean are greater than those of arithmetic mean". Discuss.

10. Show that the Laspeyres's index has an upward bias and the Paasche's index has a downward bias. Under what conditions the two index numbers will be equal?

Answers: Self Assessment

1. homogeneous
2. weighted
3. barometers
4. level of an activity
5. base year
6. Current year
7. weights
8. price
9. Quantity index number
10. price index numbers
11. cost of living
12. real wages
13. base
14. weighing
15. quality
16. different

11.12 Further Readings

Books

Allan & Blumon, *Elementary Statistics : A Step by Step Approach*. McGraw-Hill College, June 2003.

David & Moae, *Introduction to the Practice of Statistics*, W.H. Freeman & Co., February 2005.

James T. McClave Terry Sincich, William Mendenhall, *Statistics*, Prentice Hall, February 2005.

Mario F. Triola, *Elementary Statistics*, Addison-Wesley, January 2006.

Mark L. Berenson, David M. Revine, Tineothy C. Krehbiel, *Basic Business Statistics: Concepts & Applications*, Prentice Hall, May 2005.

Notes

Unit 12: Hypothesis Testing

CONTENTS

Objectives

After studying this unit, you will be able to:

- Identify the Steps involved in Hypothesis Testing
- Resolve the errors in Hypothesis Testing
- Describe the One Sample and Two Sample Parametric Tests
- Explain the Chi-square Test
- Recognize the conception of ANOVA

Introduction

A statistical hypothesis test is a method of making statistical decisions using experimental data. In statistics, a result is called statistically significant if it is unlikely to have occurred by chance.

Notes

The phrase "test of significance" was coined by Ronald Fisher: "Critical tests of this kind may be called tests of significance, and when such tests are available we may discover whether a second sample is or is not significantly different from the first."

Hypothesis testing is sometimes called confirmatory data analysis, in contrast to exploratory data analysis. In frequency probability, these decisions are almost always made using null-hypothesis tests; that is, ones that answer the question. Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed? One use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom.

12.1 Steps Involved in Hypothesis Testing

1. Formulate the null hypothesis, with H_0 and H_A, the alternate hypothesis. According to the given problem, H_0 represents the value of some parameter of population.
2. Select on appropriate test assuming H_0 to be true.
3. Calculate the value.
4. Select the level of significance other at 1% or 5%.
5. Find the critical region.
6. If the calculated value lies within the critical region, then reject H_0.
7. State the conclusion in writing.

12.1.1 Formulate the Hypothesis

The normal approach is to set two hypotheses instead of one, in such a way, that if one hypothesis is true, the other is false. Alternatively, if one hypothesis is false or rejected, then the other is true or accepted. These two hypotheses are:

1. Null hypothesis
2. Alternate hypothesis

Let us assume that the mean of the population is m_0 and the mean of the sample is x. Since we have assumed that the population has a mean of m_0, this is our null hypothesis. We write this as $H_0 m = m_0$, where H_0 is the null hypothesis. Alternate hypothesis is $H_A = m$. The rejection of null hypothesis will show that the mean of the population is not m_0. This implies that alternate hypothesis is accepted.

12.1.2 Significance Level

Having formulated the hypothesis, the next step is its validity at a certain level of significance. The confidence with which a null hypothesis is accepted or rejected depends upon the significance level. A significance level of say 5% means that the risk of making a wrong decision is 5%. The researcher is likely to be wrong in accepting false hypothesis or rejecting a true hypothesis by 5 out of 100 occasions. A significance level of say 1% means, that the researcher is running the risk of being wrong in accepting or rejecting the hypothesis is one of every 100 occasions. Therefore, a 1% significance level provides greater confidence to the decision than 5% significance level.

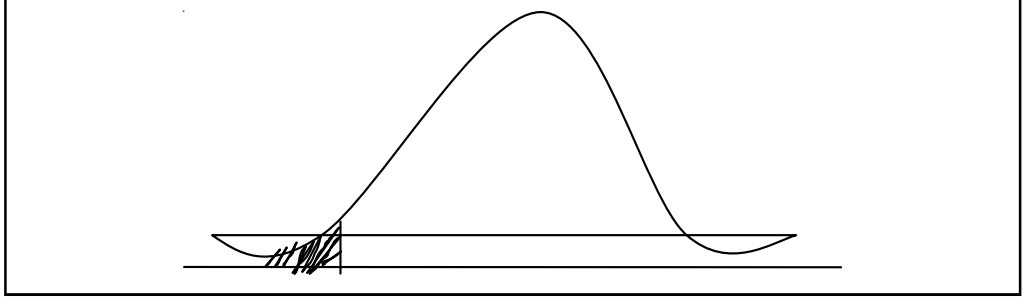
Notes

There are two types of tests.

One-tailed and Two-tailed Tests

A hypothesis test may be one-tailed or two-tailed. In one-tailed test the test-statistic for rejection of null hypothesis falls only in one-tailed of sampling distribution curve.

Figure 12.1



Example:

1. In a right side test, the critical region lies entirely in the right tail of the sample distribution. Whether the test is one-sided or two-sided – depends on alternate hypothesis.
2. A tyre company claims that mean life of its new tyre is 15,000 km. Now the researcher formulates the hypothesis that tyre life is = 15,000 km.

A two-tailed test is one in which the test statistics leading to rejection of null hypothesis falls on both tails of the sampling distribution curve as shown. One-tailed test is used when the researcher's interest is primarily on one side of the issue.

Example: "Is the current advertisement less effective than the proposed new advertisement"?

A two-tailed test is appropriate, when the researcher has no reason to focus on one side of the issue.

Example:

1. "Are the two markets - Mumbai and Delhi different to test market a product?"
2. A product is manufactured by a semi-automatic machine. Now, assume that the same product is manufactured by the fully automatic machine. This will be two-sided test, because the null hypothesis is that "the two methods used for manufacturing the product do not differ significantly".

$\therefore$ $H_0 = m_1 = m_2$

Sign of alternate hypothesis	Type of test
= /	Two-sided
<	One-sided to right
>	One-sided to left

Notes

Degree of Freedom

It tells the researcher the number of elements that can be chosen freely.

Example: a + b/2 = 5. fix a = 3, b has to be 7.

Therefore, the degree of freedom is 1.

Select Test Criteria

If the hypothesis pertains to a larger sample (30 or more), the Z-test is used. When the sample is small (less than 30), the T-test is used.

Compute

Carry out computation.

Make Decisions

Accepting or rejecting of the null hypothesis depends on whether the computed value falls in the region of rejection at a given level of significance.

Task Discuss when would you prefer two tailed test to one tailed test.

Self Assessment

Fill in the blanks:

1. Hypothesis testing is sometimes called ..... analysis.
2. The confidence with which a null hypothesis is accepted or rejected depends upon the .....
3. The rejection of null hypothesis means that the ..... hypothesis is accepted.

12.2 Errors in Hypothesis Testing

There are two types of errors:

1. Hypothesis is rejected when it is true.
2. Hypothesis is not rejected when it is false.

(1) is called Type 1 error (a), (2) is called Type 2 error (b). When a = 0.10 it means that true hypothesis will be accepted in 90 out of 100 occasions. Thus, there is a risk of rejecting a true hypothesis in 10 out of every 100 occasions. To reduce the risk, use a = 0.01 which implies that we are prepared to take a 1% risk i.e., the probability of rejecting a true hypothesis is 1%. It is also possible that in hypothesis testing, we may commit Type 2 error (b) i.e., accepting a null hypothesis which is false.

Notes The only way to reduce Type 1 and Type 2 error is by increasing the sample size.

Notes

Example of Type 1 and Type 2 error

Type 1 and Type 2 error is presented as follows. Suppose a marketing company has 2 distributors (retailers) with varying capabilities. On the basis of capabilities, the company has grouped them into two categories (1) Competent retailer (2) Incompetent retailer. Thus R1 is a competent retailer and R2 is an incompetent retailer. The firm wishes to award a performance bonus (as a part of trade promotion) to encourage good retailership. Assume that two actions A1 and A2 would represent whether the bonus or trade incentive is given and not given. This is shown as follows:

Action	(R1) Competent retailer	(R2) Incompetent retailer
A 1 performance bonus is awarded	Correct decision	Incorrect decision error (β)
A 2 performance bonus is not awarded	Incorrect decision error (α)	Correct decision

When the firm has failed to reward a competent retailer, it has committed type-2 error. On the other hand, when it was rewarded to an incompetent retailer, it has committed type-1 Error.

Self Assessment

Fill in the blanks:

4. Hypothesis is rejected when it is true is called ……error.
5. Hypothesis is not rejected when it is false is called ……..error

12.3 Parametric Tests

Parametric tests have following advantages:

1. Parametric tests are more powerful. The data in this test is derived from interval and ratio measurement.
2. In parametric tests, it is assumed that the data follows normal distributions. Examples of parametric tests are
   (a) Z-Test,
   (b) T-Test and
   (c) F-Test.
3. Observations must be independent i.e., selection of any one item should not affect the chances of selecting any others be included in the sample.

Did u know? **What is univariate/bivariate data analysis?**

Univariate

If we wish to analyse one variable at a time, this is called univariate analysis. Example: Effect of sales on pricing. Here, price is an independent variable and sales is a dependent variable. Change the price and measure the sales.

Bivariate

The relationship of two variables at a time is examined by means of bivariate data analysis.

If one is interested in a problem of detecting whether a parameter has either increased or decreased, a two-sided test is appropriate.

Parametric tests are of following types:

Notes

12.3.1 One Sample Test

One sample tests can be categorized into 2 categories.

z Test

1. When sample size is > 30

   P_1 = Proportion in sample 1

   P_2 = Proportion in sample 2

Example: You are working as a purchase manager for a company. The following information has been supplied by two scooter tyre manufacturers.

	Company A	Company B
Mean life (in km)	13000	12000
S.D (in km)	340	388
Sample size	100	100

In the above, the sample size is 100, hence a Z-test may be used.

2. Testing the hypothesis about difference between two means: This can be used when two population means are given and null hypothesis is $H_o : P_1 = P_2$.

Example: In a city during the year 2000, 20% of households indicated that they read Femina magazine. Three years later, the publisher had reasons to believe that circulation has gone up. A survey was conducted to confirm this. A sample of 1,000 respondents were contacted and it was found 210 respondents confirmed that they subscribe to the periodical 'Femina'. From the above, can we conclude that there is a significant increase in the circulation of 'Femina'?

Solution:

We will set up null hypothesis and alternate hypothesis as follows:

Null Hypothesis is $H_0 \times \mu = 15\%$

Alternate Hypothesis is $H_A \times \mu > 15\%$

This is a one-tailed (right) test.

$$Z = \frac{\frac{210}{1000} - 0.20}{\sqrt{\frac{0.20(1-0.20)}{1000}}}$$

$$Z = \frac{0.21 - 0.20}{\sqrt{\frac{0.2 \times 0.8}{1000}}}$$

$$= \frac{0.01 - \mu}{\sqrt{\frac{0.16}{1000}}}$$

Notes

$$= \frac{0.1}{\frac{0.4}{31.62}}$$

$$= \frac{0.1}{0.012} = 8.33$$

As the value of Z at 0.05 =1.64 and calculated value of Z falls in the rejection region, we reject null hypothesis, and therefore we conclude that the sale of 'Femina' has increased significantly.

T-test (Parametric Test)

T-test is used in the following circumstances: When the sample size n < 30.

Example:

1. A certain pesticide is packed into bags by a machine. A random sample of 10 bags are drawn and their contents are found as follows: 50, 49, 52, 44, 45, 48, 46, 45, 49, 45. Confirm whether the average packaging can be taken to be 50 kgs.

   In this text, the sample size is less than 30. Standard deviations are not known using this test. We can find out if there is any significant difference between the two means i.e. whether the two population means are equal.

2. There are two nourishment programmes 'A' and 'B'. Two groups of children are subjected to this. Their weight is measured after six months. The first group of children subjected to the programme 'A' weighed 44, 37, 48, 60, 41 kgs. at the end of programme. The second group of children were subjected to nourishment programme 'B' and their weight was 42, 42, 58, 64, 64, 67, 62 kgs. at the end of the programme. From the above, can we conclude that nourishment programme 'B' increased the weight of the children significantly, given a 5% level of confidence.

Null Hypothesis: There is no significant difference between Nourishment programme 'A' and 'B'.

Alternative Hypothesis: Nourishment programme B is better than 'A' or Nourishment programme 'B' increase the children's weight significantly.

Solution:

	Nourishment programme A			**Nourishment programme B**	
X	$x - \bar{x}$ **= (x - 46)**	$(x - \bar{x})^2$	y	$y - \bar{y}$ **= (y - 57)**	$(y - \bar{y})^2$
44	-2	4	42	-15	225
37	-9	81	42	-15	225
48	2	4	58	1	1
60	14	196	64	7	49
41	-5	25	64	7	49
			67	10	100
			62	5	25
230	**0**	**310**	**399**	**0**	**674**

Notes

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Here $n_1 = 5 \quad n_2 = 7$

$\Sigma x = 230, \quad \Sigma y = 399$

$\Sigma\left(x - \bar{x}\right)^2 = 310$, $\Sigma\left(y - \bar{y}\right)^2 = 399$

$$\bar{x} = \frac{\Sigma x}{n_1} = \frac{230}{5} = 46$$

$$\bar{y} = \frac{\Sigma y}{n_2} = \frac{399}{7} = 57$$

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left\{\sum\left(x - \bar{x}\right)^2 + \sum\left(y - \bar{y}\right)^2\right\}$$

D.F. = $(n_1 + n_2 - 2) = (5 + 7 - 2) = 10$

$$s^2 = \frac{1}{10}\{310 + 674\} = 98.4$$

$$t = \frac{46 - 57}{\sqrt{98.4 \times \left(\frac{1}{5} + \frac{1}{7}\right)}}$$

$$= \frac{-11}{\sqrt{98.4 \times \left(\frac{12}{35}\right)}}$$

$$= \frac{-11}{\sqrt{33.73}} = -\frac{11}{5.8}$$

$= -1.89$

t at 10 d.f. at 5% level is 1.81.

Since, calculated t is greater than 1.81, it is significant. Hence H_A is accepted. Therefore the two nutrition programmes differ significantly with respect to weight increase.

Two Tailed t-Test

When two samples are related we use paired t-test for judging the significance of the mean of difference of the two related samples. It can also be used for judging the significance of the coefficients of simple and partial correlations.

The t-test is performed using the following formula;

$$t = r_{yx}\sqrt{\frac{n-2}{1-r_{yx}}}$$

Notes

Where, (n – 2) is degrees of freedom, r_{yx} is coefficient of correlation between x and y. The computed value of t is compared with its table value. If the computed value is less than the table value the null hypothesis is accepted or rejected otherwise at a given level of significance.

Example: A study of weight of 18 pairs of male and female employees in a company shows that coefficient of correlation is 0.52. Test the significance of correlation.

Solution:

Applying t test:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$r = 0.52, n = 18$$

$$t = 0.52\sqrt{\frac{18-2}{1-(0.52)^2}}$$

$$= \frac{0.52 \times 4}{0.854} = 2.44$$

$$\nu = (n - 2) = (18 - 2) = 16$$

$$\nu = 16, t_{0.05} = 2.12$$

The calculated value of t is greater than the table value. The given value of r is significant.

12.3.2 Two Sample Test

Two sample test if known as F test

F-Test

Let there be two independent random samples of sizes n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2 respectively. Further, let $s_1^2 = \frac{1}{n_1 - 1}\sum(X_{1i} - \bar{X}_1)^2$ and $s_2^2 = \frac{1}{n_2 - 1}\sum(X_{2i} - \bar{X}_2)^2$ be the variances of the first sample and the second samples respectively. Then F - statistic is defined as the ratio of two χ^2 - variates. Thus, we can write

$$F = \frac{\frac{\chi^2_{n_1-1}}{n-1}}{\frac{\chi^2_{n_2-1}}{n_2-1}} = \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)s_2^2}{\sigma_2^2}/(n_2-1)} = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

Features of F- distribution

1. This distribution has two parameters v_1 (= n_1 – 1) and v_2 (= n_2 – 1).

2. The mean of F - variate with v_1 and v_2 degrees of freedom is $\frac{v_2}{v_2 - 2}$ and standard error is

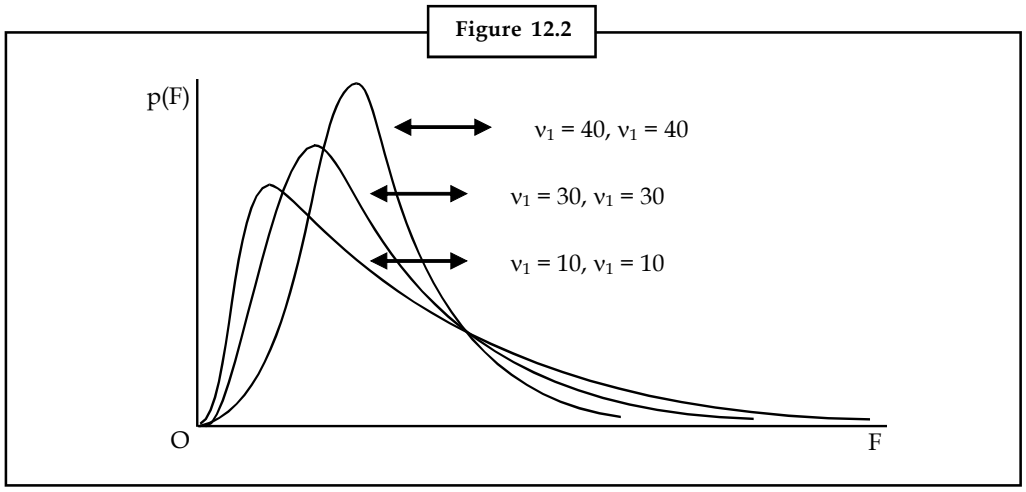
$$\left(\frac{v_2}{v_2 - 2}\right)\sqrt{\frac{2(v_1 + v_2 - 2)}{v_1(v_2 - 4)}}$$

Notes

We note that the mean will exist if $v_2 > 2$ and standard error will exist if $v_2 > 4$. Further, the mean > 1.

3. The random variate F can take only positive values from 0 to ∞. The curve is positively skewed.
4. For large values of ν_1 and ν_2, the distribution approaches normal distribution.
5. If a random variate follows t-distribution with ν degrees of freedom, then its square follows F-distribution with 1 and ν d.f. i.e. $t^2_\nu = F_{1,\nu}$
6. F and χ^2 are also related as $F_{\nu_1, \nu_2} = \frac{(\chi^2_{\nu_1})}{\nu_1}$ as $\nu_2 \to \infty$

Figure 12.2



Self Assessment

Fill in the blanks:

6. The relationship of two variables at a time is examined by means of ..... data analysis.
7. The data in parametric test is derived from interval and …….measurement.
8. One sample tests can be categorized into ……categories

12.4 Chi-square Test

A chi-square test (also chi-squared or χ^2 test) is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true, or any in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough.

Caution One case where the distribution of the test statistic is an exact chi-square distribution is the test that the variance of a normally-distributed population has a given value based on a sample variance. Such a test is uncommon in practice because values of variances to test against are seldom known exactly.

Notes

It is used in the following circumstances:

1. Sample observations should be independent i.e. two individual items should be included twice in a sample.
2. The sample should contain at least 50 observations

   or

   total frequency should be greater than 50.
3. There should be a minimum of five observations in any cell. This is called cell frequency constraint.

For instance: Chi-square

Persons	Age Group				Total
	Under 20-40	**20-40**	**41-50**	**51 & Over**	
Liked the car	146	78	48	28	300
Disliked the car	54	52	32	62	200
Total	200	130	80	90	500

Is there any significant difference between the age group and preference for the car?

Example: A company marketing tea claims that 70% of population in a metro drinks a particular brand (Wood Smoke) of tea. A competing brand challenged this claim. They took a random sample of 200 families to gather data. During the study period, it was found that 130 families were using this brand of tea. Will it be correct on the part of competitor to conclude that the claim made by the company does not holds good at 5% level of significance?

Solution:

Hypothesis H_0 – People who drink Wood Smoke brand is 70%.

H_0 – People who drink Wood Smoke brand is not 70%.

If the hypothesis is true then number of consumers who drink this particular brand is 200 × 0.7 = 140.

Those who do not drink that brand are 200 × 0.3 = 60

Degree of freedom = D = 2 – 1 = 1, since there are two groups.

Group	Observed (O)	Expected (E)	O-E	$(O-E)^2$	$(O-E)^2/E$
Those who drink branded tea	130	140	-10	100	0.714
Those who did not drink branded tea	70	60	+10	100	1.667
	200	**200**	**0**		

$$\lambda^2 = \frac{(0-E)}{E} = 2.381$$

A 0.5 level of significance of for 1 d.f. is equal to 3.841 (From tables). The calculated value is 2.381 is lower. Therefore, we accept the hypothesis that 70% of the people in that metro drink Wood Smoke branded tea.

Notes

Self Assessment

Fill in the blanks:

9. A chi-square test is used when sample observations should be ……

10. For applying chi-square test , sample should contain at least …… observations

12.5 ANOVA

ANOVA is a statistical technique. It is used to test the equality of three or more sample means. Based on the means, inference is drawn whether samples belongs to same population or not.

Notes ***Conditions for using ANOVA***

1. Data should be quantitative in nature.
2. Data normally distributed.
3. Samples drawn from a population follow random variation.

ANOVA can be discussed in two parts:

1. One-way classification
2. Two and three-way classification.

12.5.1 One-way ANOVA

Following are the steps followed in ANOVA:

1. Calculate the variance between samples.
2. Calculate the variance within samples.
3. Calculate F ratio using the formula.

   F = Variance between the samples/Variance within the sample
4. Compare the value of F obtained above in (3) with the critical value of F such as 5% level of significance for the applicable degree of freedom.
5. When the calculated value of F is less than the table value of F, the difference in sample means is not significant and a null hypothesis is accepted. On the other hand, when the calculated value of F is more than the critical value of F, the difference in sample means is considered as significant and the null hypothesis is rejected.

Example: ANOVA is useful.

1. To compare the mileage achieved by different brands of automotive fuel.
2. Compare the first year earnings of graduates of half a dozen top business schools.

Application in Market Research

Consider the following pricing experiment. Three prices are considered for a new toffee box introduced by Nutrine company. Price of three varieties of toffee boxes are ₹ 39, ₹ 44 and ₹ 49. The idea is to determine the influence of price levels on sales. Five supermarkets are selected to exhibit these toffee boxes. The sales are as follows:

Notes

Price (₹)	1	2	3	4	5	Total	Sample mean $\bar{x}$	
39	8	12	10	9	11	50		10
44	7	10	6	8	9	40		8
49	4	8	7	9	7	35		7

What the manufacturer wants to know is: (1) Whether the difference among the means is significant? If the difference is not significant, then the sale must be due to chance. (2) Do the means differ? (3) Can we conclude that the three samples are drawn from the same population or not?

Example: In a company there are four shop floors. Productivity rate for three methods of incentives and gain sharing in each shop floor is presented in the following table. Analyze whether various methods of incentives and gain sharing differ significantly at 5% and 1% F-limits.

Shop Floor	Productivity rate data for three methods of incentives and gain sharing		
	X_1	X_2	X_3
1	5	4	4
2	6	4	3
3	2	2	2
4	7	6	3

Solution:

Step 1: Calculate mean of each of the three samples (i.e., x_1, x_2 and x_3, i.e. different methods of incentive gain sharing).

$$\bar{X}_1 = \frac{5+6+2+7}{4} = 5$$

$$\bar{X}_2 = \frac{4+3+2+3}{4} = 3$$

$$\bar{X}_3 = \frac{4+3+2+3}{4} = 3$$

Step 2: Calculate mean of sample means i.e., $\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{K}$

where, K denotes Number of samples $= \frac{5+3+3}{3} = 4$(approximated)

Step 3: Calculate sum of squares (s.s.) for variance between and within the samples.

ss between = $n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + n_3(\bar{x}_3 - \bar{\bar{x}})^2$

ss within = $\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2i} - \bar{x}_2)^2 + \Sigma(x_{3i} - \bar{x}_3)^2$

Sum of squares (ss) for variance between samples is obtained by taking the deviations of the sample means from the mean of sample means () and by calculating the squares of such deviation,

Notes

which are multiplied by the respective number of items or categories in the samples and then by obtaining their total. Sum of squares(ss) for variance within samples is obtained by taking deviations of the values of all sample items from corresponding sample means and by squaring such deviations and then totalling them. For our illustration then

$$\text{ss between} = 4(5-4)^2 + 4(4-4)^2 + 4(3-4)^2$$

$$= 4 + 0 + 4 = 8$$

$$\text{ss within} = \frac{\{(5-5)^2+(6-5)^2+(2-5)^2+(7-5)^2\}}{\Sigma(x_{1i}-\bar{x}_1)^2} + \frac{\{(4-4)^2+(4-4)^2+(2-4)^2+(6-4)^2\}}{\Sigma(x_{2i}-\bar{x}_2)^2}$$

$$+ \frac{\{(4-3)^2+(3-3)^2+(2-3)^2+(3-3)^2\}}{\Sigma(x_{3i}-\bar{x}_3)^2}$$

$$= (0+1+9+4)+(0+0+4+4)+(1+0+1+0)$$

$$= 14 + 8 + 2$$

$$= 24$$

Step 4: ss of total variance which is equal to total of s.s. between and ss within and is denoted by formula as follows:

$$\Sigma(x_{ij}-\bar{\bar{x}})^2$$

where

i = 1.23

j = 1.23

for our example, total ss will thus be:

$$[\{(5-4)^2+(6-4)^2+(2-4)^2+(7-4)^2\}+\{(4-4)^2+(4-4)^2+(2-4)^2+(6-4)^2\}$$

$$+ \{(4-4)^2+(3-4)^2+(2-4)^2+(3-4)^2\}]$$

$$= \{(1+4+4+9)+(0+0+4+4)+(0+1+4+1)\}$$

$$= 08 + 8 + 6 = 32$$

We will, however, get the same value if we simply total respective values of ss between and ss within. For our example, ss between is 8 and ss within is 24, thus ss of total variance is 32 (8+24).

Step 5: Ascertain degrees of freedom and mean square (MS) between and within the samples. Degrees of freedom (df) for between samples and within samples are computed differently as follows.

For between samples, df is (k-1), where k' represents number of samples (for us it is 3). For within samples df is (n-k), where 'n' represents total number of items in all the samples (for us it is 12).

Mean squares (MS) between and within samples are computed by dividing the ss between and ss within by respective degrees of freedom. Thus for our example:

(i) $\text{MS between} = \frac{\text{ss between}}{(k-1)} = \frac{8}{2} = 4$

where (K – 1) is the df.

Notes

(ii) MS within = $\frac{\text{ss within}}{(n-k)} = \frac{24}{9} = 2.67$

where (n – k) is the df.

Step 6: Now we will have to compute F ratio by analysing our samples. The formula for computing 'F' ratio is: $\frac{\text{ss between}}{\text{ss within}}$

Thus for our example, F ratio = $\frac{4.00}{2.67} = 1.5$

Step 7: Now we will have to analyze whether various methods of incentives and gain sharing differ significantly at 5% and 1% 'F' limits. For this, we need to compare observed 'F' ratio with 'F' table values. When observed 'F' value at given degrees of freedom is either equal to or less than the table value, difference is considered insignificant. In reverse cases, i.e., when calculated 'F' value is higher than table-F value, the difference is considered significant and accordingly we draw our conclusion.

For example, our observed 'F' ratio at degrees of freedom (v_1* & v_2**, i.e., and 9) is 1.5. The table value of F at 5% level with df 2 and 9 ($v_1 = 2$, $v_2 = 9$) is 4.26. Since the table value is higher than the observed value, difference in rate of productivity due to various methods of incentives and gain sharing is considered insignificant. At 1% level with df 2 and 9, we get the table value of F as 8.02 and we draw the same conclusion.

We can now draw an ANOVA table as follows to show our entire observation.

Variation	SS	df	MS	F-ratio	Table value of F	
					5%	1%
Between sample	8	(k–1)= (3–1)=2	ss between (k–1) = 8/2 = 4	MS between MS within = 4/2.67 =1.5	F (v_1, v_2) =F (2,9) = 4.26	F (v_1, v_2) =F(2,9) 8.02
Within simple	24	(n–k)= (12–3) = 9	ss.within (n–k) = 24/9 = 2.67			

12.5.2 Two-way ANOVA

The procedure to be followed to calculate variance is the same as it is for the one-way classification. The example of two-way classification of ANOVA is as follows:

Suppose, a firm has four types of machines – A, B, C and D. It has put four of its workers on each machines for a specified period, say one week. At the end of one week, the average output of each worker on each type of machine was calculated. These data are given below:

Average Production by the Type of Machine

	A	B	C	D
Worker 1	25	26	23	28
Worker 2	23	22	24	27
Worker 3	27	30	26	32
Worker 4	29	34	27	33

Notes

The firm is interested in knowing:

1. Whether the mean productivity of workers is significantly different.
2. Whether there is a significant difference in the mean productivity of different types of machines.

Example: Company 'X' wants its employees to undergo three different types of training programme with a view to obtain improved productivity from them. After the completion of the training programme, 16 new employees are assigned at random to three training methods and the production performance were recorded.

The training managers problem is to find out if there are any differences in the effectiveness of the training methods? The data recorded is as under:

Daily Output of New Employees

Method 1	15	18	19	22	11	
Method 2	22	27	18	21	17	
Method 3	18	24	19	16	22	15

Following steps are followed.

1. Calculate Sample mean i.e. $\bar{x}$
2. Calculate General mean i.e. $\bar{\bar{x}}$
3. Calculate variance between columns using the formula $\bar{\sigma}^2 = \frac{\sum n_i \left(x_i - \bar{\bar{x}}\right)^2}{k-1}$

   where K = $(n_1 + n_2 + n_3 - 3)$.
4. Calculate sample variance. It is calculated using formula:

   Sample variance $s_i^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}$ where n is No. of observation under each method.
5. Calculate variance within columns using the formula $\bar{\sigma}^2 = \frac{\sum n_i - 1}{n_r - k}$
6. Calculate F using the ratio F = $\left(\frac{\text{between column variance}}{\text{within column variance}}\right)$
7. Calculate the number of degree of freedom in the numerator F ratio using equation, d.f = (No. of samples –1).
8. Calculate the number of degree of freedom in the denominator of F ratio using the equation d.f = $S(n_i - k)$
9. Refer to F table f8 find value.
10. Draw conclusions.

Notes

Solution:

Method 1	Method 2	Method 3
15	22	24
18	27	19
19	18	16
22	21	22
11	17	15
		18
85	**105**	**114**

1. Sample mean is calculated as follows:

$$\bar{x}_1 = \frac{85}{5} = 17, \bar{x}_2 = \frac{105}{5} = 21, \bar{x}_3 = \frac{114}{6} = 19$$

2. Grand mean

$$\bar{\bar{x}} = \frac{15+18+19+22+11+22+27+18+21+17+24+19+16+22+15+18}{16} = \frac{304}{16} = 19$$

3. Calculate variance between columns:

n	$\bar{x}$	$\bar{x}$	$\bar{x} - \bar{\bar{x}}$	$(\bar{x} - \bar{\bar{x}})^2$	$n(\bar{x} - \bar{\bar{x}})^2$
5	17	19	-2	4	5 × 4 = 20
5	21	19	2	4	5 × 4 = 20
6	19	19	0	0	6 × 0 = 0
				$\sum n_i(\bar{x}_1 - \bar{\bar{x}})^2$	**= 40**

$$\bar{\sigma}^2 = \frac{\sum n_i\left(x_i - \bar{\bar{x}}\right)^2}{k-1} = \frac{40}{3-1} = 20$$

Variance between column = 20

4. Calculation sample variance:

Training method -1		**Training method -2**		**Training method -3**	
$x - \bar{x}$	$(x - \bar{x})^2$	$x - \bar{x}$	$(x - \bar{x})^2$	$x - \bar{x}$	$(x - \bar{x})^2$
15-17	$(-2)^2 = 4$	22-21	$(1)^2 = 1$	18-19	$(1)^2 = 1$
18-17	$(1)^2 = 1$	27-21	$(6)^2 = 36$	24-19	$(5)^2 = 25$
19-17	$(2)^2 = 4$	18-21	$(-3)^2 = 9$	19-19	$(0)^2 = 0$
22-17	$(5)^2 = 25$	21-21	$(0)^2 = 1$	16-19	$(-3)^2 = 9$
11-17	$(-6)^2 = 36$	17-21	$(-4)^2 = 16$	22-19	$(3)^2 = 9$
				15-19	$(-4)^2 = 16$
	$\sum(x - \bar{x})^2 = 70$		$\sum(x - \bar{x})^2 = 62$		$\sum(x - \bar{x})^2 = 60$

Sample variance = $\frac{\sum\left(x-\bar{x}\right)^2}{n-1} = \frac{70}{5-1}$, $\frac{\sum\left(x-\bar{x}\right)^2}{n-1} = \frac{62}{5-1}$, $\frac{\sum\left(x-\bar{x}\right)^2}{n-1} = \frac{60}{5-1}$

Notes

$s_1^2 = \frac{70}{4} = 17.5,\ s_2^2 = \frac{62}{4} = 15.5,\ s_3^2 = \frac{60}{5} = 12$

5. Within column variance $\overline{\sigma}^2 = \sum\left(\frac{n_i - 1}{n_i - k}\right)s_1^2$

$= \left(\frac{5-1}{16-3}\right)\times 17.5 + \left(\frac{5-1}{16-3}\right)\times 15.5 + \left(\frac{6-1}{16-3}\right)\times 12$

$= \left(\frac{4}{13}\right)\times 17.5 + \left(\frac{4}{13}\right)\times 15.5 + \frac{5}{13}\times 12$

Within column variance $= \frac{192}{13} = 14.76$

6. $F = \frac{\text{Between column variance}}{\text{Within column variance}} = \frac{20}{14.76} = 1.354$

7. d.f. of Numerator = (3 - 1) = 2.

8. d.f. of Denominator = $Sn_1 - k$ = (5 - 1) + (5 - 1) + (6 - 1) = 16 - 3 = 13.

9. Refer to table using d.f. = 2 and d.f. = 13.

10. The value is 3.81. This is the upper limit of acceptance region. Since calculated value 1.354 lies within it we can accept H_0, the null hypothesis.

Conclusion: There is no significant difference in the effect of the three training methods.

Example: Let us now frame a problem to study the effects of incentive and gain sharing and level of technology (independent variables) on productivity rate (dependent variable).

Productivity Rate Data of Workers of M/s. XYZ & Co.

Level of Technology	Incentive and gain sharing		
	A	B	C
W	4	3	3
X	5	3	2
Y	1	1	1
Z	6	5	2

Solution:

1. Total values (T) of individual item = 36, n = 12

2. Correction factor = $\frac{(T)^2}{n} = \frac{36\times 36}{12}$

= 108

3. Total ss = (16 + 9 + 9 + 25 + 9 + 4 + 1 + 1 + 1 + 36 + 25 + 4)

= 140 – 108 = 32

Notes

4. ss between columns:

$$= \left[\frac{16\times16}{4}+\frac{12\times12}{4}+\frac{8\times8}{4}\right]-108$$

5. ss between rows:

$$= \left[\frac{10\times10}{3}+\frac{10\times10}{3}+\frac{3\times3}{3}+\frac{13\times13}{3}\right]-108$$

$$= \left[\frac{100}{3}+\frac{100}{3}+\frac{9}{3}+\frac{169}{3}\right]-108$$

= [33.33 + 33.33 + 3 + 56.33] - 108

= 126 - 108

= 18 (after adjusting fraction)

6. ss residual:

= Total ss - (ss between column + ss between rows)

= 32 – (8 +18) = 6.

Now we need to set up ANOVA table.

Variation source	SS	d.f	M.S	F ratio	5%	1%
Between columns	8	(c–1) = 2	8/2=4	4/1=4	F (2, 6) = 5.14	F (2, 6) = 10.92
Between rows	18	(r–1) = 3	18/3=6	6/1=6	F (3, 6) = 4.76	F (3, 6) = 9.78
Residual	6	(c–1) x (r–1) = 6	6/6=1			

From the ANOVA table, we find that differences related to varieties of incentives and gain sharing are insignificant at 5% level as the calculated F-ratio, i.e., 4 is less than table value of F, which is 5.14. However differences are significant for different levels of technology at 5% level as the observed F ratio is higher than table value of F. At 1% level, however, differences are insignificant.

Self Assessment

Fill in the blanks:

11. ..... is used to test the equality of three or more sample means.

12. For using ANOVA Data should be ……..in nature

12.6 Non-parametric Test

Notes

Non-parametric tests are used to test the hypothesis with nominal and ordinal data.

1. We do not make assumptions about the shape of population distribution.
2. The hypothesis of non-parametric test is concerned with something other than the value of a population parameter.
3. Easy to compute. There are certain situations particularly in marketing research, where the assumptions of parametric tests are not valid. Example: In a parametric test, we assume that data collected follows a normal distribution. In such cases, non-parametric tests are used. Example of non-parametric tests are Binomial test, Mann-Whitney U test, Sign test, etc. A binomial test is used when the population has only two classes such as male, female; buyers, non-buyers, success, failure etc. All observations made about the population must fall into one of the two tests. The binomial test is used when the sample size is small.

Did u know? Non-parametric tests are distribution-free tests.

Advantages

1. They are quick and easy to use.
2. When data are not very accurate, these tests produce fairly good results.

Disadvantage

Non-parametric test involves the greater risk of accepting a false hypothesis and thus committing a Type 2 error.

12.6.1 One Sample Tests

The following are the main examples of one sample non-parametric tests:

Cox and Stuart Test

This test is used to examine the presence of trends. A set of numbers is said to show upward trend if the latter numbers in the sequence are greater than the former numbers. And similarly, one can define a downward trend. How to examine whether a trend is noticeable in a sequence?

Example: Suppose a marketer wants to examine whether its sales are showing a trend or just fluctuating randomly. Suppose the company has gathered the monthly sales figures during the past one year month-wise:

Month	1	2	3	4	5	6	7	8	9	10	11	12
Sales	200	250	280	300	320	278	349	268	240	318	220	380

From the given data, analyse the sales trend.

Notes

Sign-test

Sign-test is used with matched pairs. The test is used to identify the pairs and decide whether the pair has more or less similar characteristics.

Example: Suppose, an experiment on the effect of brand name on quality perceptions is to be conducted. 10 persons are selected and asked to taste and compare the two products (beverage). One of them is identified as branded well known beverage, and the other is a new beverage. In reality, the samples are identical. The respondents who tested were asked to rate the two samples on an ordinal scale. Two hypotheses are set up as follows:

H_0 - there is no difference between the perceived qualities of two beverages.

H_A - there is a difference in the perceived qualities of two beverages.

12.6.2 Two Sample Tests

The following are the main examples of two sample non-parametric tests:

Mann Whitney "U" Test

(Rank Sum test)

This test is used to determine whether two independent samples have been drawn from the same population. Suppose an experiment has obtained two sets of samples from two populations and the study wishes to examine whether the two populations are identical.

Example: A computer company XYZ would like to choose the performance of programmers, working in 2 branches, located in different cities. The performance indices of employees:

Branch A	Branch B
84	76
68	77
78	64
49	62
45	53

To find out whether there is any difference in the performance indices of employees of the two branches.

Kolmogorov-Smirnov Test

This is used for examining the efficacy of fit between observed samples and expected frequency distribution of data when the variable is in the ordinal scale.

Example: A manufacturer of cosmetics wants to test four different shades of the liquid foundation compound – very light, light, medium and dark. The company has hired a market research agency to determine whether any distinct preference exists towards either extreme. If so, the company will manufacture only the preferred shade, otherwise, the company is planning to market all shades. Suppose, out of a sample of hundred, 50 preferred "very light shade" 30 liked light shade, 15 the medium shade, and 50 dark shades. Do you think the results show any kind of preference?

Since the shade represents ordering (rank), this test can be used to find the preference.

Notes

12.6.3 K Sample Test

We can use the Mann Whitney test; when two populations are involved, the Kruskal-Wallis test is used, when more than two populations are involved. This test will enable us to know whether independent samples have been drawn from the same population or from different populations having the same distribution. This test is an extension of "Mann Whitney test".

This is a type of Rank Sum test. This test is used to find out whether two or more independent samples are drawn from an identical population. This test is also called the H Test. Mann Whitney test is used when only two populations are involved and Kruskal- Wallis test is used when more than two populations are involved.

Example: In an assembling unit, three different workers do assembly work in shifts. The data is tabulated as follows:

Shift No.	Worker-1	Worker-2	Worker-3
1	25	28	29
2	31	28	30
3	35	29	27
4	33	28	36
5	35	32	31
6	31	32	34

Check whether there is any difference in the production quantum of the three workers:

Example: (Kruskal-Wallis Test, H-Test)

Let us assume that there are three categories of workers involved in a building construction. The wages depends on the skills possessed by them and their availability. The wages of three categories, namely painter carpenter and plumber are as follows:

Item	Sample 1 Daily wages (Painter ₹)	Sample 2 Daily wages (Carpenter ₹)	Sample 3 Daily wages (Plumber ₹)
1	64	72	51
2	66	74	52
3	72	75	54
4	74	78	56
5		80	

Use H-test and state whether the three populations are same or different.

Solution:

H_0 - The wages of the three occupations are the same.

H_1 - The wages of the three occupations is not the same.

Notes

Item	Wage-Painter ₹/day		Wage-Carpenter ₹/day		Wage-Plumber ₹/day	
	₹	Rank	₹	Rank	₹	Rank
1	64	5	72	7.5	51	1
2	66	6	74	9.5	52	2
3	72	7.5	75	11	54	3
4	74	9.5	78	12	56	4
5			80	13		
Total	**276**	**$R_1 = 28$**	**379**	**$R_2 = 53$**	**213**	**$R_3 = 10$**

$n_1 = 4,\ n_2 = 5,\ n_3 = 4$

$n = n_1 + n_2 + n_3 = 4 + 5 + 4 = 13$

$R_1 = 28, R_2 = 53, R_3 = 10$

$$H = \frac{12}{n(n+1)}\sum\left[\frac{R_1^{\,2}}{n_1}\right] - 3(n+1)$$

$$H = \frac{12}{13(13+1)}\sum\left[\frac{28^2}{4} + \frac{53^2}{5} + \frac{10^2}{4}\right] - 3(3+1) = 9.61$$

At 5% level of significance, for d.f. = (3 - 1) = 2, the table value is 5.991. Computed value 9.61 is greater.

Conclusion: Reject the Null hypothesis that the three populations are different.

Self Assessment

Fill in the blanks:

13. ..... Test is used to determine whether two independent samples have been drawn from the same population.
14. ..... Test is used for examining the efficacy of fit between observed samples and expected frequency distribution.
15. Sign-test is used with ..... pairs.
16. Non-parametric tests are used to test the hypothesis with ..... and ..... data.
17. ..... Test is used to examine the presence of trends.
18. ..... test involves the greater risk of accepting a false hypothesis.

12.7 Summary

- Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true.
- The usual process of hypothesis testing consists of four steps.
- Formulate the null hypothesis and the alternative hypothesis.

Notes

- Identify a test statistic that can be used to assess the truth of the null hypothesis.
- Compute the P-value, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true.
- The smaller the ρ-value, the stronger the evidence against the null hypothesis.
- Compare the ρ-value to an acceptable significance value α.
- If $p \leq \alpha$, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

12.8 Keywords

Alternate Hypothesis: An alternative hypothesis is one that specifies that the null hypothesis is not true. The alternative hypothesis is false when the null hypothesis is true, and true when the null hypothesis is false.

ANOVA: It is a statistical technique used to test the equality of three or more sample means.

Degree of Freedom: It is the consideration that tells the researcher the number of elements that can be chosen freely.

Null Hypothesis: The null hypothesis is a hypothesis which the researcher tries to disprove, reject or nullify.

Significance Level: Significance level is the criterion used for rejecting the null hypothesis.

12.9 Review Questions

1. What hypothesis, test and procedure would you use when an automobile company has manufacturing facility at two different geographical locations? Each location manufactures two-wheelers of a different model. The customer wants to know if the mileage given by both the models is the same or not. Samples of 45 numbers may be taken for this purpose.
2. What hypothesis, test and procedure would you use when a company has 22 sales executives? They underwent a training programme. The test must evaluate whether the sales performance is unchanged or improved after the training programme.
3. What hypothesis, test and procedure would you use in A company has three categories of managers:
   (a) With professional qualifications but without work experience.
   (b) With professional qualifications accompanied by work experience.
   (c) Without professional qualifications but with work experience.
4. Each person in a random sample of 50 was asked to state his/her sex and preferred colour. The resulting frequencies are shown below.

Colour		**Red**	**Blue**	**Green**
Sex	**Male**	5	14	6
	Female	15	6	4

A chi-square test is used to test the null hypothesis that sex and preferred colour are independent. Will you reject at the null hypothesis 0.005 level? Why/Why not?

Notes

5. Are all employees equally prone to having accidents? To investigate this hypothesis, Parry (1985) looked at a light manufacturing plant and classified the accidents by type and by age of the employee.

Age	Accident Type		
	Sprain	Burn	Cut
Under 25	9	17	5
25 or over	61	13	12

A chi-square test gave a test-statistic of 20.78. If we test at a =.05, does the proportion of sprain, cuts and burns seems to be similar for both age classes? Why/why not?

6. In hypothesis testing, if β is the probability of committing an error of Type II. The power of the test, 1 – β is then the probability of rejecting H_0 when H_A is true or not? Why?

7. In a statistical test of hypothesis, what would happen to the rejection region if α, the level of significance, is reduced?

8. During the pre-flight check, Pilot Mohan discovers a minor problem - a warning light indicates that the fuel gauge may be broken. If Mohan decides to check the fuel level by hand, it will delay the flight by 45 minutes. If he decides to ignore the warning, the aircraft may run out of fuel before it gets to Mumbai. In this situation, what would be:

   (a) the appropriate null hypothesis? and;

   (b) a type I error?

9. Can the probability of a Type II error be controlled by the sample size? Why/ why not?

10. A research biologist has carried out an experiment on a random sample of 15 experimental plots in a field. Following the collection of data, a test of significance was conducted under appropriate null and alternative hypotheses and the P-value was determined to be approximately .03. What does this indicate with respect to the hypothesis testing?

11. Two samples were drawn from a recent survey, each containing 500 hamlets. In the first sample, the mean population per hamlet was found to be 100 with a S.D. of 20, while in the second sample the mean population was 120 with a S.D. 15. Do you find the averages of the samples to be statistically significant?

12. A simple random sample of size 100 has a mean of 15, the population variance being 25. Find an interval estimate of the population mean with a confidence level of (i) 99% and (ii) 95%.

13. A population consists of five numbers 2, 3, 6, 8, 11. Consider all possible samples of size two which can be drawn with replacement from this population. Calculate the S.E. of sample means.

14. A certain drug is claimed to be effective in curing colds; half of them were given sugar pills. The patients' reactions to the treatment are recorded in the following table.

	Helped	Harmed	No effect
Drug	52	10	18
Sugar pills	44	10	26

Test the hypothesis that the drug is no better than the sugar pills for curing colds. (The 5 % value of x^2 for v = 2 = 5.991)

15. A random sample of 640 persons from a village provided the following information:

Notes

Effect of Influenza	New drug administered	New drug not administered	Total
Attacked	100	60	160
Not attacked	200	280	480
Total	300	340	640

Test whether the new drug was effective in preventing the attack of influenza.

Answers: Self Assessment

1. confirmatory data
2. significance level
3. alternate
4. Type 1
5. Type 2
6. bivariate
7. ratio
8. 2
9. independent
10. 50
11. ANOVA
12. quantitative
13. Mann Whitney "U"
14. Kolmogorov-Smirnov
15. matched
16. nominal, ordinal
17. Cox and Stuart
18. Non-parametric

12.10 Further Readings

Books

Abrams, M.A, *Social Surveys and Social Action*, London: Heinemann, 1951.

Arthur, Maurice, *Philosophy of Scientific Investigation*, Baltimore: John Hopkins University Press, 1943.

R.S. Bhardwaj, *Business Statistics*, Excel Books, New Delhi, 2008.

S.N. Murthy and U. Bhojanna, *Business Research Methods*, Excel Books, 2007.

Notes

Unit 13: Multivariate Analysis

CONTENTS

Objectives

After studying this unit, you will be able to:

- Explain the concept of multivariate analysis
- Classify the multivariate analysis
- Define the Discriminant Analysis and Conjoint Analysis
- Discuss the Factor Analysis and Cluster Analysis
- State the Multidimensional Scaling (MDS)

Introduction

As the name indicates, multivariate analysis comprises a set of techniques dedicated to the analysis of data sets with more than one variable. Several of these techniques were developed recently in part because they require the computational capabilities of modern computers. Multivariate analysis (MVA) is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical variable at a time. In design and analysis, the technique is used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest. Sometimes, the marketers will come across situations, which are complex involving two or more variables. Hence, bivariate analysis deals with this type of situation. Chi-Square is an example of bivariate analysis.

13.1 Multivariate Analysis

Notes

In multivariate analysis, the number of variables to be tackled are many.

Example: The demand for television sets may depend not only on price, but also on the income of households, advertising expenditure incurred by TV manufacturer and other similar factors. To solve this type of problem, multivariate analysis is required.

Classification

Multiple-variate analysis: This can be classified under the following heads:

1. Multiple regression
2. Discriminant analysis
3. Conjoint analysis
4. Factor analysis
5. Cluster analysis
6. Multidimensional scaling.

13.1.1 Multiple Regression

In the case of simple linear regression, one variable, say, X_1 is affected by a linear combination of another variable X_2 (we shall use X_1 and X_2 instead of Y and X used earlier). However, if X_1 is affected by a linear combination of more than one variable, the regression is termed as a multiple linear regression.

Let there be k variables X_1, X_2 ..... X_k, where one of these, say X_j, is affected by the remaining k – 1 variables. We write the typical regression equation as

$$X_{jc} = a_{j\times 1,\,2,\,\ldots.\,j-1,\,j+1,\,\ldots.\,k} + b_{j\,1.2,3,\,\ldots.\,j-1,\,j+1,\,\ldots.k}X_1 + b_{j\,2.1,\,3,\,\ldots.\,j-1,\,j+1,\,\ldots.k}X_2 + \ldots\ldots(j = 1, 2, \ldots. k).$$

Here $a_{j.1,2,\,\ldots.}$, $b_{j1.2,\,3,\,\ldots.}$ ..... etc. are constants. The constant $a_{j.1,2,\,\ldots.}$ is interpreted as the value of X_j when X_2, X_3, ... X_{j-1}, X_{j+1} ..... X_k are all equal to zero. Further, $b_{j1.2,3,\,\ldots.\,j-1,\,j+1,\,\ldots.k}$, $b_{j2.1,3,\,\ldots.\,j-1,\,j+1,\,\ldots.k}$ etc., are (k – 1) partial regression coefficients of regression of X_j on X_1, X_2 ..... X_{j-1}, X_{j+1} ..... X_k.

For simplicity, we shall consider three variables X_1, X_2 and X_3. The three possible regression equations can be written as

$$X_{1c} = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \quad \ldots. (1)$$

$$X_{2c} = a_{2.13} + b_{21.3}X_1 + b_{23.1}X_3 \quad \ldots. (2)$$

$$X_{3c} = a_{3.12} + b_{31.2}X_1 + b_{32.1}X_2 \quad \ldots. (3)$$

Given n observations on X_1, X_2 and X_3, we want to find such values of the constants of the regression equation so that $\sum_{i=1}^{n}\left(X_{ij} - X_{ijc}\right)^2$, j = 1, 2, 3, is minimised.

For convenience, we shall use regression equations expressed in terms of deviations of variables from their respective means. Equation (1), on taking sum and dividing by n, can be written as

$$\frac{\sum X_{1c}}{n} = a_{1.23} + b_{12.3}\frac{\sum X_2}{n} + b_{13.2}\frac{\sum X_3}{n}$$

Notes

or $$\bar{X}_1 = a_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 \quad \text{.... (4)}$$

Note: $\Sigma X_1 = \Sigma X_{1c}$.

Subtracting (4) from (1), we have

$$X_{1c} - \bar{X}_1 = b_{12.3}\left(X_2 - \bar{X}_2\right) + b_{13.2}\left(X_3 - \bar{X}_3\right) \text{ or } x_{1c} = b_{12.3}x_2 + b_{13.2}x_3 \quad \text{.... (5)}$$

where $X_{1c} - \bar{X}_1 = x_{1c}$, $X_2 - \bar{X}_2 = x_2$ and $X_3 - \bar{X}_3 = x_3$.

Similarly, we can write equations (2) and (3) as

$$x_{2c} = b_{21.3}x_1 + b_{23.1}x_3 \quad \text{.... (6)}$$

and $$x_{3c} = b_{31.2}x_1 + b_{32.1}x_2, \text{ respectively.} \quad \text{.... (7)}$$

Notes The subscript of the coefficients preceding the dot are termed as primary subscripts while those appearing after it are termed as secondary subscripts. The number of secondary subscripts gives the order of the regression coefficient, e.g., $b_{12.3}$ is regression coefficient of order one, etc.

Least Square Estimates of Regression Coefficients

Let us first estimate the coefficients of regression equation (5). Given n observations on each of the three variables X_1, X_2 and X_3, we have to find the values of the constants $b_{12.3}$ and $b_{13.2}X_3$ so that is minimised. Using method of least squares, the normal equations can be written as

$$\sum x_1x_2 = b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 \quad \text{.... (8)}$$

$$\sum x_1x_3 = b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 \quad \text{.... (9)}$$

Solving the above equations simultaneously, we get

$$b_{12.3} = \frac{\left(\sum x_1x_2\right)\left(\sum x_3^2\right) - \left(\sum x_1x_3\right)\left(\sum x_2x_3\right)}{\left(\sum x_2^2\right)\left(\sum x_3^2\right) - \left(\sum x_2x_3\right)^2} \quad \text{.... (10)}$$

$$b_{13.2} = \frac{\left(\sum x_1x_3\right)\left(\sum x_2^2\right) - \left(\sum x_1x_2\right)\left(\sum x_2x_3\right)}{\left(\sum x_2^2\right)\left(\sum x_3^2\right) - \left(\sum x_2x_3\right)^2} \quad \text{.... (11)}$$

Using equation (4), we can find $a_{1.23} = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13.2}\bar{X}_3$.

Notes

1. Various sums of squares and sums of products of deviations, used above, can be computed using the formula $\sum x_px_q = \sum X_pX_q - \frac{\left(\sum X_p\right)\left(\sum X_q\right)}{n}$. For example, put p = 1 and q = 2 in the formula to obtain SX_1X_2 and put $p = q = 2$, to obtain $\sum x_2^2$, etc.

Contd...

Notes

2. The fact that a regression coefficient is independent of change of origin can also be utilised to further simplify the computational work.

3. The regression coefficients of equations (2) and (3) can be written by symmetry as given below:

$$b_{21.3} = \frac{\left(\sum x_2x_1\right)\left(\sum x_3^2\right)-\left(\sum x_2x_3\right)\left(\sum x_1x_3\right)}{\left(\sum x_1^2\right)\left(\sum x_3^2\right)-\left(\sum x_1x_3\right)^2}$$

$$b_{23.1} = \frac{\left(\sum x_2x_3\right)\left(\sum x_1^2\right)-\left(\sum x_2x_1\right)\left(\sum x_1x_3\right)}{\left(\sum x_1^2\right)\left(\sum x_3^2\right)-\left(\sum x_1x_3\right)^2}$$

Further, $b_{31.2} = b_{13.2}$ and $b_{32.1} = b_{23.1}$ and the expressions for the constant terms are $a_{2.13} = \bar{X}_2 - b_{21.3}\bar{X}_1 - b_{23.1}\bar{X}_3$ and $a_{3.12} = \bar{X}_3 - b_{31.2}\bar{X}_1 - b_{32.1}\bar{X}_2$ respectively.

Example: Fit a linear regression of rice yield (X_1 quintals) on the use of fertiliser (X_2 kgs per acre) and the amount of rain fall (X_3 inches), from the following data:

X_1	45	50	55	70	75	75	85
X_2	25	35	45	55	65	75	85
X_3	31	28	32	32	29	27	31

Estimate the yield when $X_2 = 60$ and $X_3 = 25$.

Solution:

Calculation Table

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_1^2	X_2^2	X_3^2
45	25	31	1125	1395	775	2025	625	961
50	35	28	1750	1400	980	2500	1225	784
55	45	32	2475	1760	1440	3025	2025	1024
70	55	32	3850	2240	1760	4900	3025	1024
75	65	29	4875	2175	1885	5625	4225	841
75	75	27	5625	2025	2025	5625	5625	729
85	85	31	7225	2635	2635	7225	7225	961
455	**385**	**210**	**26925**	**13630**	**11500**	**30925**	**23975**	**6324**

From the above table we compute the following sums of product and sums of squares:

$$\Sigma x_1x_2 = \Sigma X_1 \Sigma X_2 - \frac{(\Sigma X_1)(\Sigma X_2)}{n} = 26925 - \frac{455 \times 385}{7} = 1900$$

$$\Sigma x_1x_3 = \Sigma X_1X_3 - \frac{(\Sigma X_1)(\Sigma X_3)}{n} = 13630 - \frac{455 \times 210}{7} = -20$$

Notes

$$\Sigma x_2 x_3 = \Sigma X_2 X_3 - \frac{(\Sigma X_2)(\Sigma X_3)}{n} = 11500 - \frac{385 \times 210}{7} = -50$$

$$\Sigma x_2^2 = \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n} = 23975 - \frac{385^2}{7} = 2800$$

$$\Sigma X_3^2 = \Sigma X_3^2 - \frac{(\Sigma X_3)^2}{n} = 6324 - \frac{210^2}{7} = 24$$

Substituting these values in equations (10) and (11), we get

$$b_{12.3} = \frac{1900 \times 24 - (-20) \times (-50)}{2800 \times 24 - (-50)^2} = 0.689$$

$$b_{13.2} = \frac{(-20) \times 2800 - 1900 \times (-50)}{2800 \times 24 - (-50)^2} = 0.603$$

Also $$\bar{X}_1 = \frac{455}{7} = 65, \bar{X}_2 = \frac{385}{7} = 55, \bar{X}_3 = \frac{210}{7} = 30$$

Thus $$a_{1.23} = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13.2}\bar{X}_3 = 65 - 0.689 \times 55 - 0.603 \times 30 = 9.015$$

$\therefore$ The fitted regression of X_1 on X_2 and X_3 is $X_{1c} = 9.015 + 0.689X_2 + 0.603X_3$

The estimate of yield (X_{1c}) when $X_2 = 60$ and $X_3 = 25$ is

$X_{1c} = 9.015 + 0.689 \times 60 + 0.603 \times 25 = 65.43$ quintals.

Alternatively to simplify calculation work, we change origin of the three variable as

$$U_1 = X_1 - 65,\ U_2 = X_2 - 55 \text{ and } U_3 = X_3 - 30.$$

U_1	U_2	U_3	U_1U_2	U_1U_3	U_2U_3	U_1^2	U_2^2	U_3^2
– 20	– 30	1	600	– 20	– 30	400	900	1
– 15	– 20	– 2	300	30	40	225	400	4
– 10	– 10	2	100	– 20	– 20	100	100	4
5	0	2	0	10	0	25	0	4
10	10	– 1	100	– 10	– 10	100	100	1
10	20	– 3	200	– 30	– 60	100	400	9
20	30	1	600	20	30	400	900	1
0	**0**	**0**	**1900**	**– 20**	**– 50**	**1350**	**2800**	**24**

Note: Since $\Sigma U_i = 0$, $\therefore u_i = U_i - \bar{U} = U_i$, $i = 1, 2, 3.$ $\left(\because \bar{U}_i = 0\right)$

Hence $$b_{12.3} = \frac{1900 \times 24 - (-20)(-50)}{2800 \times 24 - (-50)^2} = 0.689$$

$$b_{13.2} = \frac{(-20) \times 2800 - 1900 \times (-50)}{2800 \times 24 - (-50)^2} = 0.603$$

Further, we have

Notes

$$\bar{X}_1 = \frac{\sum U_1}{n} + 65 = 65, \bar{X}_2 = \frac{\sum U_2}{n} + 55 = 55 \text{ and } \bar{X}_3 = \frac{\sum U_3}{n} + 30 = 30$$

Caution The above method should be used when mean of all the variables are integers.

Alternative Method

The coefficients of the regression equation $X_{1c} = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$ can also be obtained by simultaneously solving the following normal equations:

$$\Sigma X_1 = n \cdot a_{1.23} + b_{12.3}\Sigma X_2 + b_{13.2}\Sigma X_3$$

$$\Sigma X_1 X_2 = a_{1.23}\Sigma X_2 + b_{12.3}\Sigma X_2^2 + b_{13.2}\Sigma X_2 X_3$$

$$\Sigma X_1 X_3 = a_{1.23}\Sigma X_3 + b_{12.3}\Sigma X_2 X_3 + b_{13.2}\Sigma X_3^2$$

Self Assessment

Fill in the blanks:

1. Regression coefficient is independent of change of .....
2. In the case of ……regression, one variable is affected by a linear combination of another variable.
3. ……..analysis is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical variable at a time.

13.2 Discriminant Analysis

In this analysis, two or more groups are compared. In the final analysis, we need to find out whether the groups differ one from another.

Example: Where discriminant analysis is used

1. Those who buy our brand and those who buy competitors' brand.
2. Good salesman, poor salesman, medium salesman
3. Those who go to Food World to buy and those who buy in a Kirana shop.
4. Heavy user, medium user and light user of the product.

Suppose there is a comparison between the groups mentioned as above along with demographic and socio-economic factors, then discriminant analysis can be used. One way of doing this is to proceed and calculate the income, age, educational level, so that the profile of each group could be determined. Comparing the two groups based on one variable alone would be informative but it would not indicate the relative importance of each variable in distinguishing the groups. This is because several variables within the group will have some correlation which means that one variable is not independent of the other.

If we are interested in segmenting the market using income and education, we would be interested in the total effect of two variables in combinations, and not their effects separately. Further, we would be interested in determining which of the variables are more important or

Notes

had a greater impact. To summarize, we can say, that Discriminant Analysis can be used when we want to consider the variables simultaneously to take into account their interrelationship.

Like regression, the value of dependent variable is calculated by using the data of independent variable.

$$Z = b_1x_1 + b_2x_2 + b_3x_3 + ...$$

Z = Discriminant score

b_1 = Discriminant weight for variable

x = Independent variable

As can be seen in the above, each independent variable is multiplied by its corresponding weightage.

This results in a single composite discriminant score for each individual. By taking the average of discriminant score of the individuals within a certain group, we create a group mean. This is known as centroid. If the analysis involves two groups, there are two centroids. This is very similar to multiple regression, except that different types of variables are involved.

Application

A company manufacturing FMCG products introduces a sales contest among its marketing executives to find out "How many distributors can be roped in to handle the company's product". Assume that this contest runs for three months. Each marketing executive is given target regarding number of new distributors and sales they can generate during the period. This target is fixed and based on the past sales achieved by them about which, the data is available in the company. It is also announced that marketing executives who add 15 or more distributors will be given a Maruti Omni-van as prize. Those who generate between 5 and 10 distributors will be given a two-wheeler as the prize. Those who generate less than 5 distributors will get nothing. Now assume that 5 marketing executives won a Maruti van and 4 won a two-wheeler.

The company now wants to find out, "Which activities of the marketing executive made the difference in terms of winning a prize and not winning the prize". One can proceed in a number of ways. The company could compare those who won the Maruti van against the others. Alternatively, the company might compare those who won, one of the two prizes against those who won nothing. It might compare each group against each of the other two.

Discriminant analysis will highlight the difference in activities performed by each group members to get the prize. The activity might include:

1. More number of calls made to the distributors.
2. More personal visits to the distributors with advance appointments.
3. Use of better convincing skills.

Discriminant analysis answers the following questions:

1. What variable discriminates various groups as above; the number of groups could be two or more? Dealing with more than two groups is called Multiple Discriminant Analysis (M.D.A.).
2. Can discriminating variables be chosen to forecast the group to which the brand/person/place belong to?
3. Is it possible to estimate the size of different groups?

SPSS Commands for Discriminate Analysis

Notes

Input data has to be typed in an SPSS file.

1. Click on STATISTICS at the SPSS menu bar.
2. Click on CLASSIFY followed by DISCRIMINANT.
3. Dialogue box will appear. Select the GROUPING VARIABLE. This can be done by clicking on the right arrow to transfer them from the variable list on the left to the grouping variable box on the right.
4. Define the range of values by clicking on DEFINE RANGE. Enter Minimum and Maximum value then click CONTINUE.
5. Select all the independent variable for discriminant analysis from the variable list by clicking on the arrow that transfers them to box on the right.
6. Click on STATISTICS on the lower part of main dialogue box. This will open up a smaller dialogue box.
7. Click on CLASSIFY on the lower part of the main dialogue box select SUMMARY TABLE under the heading DISPLAY in a small dialogue box that appears.
8. Click OK to get the discriminant analysis output.

Self Assessment

Fill in the blanks:

4. In discriminant analysis, ... groups are compared.
5. If the discriminant analysis involves two groups, there are ..... centroids.

13.3 Conjoint Analysis

Conjoint analysis is concerned with the measurement of the joint effect of two or more attributes that are important from the customers' point of view. In a situation where the company would like to know the most desirable attributes or their combination for a new product or service, the use of conjoint analysis is most appropriate.

Example: An airline would like to know, which is the most desirable combination of attributes to a frequent traveller: (a) Punctuality (b) Air fare (c) Quality of food served on the flight and (d) Hospitality and empathy shown.

Conjoint Analysis is a multivariate technique that captures the exact levels of utility that an individual customer places on various attributes of the product offering. Conjoint Analysis enables a direct comparison,

Example: A comparison between the utility of a price level of ₹ 400 versus ₹ 500, a delivery period of 1 week versus 2 weeks, or an after-sales response of 24 hours versus 48 hours.

Once we know the utility levels for each attribute (and at individual levels as well), we can combine these to find the best combination of attributes that gives the customer the highest utility, the second best combination that gives the second highest utility, and so on. This information is then used to design a product or service offering.

Notes

Application

Conjoint Analysis is extremely versatile and the range of applications includes virtually in any industry. New product or service design, including the concepts in the pre-prototyping stage can specifically benefit from the conjoint applications.

Some examples of other areas where this technique can be used are:

1. Designing an automobile loan or insurance plan in the insurance industry,
2. Designing a complex machine for business customers.

Process

Design attributes for a product are first identified. For a shirt manufacturer, these could be design such as designer shirts vs plain shirts, this price of ₹ 400 versus ₹ 800. The outlets can have exclusive distribution or mass distribution. All possible combinations of these attribute levels are then listed out. Each design combination will be ranked by customers and used as input data for Conjoint Analysis. Then the utility of the products relative to price can be measured.

The output is a part-worth or utility for each level of each attribute. For example, the design may get a utility level of 5 and plain, 7.5. Similarly, the exclusive distribution may have a part utility of 2, and mass distribution, 5.8. We then put together the part utilities and come up with a total utility for any product combination we want to offer, and compare that with the maximum utility combination for this customer segment.

This process clarifies to the marketer about the product or service regarding the attributes that they should focus on in the design.

If a retail store finds that the height of a shelf is an important attribute for selling at a particular level, a well-designed shelf may result from this knowledge. Similarly, a designer of clocks will benefit from knowing the utility attached by customers to the dial size, background colours, and price range of the clocks.

Approach

From a discussion with the client, identify the design attributes to be studied and the levels at which they can be offered. Then build a list of product concepts on offer. These product concepts are then ranked by customers. Once this data is available, use Conjoint Analysis to derive the part utilities of each attribute level. This is then used to predict the best product design for the given customer segment. Use the SPSS Conjoint procedure to analyse the data.

There are three steps in conjoint analysis:

1. Identification of relevant products or service attributes.
2. Collection of data.
3. Estimation of worth for the attribute chosen.

For attributes selection, the market researcher can conduct interview with the customers directly.

Example: Example of conjoint analysis for a Laptop:

For a laptop, consider 3 attributes:

1. Weight (3 Kg or 5 Kg)
2. Battery life (2 hours or 4 hours)
3. Brand name (Lenovo or Dell)

Notes

SPSS Command for Conjoint Analysis

SPSS commands for conjoint Analysis. A data file is to be created containing all possible attribute combination.

1. Ask each of the respondent to rank all the combination of attributes contained in the file. This is nomenclated at DATA FILE 1. All the rankings should be entered in another file called DATA FILE 2.
2. Now 2 files namely DATA FILE 1 and DATA FILE 2 are created.
3. A third file called SYNTAX file is to be opened. By using the FILE, OPEN command followed by syntax.
4. Type the following - conjoint plan = DATA FILE 1 SAV/DATA' DATA FILE 2 SAV/ SCORES=SCORE 1 to Score number of ranking/FACTOR VARI (DISCRETE)/PLOT ALL (Here 25 is the possible combination of attributes). Score is the term used for rankings. The no of scores will be equal to number of rankings. We should use the word RANK in the syntax instead of scores if Rankings are contained in the data file.
5. Click RUN from the menu of the syntax file that was created click all in the menu which appears on the screen. If the syntax is correct, the output for conjoint will appear.

Task Rank order the following combination of these characteristics:

1 = Most preferred, 8 = Least preferred

Combination	Rank
3 Kg, 2 hours, Lenovo	4
5 Kg, 4 hours, Dell	5
5 Kg, 2 hours, Lenovo	8
3 Kg, 4 hours, Lenovo	3
3 Kg, 2 hours, Dell	2
5 Kg, 4 hours, Lenovo	7
5 Kg, 2 hours, Dell	6
3 Kg, 4 hours, Dell	1

One combination 3 kg, 4 hours, Dell clearly dominates and 5 kg, 2 hours, Lenovo is least preferred.

Let us now take the average rank for 3 kg option = 4 + 3 + 2 + 1/4 = 2.5

For 5 kg option average rank is 5 + 8 + 7 + 6/4 = 6.5

For 4 hour option 5 + 3 + 7 + 1/4 = 4

For 2 hour option 4 + 8 + 2 + 6/4 = 5

For Dell 5 + 6 + 1 + 2/4 = 3.5

For Lenovo 5.5

Looking at the difference in average ranks, the most important characteristic to this respondent is weight = 4, followed by brand name = 2 and battery life = 1.

Notes

Self Assessment

Fill in the blanks:

6. ..... analysis is concerned with the measurement of the joint effect of two or more attributes.
7. For ..... selection, the market researcher can conduct interview with the customers directly.
8. The ..... is a part-worth or utility for each level of each attribute.

13.4 Factor Analysis

The main purpose of Factor Analysis is to group large set of variable factors into fewer factors. Each factor will account for one or more component. Each factor a combination of many variables. There are two most commonly employed factor analysis procedures or methods. They are:

1. Principle component analysis
2. Common factor analysis.

When the objective is to summarise information from a large set of variables into fewer factors, principle component factor analysis is used. On the other hand, if the researcher wants to analyse the components of the main factor, common factor analysis is used.

Example: Common factor – Inconvenience inside a car. The components may be:

1. Leg room
2. Seat arrangement
3. Entering the rare seat
4. Inadequate dickey space
5. Door locking mechanism.

13.4.1 Principle Component Factor Analysis

Purposes: Customer feedback about a two-wheeler manufactured by a company.

Method: The MR manager prepares a questionnaire to study the customer feedback. The researcher has identified six variables or factors for this purpose. They are as follows:

1. Fuel efficiency (A)
2. Durability (Life) (B)
3. Comfort (C)
4. Spare parts availability (D)
5. Breakdown frequency (E)
6. Price (F)

The questionnaire may be administered to 5,000 respondents. The opinion of the customer is gathered. Let us allot points 1 to 10 for the variables factors A to F. 1 is the lowest and 10 is the highest. Let us assume that application of factor analysis has led to grouping the variables as follows:

Notes

A, B, D, E into factor-1

F into Factor -2

C into Factor - 3

Factor - 1 can be termed as Technical factor;

Factor - 2 can be termed as Price factor;

Factor - 3 can be termed as Personal factor.

For future analysis, while conducting a study to obtain customers' opinion, three factors mentioned above would be sufficient. One basic purpose of using factor analysis is to reduce the number of independent variables in the study. By having too many independent variables, the M.R study will suffer from following disadvantages:

1. Time for data collection is very high due to several independent variables.
2. Expenditure increases due to the time factor.
3. Computation time is more, resulting in delay.
4. There may be redundant independent variables.

Did u know? **What is correspondence analysis?**

Correspondence analysis is a descriptive/exploratory technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns.

The results provide information which is similar in nature to those produced by Factor Analysis techniques, and they allow one to explore the structure of categorical variables included in the table. The most common kind of table of this type is the two-way frequency cross-tabulation table.

In a typical correspondence analysis, a cross-tabulation table of frequencies is first standardized, so that the relative frequencies across all cells sum to 1.0. One way to state the goal of a typical analysis is to represent the entries in the table of relative frequencies in terms of the distances between individual rows and/or columns in a low-dimensional space.

Example: Following are the data on the drinking habits of different employees in an organization:

	Drinking Habits				
Employee Group	**(1) None**	**(2) Light**	**(3) Medium**	**(4) Heavy**	**Row Totals**
(1) Senior Level Management	5	2	4	3	14
(2) Middle Level Management	4	2	5	9	20
(3) Junior Level Management	15	12	10	5	42
(4) Executives	25	20	30	15	90
(5) Other Employees	30	5	10	5	50
Column Totals	**79**	**41**	**59**	**37**	**216**

One may think of the 4 column values in each row of the table as coordinates in a 4-dimensional space, and one could compute the (Euclidean) distances between the 5 row points in the 4-

Notes

dimensional space. The distances between the points in the 4-dimensional space summarize all information about the similarities between the rows in the table above. Now suppose one could find a lower-dimensional space, in which to position the row points in a manner that retains all, or almost all, of the information about the differences between the rows. You could then present all information about the similarities between the rows (types of employees in this case) in a simple 1, 2, or 3-dimensional graph. While this may not appear to be particularly useful for small tables like the one shown above, one can easily imagine how the presentation and interpretation of very large tables (e.g., differential preference for 10 consumer items among 100 groups of respondents in a consumer survey) could greatly benefit from the simplification that can be achieved via correspondence analysis (e.g., represent the 10 consumer items in a two-dimensional space).

13.4.2 Rotation in Factor Analysis

Rotation is the step in factor analysis that permits you to identify meaningful factor names or descriptions like these.

Linear Functions of Predictors

To identify with rotation, first consider a problem that doesn't involve factor analysis. Suppose you want to predict the grades of college students (all in the same college) in many dissimilar courses, from their scores on general "verbal" and "math" skill tests. To build up predictive formulas, you have a body of past data consisting of the grades of numerous hundred previous students in these courses, plus the scores of those students on the math and verbal tests. To predict grades for present and future students, you might use these data from past students to fit a series of two-variable multiple regressions, each regression forecasting grade in one course from scores on the two skill tests.

At present suppose a co-worker suggests summing each student's verbal and math scores to obtain a composite "academic skill" score I'll call AS, and taking the difference among each student's verbal and math scores to obtain a second variable I'll call VMD (verbal-math difference). The co-worker advises running the same set of regressions to predict grades in individual courses, except using AS and VMD as predictors in each regression, instead of the original verbal and math scores. In this instance, you would get exactly the same predictions of course grades from these two families of regressions: one predicting grades in individual courses from verbal and math scores, the other predicting the identical grades from AS and VMD scores. In fact, you would get the same predictions if you formed composites of 3 math + 5 verbal and 5 verbal + 3 math, and ran a series of two-variable multiple regressions forecasting grades from these two composites. These examples are all *linear functions* of the original verbal and math scores.

The vital point is that if you have *m* predictor variables, and you replace the *m* original predictors by *m* linear functions of those predictors, you usually neither gain nor lose any information—you could if you wish use the scores on the linear functions to rebuild the scores on the original variables. But multiple regression uses whatever information you have in the optimum way (as measured by the sum of squared errors in the current sample) to forecast a new variable (e.g. grades in a particular course). Since the linear functions contain the same information as the original variables, you get the similar predictions as before.

Specified that there are lots of ways to get exactly the same predictions, is there any advantage to using one set of linear functions rather than another? Yes there is; one set might be *simpler* than another. One particular pair of linear functions may enable many of the course grades to be forecasted from just one variable (that is, one linear function) rather than from two. If we regard regressions with less predictor variables as simpler, then we can ask this question: Out of all the

Notes

possible pairs of predictor variables that would give the same predictions, which is simplest to use, in the logic of minimizing the number of predictor variables needed in the typical regression? The pair of predictor variables maximising some measure of minimalism could be said to have *simple structure*. In this example involving grades, you might be able to predict grades in some courses correctly from just a verbal test score, and predict grades in other courses accurately from just a math score. If so, then you would have achieved a "simpler structure" in your predictions than if you had used both tests for each and every predictions.

Simple Structure in Factor Analysis

The points of the preceding section are relevant when the predictor variables are factors. Think of the *m* factors F as a set of independent or predictor variables, and imagine of the *p* observed variables X as a set of dependent or criterion variables. Think a set of *p* multiple regressions, each predicting one of the variables from all *m factors*. The standardized coefficients in this set of regressions structure a *p* x *m* matrix called the *factor loading matrix*. If we replaced the original factors by a set of linear functions of those factors, we would get just the same predictions as before, but the factor loading matrix would be different. So we can ask which, of the many possible sets of linear functions we might use, produces the simplest factor loading matrix. Specially we will define simplicity as the number of zeros or near-zero entries in the factor loading matrix—the more zeros, the simpler the structure. Rotation does not alter matrix C or U at all, but does transform the factor loading matrix.

In the intense case of simple structure, each X-variable will have merely one large entry, so that all the others can be ignored. But that would be a simpler structure than you would usually expect to achieve; after all, in the real world each variable isn't in general affected by only one other variable. You then name the factors subjectively, based on an examination of their loadings.

In common factor analysis the procedure of rotation is in fact somewhat more abstract that I have implied here, since you don't actually know the individual scores of cases on factors. However, the statistics for a multiple regression that is mainly relevant here—the multiple correlation and the standardized regression slopes—can all be calculated just from the correlations of the variables and factors involved. So we can base the calculations for rotation to simple structure on just those correlations, devoid of using any individual scores.

A rotation which necessitates the factors to remain uncorrelated is an *orthogonal* rotation, while others are *oblique* rotations. Oblique rotations regularly achieve greater simple structure, though at the cost that you have to also consider the matrix of factor intercorrelations when interpreting results. Manuals are usually clear which is which, but if there is ever any ambiguity, a simple rule is that if there is any capability to print out a matrix of factor correlations, then the rotation is oblique, as no such capacity is needed for orthogonal rotations.

Self Assessment

Fill in the blanks:

9. When the objective is to summarise information from a large set of variables into fewer factors, ... analysis is used.
10. Correspondence analysis is a ..... technique.
11. In a typical correspondence analysis, a cross-tabulation table of frequencies is first .....

Notes

13.5 Cluster Analysis

Cluster Analysis is used:

1. To classify persons or objects into small number of clusters or group.
2. To identify specific customer segment for the company's brand.

Cluster Analysis is a technique used for classifying objects into groups. This can be used to sort data (a number of people, companies, cities, brands or any other objects) into homogeneous groups based on their characteristics.

The result of Cluster Analysis is a grouping of the data into groups called clusters. The researcher can analyse the clusters for their characteristics and give the cluster, names based on these.

Where can Cluster Analysis be applied?

The marketing application of cluster analysis is in customer segmentation and estimation of segment sizes. Industries, where this technique is useful include automobiles, retail stores, insurance, B-to-B, durables and packaged goods. Some of the well-known frameworks in consumer behaviour (like VALS) are based on value cluster analysis.

Cluster Analysis is applicable when:

1. An FMCG company wants to map the profile of its target audience in terms of life-style, attitude and perceptions.
2. A consumer durable company wants to know the features and services a consumer takes into account, when purchasing through catalogues.
3. A housing finance corporation wants to identify and cluster the basic characteristics, life-styles and mindset of persons who would be availing housing loans. Clustering can be done based on parameters such as interest rates, documentation, processing fee, number of installments etc.

Process

There are two ways in which Cluster Analysis can be carried out:

1. First, objects/respondents are segmented into a pre-decided number of clusters. In this case, a method called non-hierarchical method can be used, which partitions data into the specified number of clusters
2. The second method is called the hierarchical method.

The above two are basic approaches used in cluster analysis. This can be used to segment customer groups for a brand or product category, or to segment retail stores into similar groups based on selected variables.

Interpretation of Results

Ideally, the variables should be measured on an interval or ratio scale. This is because the clustering techniques use the distance measure to find the closest objects to group into a cluster. An example of its use can be clustering of towns similar to each other which will help decide where to locate new retail stores.

If clusters of customers are found based on their attitudes towards new products and interest in different kinds of activities, an estimate of the segment size for each segment of the population can be obtained, by looking at the number of objects in each cluster.

Notes

Marketing strategies for each segment are fine-tuned based on the segment characteristics. For instance, a segment of customers, like sports car, get a special promotional offer during specific period.

Example: In cluster analysis, the following five steps to be used:

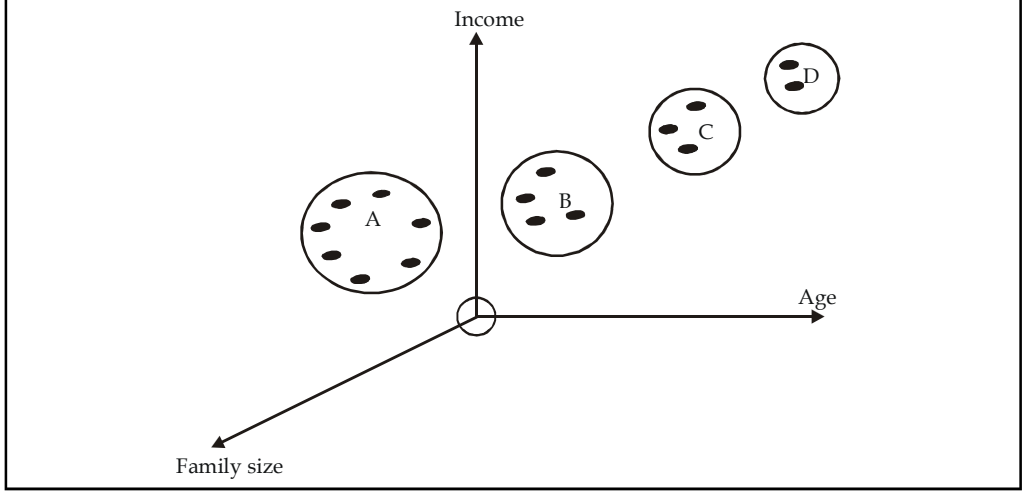
1. Selection of the sample to be clustered (buyers, products, employees).
2. Definition on which the measurement to be made (E.g.: product attributes, buyer characteristics, employees' qualification).
3. Computing the similarities among the entities.
4. Arrange the cluster in a hierarchy.
5. Cluster comparison and validation.

Did u know? Names can also be given to clusters to describe each one. For example, there can be a cluster called "neo-rich". Segments are prioritised based on their estimated size.

Cluster Analysis on Three Dimensions

The example below shows Cluster Analysis based on three dimensions age, income and family size. Cluster Analysis is used to segment the car-buying population in a Metro. For example "A" might represent potential buyers of low end cars. Example: Maruti 800 (for common man). These are people who are graduating from the two-wheeler market segment. Cluster "B" may represent mid-population segment buying Zen, Santro, Alto etc. Cluster "C" represents car buyers, who belong to upper strata of society. Buyers of Lancer, Honda city etc. Cluster "D" represents the super-rich cluster, i.e. Buyers of Benz, BMW, etc.

Figure 13.1: Matching Measure



Example: Suppose there are five attributes, 1 to 5, on which we are judging two objects A and B. The existence of an attribute may be indicated by 1 and its absence by 0. In this way, two objects are viewed as similar if they share common attributes.

Notes

Attribute	1	2	3	4	5	6	7
Brand - A	1	0	0	1	0	0	1
Brand - B	0	0	1	1	1	0	0

One measure of simple matching S is given by:

$$S = \frac{a+d}{a+b+c+d}$$

Where

a = No. of attributes possessed by brands A and B

b = No. of attributes possessed by brand A but not by brand B

c = No. of attributes possessed by brand B but not by brand A

d = No. of attributes not possessed by both brands.

Substituting, we get $S = \frac{1+2}{1+2+2+2} = \frac{3}{7} = 0.43$

A and B's association is to be the extent of 43%.

It is now clear that object A possess attributes 1, 4, and 7 while object B possess the attributes 3, 4 and 5. A glance at the above table will indicate that objects A and B are similar in respect of 2 (0 & 0), 6 (0 & 0) and 4 (1 & 1). In respect of other attributes, there is no similarity between A and B. Now we can arrive at a simple matching measure by (a) counting up the total number of matches - either 0, 0 or 1, (b) dividing this number by the total number of attributes.

Symbolically SAB = M/N

SAB = Similarity between A and B

M = Number of attributes held in common (0 or 1)

N = Total number of attributes

SAB = 3/7 = 0.43

i.e., A & B are similar to the extent of 43%.

SPSS Command for Cluster Analysis

Stage 1

Enter the input data along with variable and value labels in an SPSS file.

1. Click on STATISTICS at the spss menu bar.
2. Click on CLASSIFY followed by HIERARCHICAL CLUSTER.
3. Dialogue box will appear select all the variables which are required to be used in cluster analysis. This can be done by clicking on the right arrow to transfer them from the variable list on the left.
4. Click on METHOD. The dialogue box will open. Choose "Between Groups Linkage" as the CLUSTER METHOD.
5. Click CONTINUE to return to main dialogue box.
6. Click STATISTICS on the main dialogue box. Choose "Agglomeration schedule" so that it will appear in the final output click CONTINUE.

Notes

7. Choose DENDROGRAM then on the box called ICICLE, Choose "All Clusters" and "Vertical".
8. Click OK on the main dialogue box to get the output of the hierarchical cluster analysis.

Stage 2

This stage is used to know how many clusters are required. This stage is called K- MEANS CLUSTERING.

1. Click CLASSIFY, followed by K- FANS CLUSTER desired.
2. Fill in the desired number of clusters that has been identified from stage 1.
3. Click OPTIONS on the main dialogue box. Select "Initial Cluster Centers". Then click CONTINUE to return to the main dialogue box.
4. Click OK on the main dialogue box to get the output which has final clusters.

Self Assessment

Fill in the blanks:

12. ..... Analysis is a technique used for classifying objects into groups.
13. The ..... application of cluster analysis is in customer segmentation and estimation of segment sizes.

13.6 Multidimensional Scaling (MDS)

In addition to fulfilling the goals of detecting underlying structure and data reduction that is shares with other methods, multidimensional scaling (MDS) provides the researcher with a spatial representation of data that can facilitate interpretation and reveal relationships. Therefore, we can define MDS as "a set of multivariate statistical methods for estimating the parameters in and assessing the fit of various spatial distance models for proximity data."

The spatial display of data provided by MDS is why it is also sometimes referred to as perceptual mapping. MDS has much more flexibility about the types of data that can be used to generate the solution. Almost any measures of similarity and dissimilarity can be used, depending on what your statistical computer software will accept.

Types of MDS

In general, there are two types of MDS:

1. Metric
2. Non-metric

Metric MDS makes the assumption that the input data is either ratio or interval data, while the non-metric model requires simply that the data be in the form of ranks. Therefore, the non-metric model has more fewer restrictions than the metric model, but also less rigor. One technique to use if you are unsure whether your data is ordinal or can be considered interval is to try both metric and non-metric models. If the results are very close, the metric model may be used.

An advantage of the non-metric models is that they permit the researcher to categorize and examine preference data, such as the kind obtained in marketing studies or other areas where comparisons are useful.

Another technique, correspondence analysis, can work with categorical data, i.e., data at the nominal level of measurement, however that technique will not be described here.

Notes

Notes **Similarities and Differences between Factor Analysis and MDS**

We have already seen that MDS can accept more different measures of similarity and dissimilarity than factor analysis techniques can. In addition, there are some differences in terminology. These differences reflect the origin of MDS in the field of psychology. The measure corresponding to factors are called alternatively dimensions or stimulus coordinates.

The output of MDS looks very similar to that of factor analysis and the determination of the optimal number of dimensions is handled in much the same way.

Steps in using MDS

There are four basic steps in MDS:

1. Data collection and formation of the similarity/dissimilarity matrix
2. Extraction of stimulus coordinates
3. Decision about the number of stimulus coordinates that represent the data
4. Rotation and interpretation

Example: Let us say that you have a matrix of distances between a number of major cities, such as you might find on the back of a road map. These distances can be used as the input data to derive an MDS solution. When the results are mapped in two dimensions, the solution will reproduce a conventional map, except that the MDS plot might need to be rotated so that the north-south and east-west dimensions conform to expectations. However, the once the rotation is completed, the configuration of the cities will be spatially correct.

Self Assessment

Fill in the blanks:

14. An advantage of the non-metric models is that they permit the researcher to ..... and ..... preference data.
15. The spatial display of data provided by MDS is also sometimes referred to as ……..

13.7 Summary

- Multivariate analysis is used if there are more than 2 variables.
- Some of the multi variate analysis are discriminant analysis, Factor analysis, Cluster analysis, conjoint analysis, and multi dimensional scaling.
- In discriminant analysis, it is verified whether the 2 groups differ from one another.
- Factor analysis is used to reduce large no of various factors into fewer variables cluster analysis is used to segmenting the market or to identify the target group.
- Regression is a term used for predicting the value of one variable from the other.
- Least square method is used to fit the line.

Notes

- MDS as a set of multivariate statistical methods for estimating the parameters in and assessing the fit of various spatial distance models for proximity data.
- The output of MDS looks very similar to that of factor analysis and the determination of the optimal number of dimensions is handled in much the same way.

13.8 Keywords

Cluster Analysis: Cluster Analysis is a technique used for classifying objects into groups.

Conjoint Analysis: Conjoint analysis is concerned with the measurement of the joint effect of two or more attributes that are important from the customers' point of view.

Discriminant Analysis: In this analysis, two or more groups are compared. In the final analysis, we need to find out whether the groups differ one from another.

Factor Analysis: Factor Analysis is the analysis whose main purpose is to group large set of variable factors into fewer factors.

Multivariate Analysis: In multi variate analysis, the number of variables to be tackled are many.

13.9 Review Questions

1. Which technique would you use to measure the joint effect of various attributes while designing an automobile loan and why?
2. Do you think that the conjoint analysis will be useful in any manner for an airline? If yes how, if no, give an example where you think the technique is of immense help.
3. In your opinion, what are the main advantages of cluster analysis?
4. Which analysis would you use in a situation when the objective is to summarise information from a large set of variables into fewer factors? What will be the steps you would follow?
5. Which analysis would answer if it is possible to estimate the size of different groups?
6. Which analysis would you use to compare a good, bad and a mediocre doctor and why?
7. Analyse the weakness of principle component factor analysis.
8. Which multivariate analysis would you apply to identify specific customer segment for a company's brand and why?
9. Critically evaluate multidimensional scaling.
10. In your opinion what will be the disadvantages of having too many independent variables in an MR study?
11. The following constructs of intelligence measures of management students:

Variable	Load
CGPA	0.60 x F
Problem Solving Skills	0.75 x F
Communication Skills	0.85 x F

This table tells us communication skill score loads highly on intelligence factor of management students, followed by problem solving skills and CGPA. These loads or

Notes

weights are correlations, i.e., the correlations between communication skills and the factor. But here we have only three variables and only one factor. In real life we may have many variables and more factors. Whatever may be the case, the basic ideas remain the same.

Suppose we want to recruit management trainees from the campus and as a selection process, we need to consider the following variables.

X_1 = CGPA

X_2 = Problem Solving Skills

X_3 = Communication Skills

X_4 = Knowledge Test Score

X_5 = GD Score

X_6 = Personal Interview Score

12. People have been rated on their suitability for an advanced training course in computer programming on the basis of six ratings given by their manager (rated 1=low to 20=high):

(a) Intellect

(b) Interest in doing the course

(c) Experience of computer programming

(d) Likelihood of them staying with the company

(e) Commitment to the company

(f) Loyalty to their team and two other ratings:

(g) Number of GCSEs

(h) Score on a computer programming aptitude test

The training department believe that these are really measuring only three things; intellect, computer programming experience and loyalty, and want you to carry out a factor analysis to explore that hypothesis. Describe the decisions you would have to make in carrying out a factor analysis and what the results would be likely to tell you.

13. Six observations on two variables are available, as shown in the following table:

Obs.	X_1	X_2
a	3	2
b	4	1
c	2	5
d	5	2
e	1	6
f	4	2

(a) Plot the observations in a scatter diagram. How many groups would you say there are, and what are their members?

(b) Apply the nearest neighbor method and the squared Euclidean distance as a measure of dissimilarity. Use a dendrogram to arrive at the number of groups and their membership.

14. Six observations on two variables are available, as shown in the following table:

Notes

Obs.	X_1	X_2
a	-1	-2
b	0	0
c	2	2
d	-2	-2
e	1	-1
f	1	2

(a) Plot the observations in a scatter diagram. How many groups would you say there are, and what are their members?

(b) Apply the nearest neighbor method and the Euclidean distance as a measure of dissimilarity.

Answers: Self Assessment

1. origin
2. simple linear
3. Multivariate
4. two or more
5. two
6. Conjoint
7. attributes
8. output
9. principle component factor
10. descriptive/exploratory
11. standardized
12. Cluster
13. marketing
14. categorize, examine
15. perceptual mapping.

13.10 Further Readings

Books

A Parasuraman, Dhruv Grewal, *Marketing Research*, Biztantra

Cisnal Peter, *Marketing Research*, MCGE.

Hague & Morgan, *Marketing Research in Practice*, Kogan page.

Paneerselvam, R, *Research Methods*, PHI.

Tull and Donalds, *Marketing Research*, MMIL.

Notes

Unit 14: Report Writing

CONTENTS

Objectives

After studying this unit, you will be able to:

- Explain the meaning and characteristics of research report
- Recognize the significance of report writing
- Describe the techniques and precaution of interpretation
- Discuss the layout of report
- Categorize different types of report

Introduction

Notes

A report is a very formal document that is written for a variety of purposes, generally in the sciences, social sciences, engineering and business disciplines. Generally, findings pertaining to a given or specific task are written up into a report. It should be noted that reports are considered to be legal documents in the workplace and, thus, they need to be precise, accurate and difficult to misinterpret.

There are three features that, together, characterize report writing at a very basic level: a predefined structure, independent sections, and reaching unbiased conclusions.

- ***Predefined structure:*** Broadly, these headings may indicate sections within a report, such as an introduction, discussion, and conclusion.
- ***Independent sections:*** Each section in a report is typically written as a stand-alone piece, so the reader can selectively identify the report sections they are interested in, rather than reading the whole report through in one go from start to finish.
- ***Unbiased conclusions:*** A third element of report writing is that it is an unbiased and objective form of writing.

14.1 Characteristics of Research Report

Characteristics feature is an integral part of the report. There is no hard and fast rule for preparing a research report. The research report will differ based on the need of the particular managers using the report. The report also depends on the philosophy of the researcher.

Example: A report prepared for a government agency will be different from the one prepared for a private organization.

In spite of the fact that, marketing report is influenced by the researcher, there are certain characteristics which the report should possess, if it is to be effectively communicated. These characteristics can be classified as:

i. Substantive characteristics

ii. Semantic characteristics.

14.1.1 Substantive Characteristics

Substantive characteristics are:

- Accuracy
- Currency
- Sufficiency
- Availability
- Relevancy

The more that the report possesses the above characteristics, the greater is its practical value in decision making.

Accuracy: Accuracy refers to the degree to which information reflects reality. Specifically, research report must accurately present both research procedure and research results. Even if the research results are not as per the expectation of the management, the researcher has the professional

Notes

obligation to present the findings accurately and objectively. Less accurate report means, injustice to the management.

Currency: Currency refers to the time span between completion of the research project and presentation of the research report to management. If the management receives the research report too late, the results are no longer valid due to environmental changes, and then the report will have no or little value for decision making. Currency is one of the reasons for orally or informally communicating preliminary research results to management to ensure timely decision making.

Sufficiency: The research report must have sufficient details, so that important and valid decision can be made. Sometimes the sample size, sample representativeness may act as a constraint for sufficient details not being available.

Example: Data required by the management, say segment wise market, whereas overall market data is available.

Notes A research report must document methodology and techniques used so that an assessment can be made regarding validity, reliability and generalizability. Therefore, sufficiency refers to whether enough information is present in the research report to enable the manager to take valid decision.

It should be remembered that sufficiency characteristic does not mean that all possible research project information must be incorporated in the research report. A researcher should include in a report only that information, which is necessary to convey complete perspective of the research project.

Availability: The fourth important characteristic of research report is that, it is available to the appropriate decision maker when they need it. Availability refers to the communication process between researcher and the decision maker. We use the word 'appropriate decision maker' to emphasize the fact "who should or who should not have access to the report". This decision is made by the management, and it is the duty of the researcher to carry out this decision. Most reports carry confidential information. Therefore, it is necessary to restrict the report availability, to individuals as well as outside of an organization to prevent the competitor from having access to it.

Relevancy: The research report should be confined to the decision issue researched. Sometimes the researcher might include some information, which he thinks is interesting, but may not have any relevance. This type of information should be excluded from the report.

Example: A researcher may be preparing a report on the audience perception of RJs (Radio Jockeys). This may be done with a view to recruit them based on the perception. In this context, a lengthy commentary on relative audience appeal of each radio station is included. This type of data may be readily available from some research agency, who is selling commercial data. Therefore, including this type of aspect may not be necessary.

14.1.2 Semantic Characteristics

Semantic characteristics are equally important in report. The report should be grammatically correct. It should be free from spelling and typing errors. This will ensure that there is no ambiguity or misunderstanding. Assistance of a proof reader, other than the researcher would be required to eliminate the above errors.

Notes

i. Creative expressions in the form of superlatives, similes should be avoided.

ii. The report should be concise.

iii. Jargon of any kind should be avoided.

iv. Common words with multiple meaning should be avoided.

v. Language of the report must be simple. For example, sentences like "illumination must be extinguished when premises are not in use" can be expressed in simple words say "switch off the lights when you leave".

vi. Avoid using 'I' 'we'. The report should be more impersonal.

vii. Sometimes, the current research uses the data of research conducted in the past. In this case it is better to use past tense than present tense.

The following are the hindrances for clarity of any research report.

- Ambiguity
- Jargon
- Misspelled words
- Excessive prediction
- Improper punctuation
- Unfamiliar words
- Clerical error

Some of the illustrations that can cause inaccuracy in report writing are given below:

- ***Addition/subtraction error:*** Assume that a survey was conducted to ascertain the income of various strata of population in a city. Suppose, it is found that 15% belong to super rich, 18% belong to rich class, 61% belong to middle class.

  By oversight the total is recorded as (15+61+18) which is not equal to hundred. This error can be corrected easily by the researcher. This type of error leads to confusion because the reader or decision maker does not know which categories are left out (may be lower middle class and lower class).

- ***Confusion between percentage and percentage points:*** Suppose the report indicates that raw material cost of a product as a percentage of total cost increased from 8 percentage points in 2003 to 10 percentage points in 2009. Therefore, the raw material cost has increased by only 2 percentage points in 6 years. The real increase is 2 percentage points or 25 percent.

- ***Wrong conclusion:*** Mr. X annual income has increased from ₹ 20,000 to ₹ 40,000 in 8 years. Therefore, the conclusion is, since income has doubled, the purchasing power also has doubled. This may not be true because due to inflation in 8 years, purchasing power might come down or money value could get eroded.

Self Assessment

Fill in the blanks:

1. The research report will differ based on the ……of the particular managers using the report.

2. Accuracy refers to the degree to which information reflects……..

Notes

3. Availability refers to the communication process between researcher and the……..

4. …….refers to the time span between completion of the research project and presentation of the research report to management

14.2 Significance of Report Writing

Preparation and presentation of a research report is the most important part of the research process. No matter how brilliant the hypothesis and how well designed is the research study, they are of little value unless communicated effectively to others in the form of a research report. Moreover, if the report is confusing or poorly written, the time and effort spent on gathering and analysing data would be wasted. It is therefore, essential to summarise and communicate the result to the management in the form of an understandable and logical research report.

Research report is regarded as a major component of the research study for the research task remains unfinished till the report has been presented and/or written. As a matter of fact even the most brilliant hypothesis, very well designed and conducted research study, and the most striking generalizations and findings are of modest value unless they are effectively communicated to others. The rationale of research is not well served unless the findings are made known to others. Research results must customarily enter the general store of knowledge. All this explains the importance of writing research report. There are people who do not consider writing of report as an essential part of the research process. But the general opinion is in favour of treating the presentation of research results or the writing of report as division and parcel of the research project. Writing of report is the final step in a research study and requires a set of skills somewhat different from those called for in respect of the former stages of research. This task should be accomplished by the researcher with extreme care; he may seek the assistance and guidance of experts for the reason.

Self Assessment

Fill in the blanks:

5. ……is regarded as a major component of the research study

6. Writing of report is the ……..step in a research study and requires a set of skills somewhat different from those called for in respect of the former stages of research.

14.3 Techniques and Precautions of Interpretation

Interpretation means bringing out the meaning of data. We can also say that interpretation is to convert data into information. The essence of any research is to do interpretation about the study. This requires a high degree of skill. There are two methods of drawing conclusions (i) induction (ii) deduction.

In the induction method, one starts from observed data and then generalisation is done which explains the relationship between objects observed.

On the other hand, deductive reasoning starts from some general law and is then applied to a particular instance i.e., deduction comes from the general to a particular situation.

Example:

Example of Induction: All products manufactured by Sony are excellent. DVD player model 2602 MX is made by Sony. Therefore, it must be excellent.

Notes

Example of Deduction: All products have to reach decline stage one day and become obsolete. This radio is in decline mode. Therefore, it will become obsolete.

During the inductive phase, we reason from observation. During the deductive phase, we reason towards the observation. Successful interpretation depends on how well the data is analysed. If data is not properly analysed, the interpretation may go wrong. If analysis has to be corrected, then data collection must be proper. Similarly, if the data collected is proper but analysed wrongly, then too the interpretation or conclusion will be wrong. Sometimes, even with the proper data and proper analysis, the data can still lead to wrong interpretation. Interpretation depends upon the experience of the researcher and methods used by him for interpretation.

Did u know? Both logic and observation are essential for interpretation.

Example: A detergent manufacturer is trying to decide which of the three sales promotion methods (discount, contest, buy one get one free) would be most effective in increasing the sales. Each sales promotion method is run at different times in different cities. The sales obtained by the different sale promotion methods is as follows.

Sales Impact of Different Sale Promotion Methods

Sales Promotion Method	Sales Associated with Sales Promotion
1	2,000
2	3,500
3	2,510

The results may lead us to the conclusion that the second sales promotion method was the most effective in developing sales. This may be adopted nationally to promote the product. But one cannot say that the same method of sales promotion will be effective in each and every city under study.

14.3.1 Basic Analysis of "Quantitative" Information

(for information other than commentary, e.g., ratings, rankings, yes's, no's, etc.)

- Make copies of your data and store the master copy away. Use the copy for making edits, cutting and pasting, etc.
- Tabulate the information, i.e., add up the number of ratings, rankings, yes's, no's for each question.
- For ratings and rankings, consider computing a mean, or average, for each question. For example, "For question #1, the average ranking was 2.4". This is more meaningful than indicating, e.g., how many respondents ranked 1, 2, or 3.
- Consider conveying the range of answers, e.g., 20 people ranked "1", 30 ranked "2", and 20 people ranked "3".

14.3.2 Basic Analysis of "Qualitative" Information

(respondents' verbal answers in interviews, focus groups, or written commentary on questionnaires):

Notes

- Read through all the data.
- Organize comments into similar categories, e.g., concerns, suggestions, strengths, weaknesses, similar experiences, program inputs, recommendations, outputs, outcome indicators, etc.
- Label the categories or themes, e.g., concerns, suggestions, etc.
- Attempt to identify patterns, or associations and causal relationships in the themes, e.g., all people who attended programs in the evening had similar concerns, most people came from the same geographic area, most people were in the same salary range, what processes or events respondents experience during the program, etc.
- Keep all commentary for several years after completion in case needed for future reference.

14.3.3 Interpreting Information

- Attempt to put the information in perspective, e.g., compare results to what you expected, promised results; management or program staff; any common standards for your products or services; original goals (especially if you're conducting a program evaluation); indications or measures of accomplishing outcomes or results (especially if you're conducting an outcomes or performance evaluation); description of the program's experiences, strengths, weaknesses, etc. (especially if you're conducting a process evaluation).
- Consider recommendations to help employees improve the program, product or service; conclusions about program operations or meeting goals, etc.
- Record conclusions and recommendations in a report, and associate interpretations to justify your conclusions or recommendations.

14.3.4 Precautions

1. Keep the main objective of research in mind.
2. Analysis of data should start from simpler and more fundamental aspects.
3. It should not be confusing.
4. The sample size should be adequate.
5. Take care before generalising of the sample studied.
6. Give due attention to significant questions.

Caution: In report writing, do not miss the significance of some answers, because they are found from very few respondents, such as "don't know" or "can't say".

Self Assessment

Fill in the blanks:

7. ……means bringing out the meaning of data.
8. Successful interpretation depends on how well the data is……...
9. In the ……method, one starts from observed data and then generalisation is done

Notes

14.4 Types of Report

There are two types of reports (1) Oral report (2) Written report.

14.4.1 Oral Report

This type of reporting is required, when the researchers are asked to make an oral presentation. Making an oral presentation is somewhat difficult compared to the written report. This is because the reporter has to interact directly with the audience. Any faltering during an oral presentation can leave a negative impression on the audience. This may also lower the self-confidence of the presenter. In an oral presentation, communication plays a big role. A lot of planning and thinking is required to decide 'What to say', 'How to say', 'How much to say'. Also, the presenter may have to face a barrage of questions from the audience. A lot of preparation is required; the broad classification of an oral presentation is as follows.

Nature of an Oral Presentation

Opening: A brief statement can be made on the nature of discussion that will follow. The opening statement should explain the nature of the project, how it came about and what was attempted.

Finding/Conclusion: Each conclusion may be stated backed up by findings.

Recommendation: Each recommendation must have the support of conclusion. At the end of the presentation, question-answer session should follow from the audience.

Method of presentation: Visuals, if need to be exhibited, can be made use of. The use of tabular form for statistical information would help the audience.

(a) What type of presentation is a root question? Is it read from a manuscript or memorized or delivered ex-tempo. Memorization is not recommended, since there could be a slip during presentation. Secondly, it produces speaker-centric approach. Even reading from the manuscript is not recommended, because it becomes monotonous, dull and lifeless. The best way to deliver in ex-tempo, is to make main points notes, so that the same can be expanded. Logical sequences should be followed.

Notes Points to remember in oral presentation:

1. Language used must be simple and understandable.
2. Time Management should be adhered.
3. Use of charts, graph, etc., will enhance understanding by the audience.
4. Vital data such as figures may be printed and circulated to the audience so that their ability to comprehend increases, since they can refer to it when the presentation is going on.
5. The presenter should know his target audience well in advance to prepare tailor-made presentation.
6. The presenter should know the purpose of report such as "Is it for making a decision", "Is it for the sake of information", etc.

Notes

14.4.2 Written Report

Following are the Various Types of Written Reports:

(A) ***Reports can be classified based on the time-interval such as:***

(1) Daily

(2) Weekly

(3) Monthly

(4) Quarterly

(5) Yearly

(B) ***Type of reports:***

(1) Short report

(2) Long report

(3) Formal report

(4) Informal report

(5) Government report

1. ***Short report:*** Short reports are produced when the problem is very well defined and if the scope is limited. For example, Monthly sales report. It will run into about five pages. It consists of report about the progress made with respect to a particular product in a clearly specified geographical locations.

2. ***Long report:*** This could be both a technical report as well as non-technical report. This will present the outcome of the research in detail.

   (a) ***Technical report:*** This will include the sources of data, research procedure, sample design, tools used for gathering data, data analysis methods used, appendix, conclusion and detailed recommendations with respect to specific findings. If any journal, paper or periodical is referred, such references must be given for the benefit of reader.

   (b) ***Non-technical report:*** This report is meant for those who are not technically qualified. E.g. Chief of the finance department. He may be interested in financial implications only, such as margins, volumes, etc. He may not be interested in the methodology.

3. ***Formal report:***

Example: The report prepared by the marketing manager to be submitted to the Vice-President (marketing) on quarterly performance, reports on test marketing.

4. ***Informal report:*** The report prepared by the supervisor by way of filling the shift log book, to be used by his colleagues.

5. ***Government report:*** These may be prepared by state governments or the central government on a given issue.

Example: Programme announced for rural employment strategy as a part of five-year plan.

Notes

Did u know? Report on children's education is a kind of government and social welfare report.

14.4.3 Distinguish between Oral and Written Report

Oral report	Written report
No rigid standard format.	Standard format can be adopted.
Remembering all that is said is difficult if not impossible. This is because the presenter cannot be interrupted frequently for clarification.	This can be read a number of times and clarification can be sought whenever the reader chooses.
Tone, voice modulation, comprehensibility and several other communication factors play an important role.	Free from presentation problems.
Correcting mistakes if any, is difficult.	Mistakes, if any, can be pinpointed and corrected.
The audience has no control over the speed of presentation.	Not applicable.
The audience does not have the choice of picking and choosing from the presentation.	The reader can pick and choose what he thinks is relevant to him. For instance, the need for information is different for technical and non-technical persons.

Self Assessment

Fill in the blanks:

10. In an oral presentation, …….plays a big role.
11. ……..report presents the outcome of the research in detail.
12. The …….statement should explain the nature of the project, how it came about and what was attempted.

14.5 Preparation of Research Report

Having decided on the type of report, the next step is report preparation. The following is the format of a research report:

1. Title Page
2. Page Contents
3. Executive Summary
4. Body
5. Conclusions and Recommendations
6. Bibliography
7. Appendix

1. ***Title Page:*** Title Page should indicate the topic on which the report is prepared. It should include the name of the person or agency who has prepared the report.

Notes

2. ***Table of Contents:*** The table of contents will help the reader to know "what the report contains". The table of contents should indicate the various parts or sections of the report. It should also indicate the chapter headings along with the page number.

Chapter no.	Title of the chapter	Page no.
	Declaration	
	Certificates	
	Acknowledgement	
	Executive summary	
1	Introduction to the project	
2	Research design and methodology	
3	Theoretical perspective of the study	
4	Company and industry profile	
5	Data analysis and interpretation	
6	Summary of findings, suggestions and conclusions	
	Bibliography	
	Appendix	

3. ***Executive Summary:*** If your report is long and drawn out, the person to whom you have prepared the report may not have the time to read it in detail. Apart from this, an executive summary will help in highlighting major points. It is a condensed version of the whole report. It should be written in one or two pages. Since top executives read only the executive summary, it should be accurate and well-written. An executive summary should help in decision-making.

   An executive summary should have,

   (a) Objectives

   (b) Brief methodology

   (c) Important findings

   (d) Key results

   (e) Conclusion

4. ***The Body:*** This section includes:

   (a) Introduction

   (b) Methodology

   (c) Limitations

   (d) Analysis and interpretations

   Introduction: The introduction must explain clearly the decision problem and research objective. The background information should be provided on the product and services provided by the organisation which is under study.

   Methodology: How you have collected the data is the key in this section. For example, Was primary data collected or secondary data used? Was a questionnaire used? What was the sample size and sampling plan and method of analysis? Was the design exploratory or conclusive?

   Limitations: Every report will have some shortcoming. The limitations may be of time, geographical area, the methodology adopted, correctness of the responses, etc.

Notes

Analysis and interpretations: collected data will be tabulated. Statistical tools if any will be applied to make analysis and to take decisions.

5. ***Conclusion and Recommendation:***

   (a) What was the conclusion drawn from the study?

   (b) Based on the study, what recommendation do you make?

6. ***Bibliography:*** If portions of your report are based on secondary data, use a bibliography section to list the publications or sources that you have consulted. The bibliography should include, title of the book, name of the journal in case of article, volume number, page number, edition, etc.

7. ***Appendix:*** The purpose of an appendix is to provide a place for material which is not absolutely essential to the body of the report. The appendix will contain copies of data collection forms called questionnaires, details of the annual report of the company, details of graphs/charts, photographs, CDs, interviewers' instructions. Following are the items to be placed in this section.

   (a) Data collection forms

   (b) Project related paper cuttings

   (c) Pictures and diagrams related to project

   (d) Any other relevant things.

Caution The date of the submission of the report is to be included in the title page of the report.

14.5.1 How to Write a Bibliography?

Bibliography, the last section of the report comes after appendices. Appendices contains questionnaires and other relevant material of the study. The bibliography contains the source of every reference used and any other relevant work that has been consulted. It imparts an authenticity regarding the source of data to the reader.

Bibliography are of different types viz., bibliography of works cited; this contains only the items referred in the text. A selected bibliography lists the items which the author thinks are of primary interest to the reader. An annotated bibliography gives brief description of each item. The method of representing bibliography is explained below.

Books

Name of the author, title of the book (underlined), publisher's detail, year of publishing, page number.

Single Volume Works. Dube, S. C. "India's Changing Villages", Routledge and Kegan Paul Ltd., 1958, p. 76.

Chapter in an Edited Book

Warwick, Donald P., "Comparative Research Methods" in Balmer, Martin and Donald Warwick (eds), 1983, pp. 315-30.

Notes

Periodicals Journal

Dawan Radile (2005), "They Survived Business World" (India), May 98, pp. 29-36.

Newspaper, Articles

Kumar Naresh, "Exploring Divestment", The Economic Times (Bangalore), August 7, 1999, p. 14.

Website

www.infocom.in.com

For citing Seminar Paper

Krishna Murthy, P., "Towards Excellence in Management" (Paper presented at a Seminar in XYZ College Bangalore, July 2000).

Task List the various abbreviations frequently used in footnotes with their meanings.

Self Assessment

Fill in the blanks:

13. The ……..should indicate the various parts or sections of the report.
14. ……..Page should indicate the topic on which the report is prepared.
15. A selected bibliography lists the items which the author thinks are of …….interest to the reader.

14.6 Style, Layout and Precautions of the Report writing

14.6.1 Style of Report Writing

Remember that the reader:

- Has short of time,
- Has many other urgent matters demanding his or her interest and attention,
- Is probably not knowledgeable concerning 'research jargon'.

Therefore, the rules are:

- Simplify. Keep to the essentials.
- Justify. Make no statement that is not based on facts and data.
- Quantify when you have the data to do so. Avoid large, small, instead, say 50%, one in three.
- Be precise and specific in your phrasing of findings.
- Inform, not impress. Avoid exaggeration.
- Use short sentences.

Notes

- Use adverbs and adjectives sparingly.
- Avoid the passive voice, if possible, as it creates vagueness (e.g., 'patients were interviewed' leaves uncertainty as to who interviewed them) and repeated use makes dull reading.
- Aim to be logical and systematic in your presentation.

Caution In report writing, be consistent in the use of tenses (past or present tense).

14.6.2 Layout of the Report

A good physical layout is important, as it will help your report:

i. Make a good initial impression,

ii. Encourage the readers, and

iii. Give them an idea of how the material has been organised so the reader can make a quick determination of what he will read first.

Particular attention should be paid to make sure there is:

i. An attractive layout for the title page and a clear table of contents.

ii. Consistency in margins and spacing.

iii. Consistency in headings and subheadings, for example, font size 16 or 18 bold, for headings of chapters; size 14 bold for headings of major sections; size 12 bold, for headings of sub-sections, etc.

iv. Good quality printing and photocopying. Correct drafts carefully with spell check as well as critical reading for clarity by other team-members, your facilitator and, if possible, outsiders.

v. Numbering of figures and tables, provision of clear titles for tables, and clear headings for columns and rows, etc.

vi. Accuracy and consistency in quotations and references.

14.6.3 Precautions in Report Writing

Endless description without interpretation is another pitfall. Tables need conclusions, not detailed presentation of all numbers or percentages in the cells which readers can see for themselves.

Notes The unit discussion, in particular, needs comparison of data, highlighting of unexpected results, your own or others' opinions on problems discovered, weighing of pro's and con's of possible solutions. Yet, too often the discussion is merely a dry summary of findings.

Neglect of qualitative data is also quite common. Still, quotes of informants as illustration of your findings and conclusions make your report lively. They also have scientific value in allowing the reader to draw his/her own conclusions from the data you present. (Assuming you are not biased in your presentation!)

Sometimes qualitative data (e.g., open opinion questions) are just coded and counted like quantitative data, without interpretation, whereas they may be providing interesting illustrations

Notes

of reasons for the behavior of informants or of their attitudes. This is serious maltreatment of data that needs correction.

The following must be avoided while preparing a report:

- The inclusion of careless, inaccurate, or conflicting data.
- The inclusion of outdated or irrelevant data.
- Facts and opinions that are not separated.
- Unsupported conclusions and recommendations.
- Careless presentation and proofreading.
- Too much emphasis on appearance and not enough on content.

Self Assessment

Fill in the blanks:

16. In a report there must be ……..in margins and spacing.
17. Aim must be logical and ……in the report presentation

14.7 Summary

- A report is a very formal document that is written for a variety of purposes, generally in the sciences, social sciences, engineering and business disciplines.
- The most important aspect to be kept in mind while developing research report, is the communication with the audience.
- Report should be able to draw the interest of the readers. Therefore, report should be reader centric.
- Other aspect to be considered while writing report are accuracy and clarity.
- The point to be remembered while doing oral presentation is language used, Time management, use of graph, purpose of the report, etc. Visuals used must be understandable to the audience.
- The presenter must make sure that presentation is completed within the time allotted. Sometime should be set apart for questions and answers.
- Written report may be classified based on whether the report is a short report or a long report. It can also be classified based on technical report or non technical report.
- Written report should contain title page, contents, executive summary. Body, conclusions and appendix. The last part is bibliography.
- The style of the report should be simple and to the essentials.
- There should not be endless description in report writing and qualitative data is not to be excluded.

14.8 Keywords

Appendix: The part of the report whose purpose is to provide a place for material which is not absolutely essential to the body of the report.

Notes

Bibliography: The section to list the publications or sources that you have consulted in preparation of report

Executive Summary: It is a condensed version of the whole report.

Informal Report: The report prepared by the supervisor by way of filling the shift log book, to be used by his colleagues

Short Report: Short reports are the reports that are produced when the problem is very well defined and if the scope is limited.

14.9 Review Questions

1. What is a research report?
2. What are the characteristics of report?
3. What is the criterion for an oral report? Explain.
4. What is meant by "consider the audience" when writing a research report.
5. On what criteria, oral report is evaluated? Suggest a suitable format.
6. Why are visual aids used in oral presentation?
7. What are the various criteria used for classification of written report?
8. What are the essential content of the following parts of research report?
   (a) Table of contents
   (b) Title page
   (c) Executive summary
   (d) Introduction
   (e) Conclusion
   (f) Appendix
9. Oral presentation requires the researcher to be good public speaker explain.
10. Explain the style and layout of report.

Answers: Self Assessment

1. need
2. reality
3. decision maker
4. Currency
5. Research report
6. final
7. Interpretation
8. analysed
9. induction
10. communication
11. Long
12. opening
13. table of contents
14. Title
15. primary
16. Consistency
17. systematic

Notes

14.10 Further Readings

Books

Abrams, M.A., *Social Surveys and Social Action*, London: Heinemann, 1951.

Arthur, Maurice, *Philosophy of Scientific Investigation*, Baltimore: John Hopkins University Press, 1943.

Bernal, J.D., *The Social Function of Science*, London: George Routledge and Sons, 1939.

Chase, Stuart, *The Proper Study of Mankind: An inquiry into the Science of Human Relations*, New York, Harper and Row Publishers, 1958.

S. N. Murthy and U. Bhojanna, *Business Research Methods*, Excel Books.

Statistical Tables

Notes

I. Logarithms

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	5 4	9 8	13 12	17 16	21 20	26 24	30 28	34 32	38 36
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4 4	8 7	12 11	16 15	20 18	23 22	27 26	31 29	35 33
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1186	3 3	7 7	11 10	14 14	18 17	21 20	25 24	28 27	32 31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3 3	6 7	10 10	13 13	16 16	19 19	23 22	26 25	29 29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3 3	6 6	9 9	12 12	15 14	19 17	22 20	25 23	28 26
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3 3	6 6	9 8	11 11	14 14	17 17	20 19	23 22	26 25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3 3	6 5	8 8	11 10	14 13	16 16	19 18	22 21	24 23
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	3 3	5 5	8 8	10 10	13 12	15 15	18 17	20 20	23 22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2 2	5 4	7 7	9 9	12 12	14 14	17 17	19 18	21 21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2 2	4 4	7 6	9 8	11 11	13 13	16 15	18 17	20 19
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5187	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	12	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	6	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8

Notes

I. Logarithms

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	2	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	2	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	2	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	9116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8570	8476	8482	8488	8494	8500	8505	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	6
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	6
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	6
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	6
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8999	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9560	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	7927	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4

II. Antilogarithms

Notes

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
.02	1047	1050	1052	1054	1057	1060	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
.03	1072	1074	1076	1079	1081	1084	1086	1089	1001	1094	0	1	1	1	1	2	2	2	2
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	2	2	2	2
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	2	2	2	2
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	2	2	2	2
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
.09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	0	1	1	1	2	2	2	3	3
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
.18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
.20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
.22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	2	2	2	3	3	4
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	4	4
.28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2329	1	1	2	2	3	3	4	4	5
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	4	5
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	4	4	5	5
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	4	4	5	5
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	4	4	5	6
.43	2692	2693	2794	2710	2716	2723	2729	2735	2742	2448	1	1	2	3	3	4	4	5	6
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
.45	2818	2825	2831	2838	2844	2852	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
.48	2920	2927	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	4	4	5	6	6
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1	1	3	3	4	4	5	6	6

Notes

II. Antilogarithms

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
.52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
.54	3567	3575	3589	3491	3499	3508	3516	3524	3532	3540	1	2	2	3	4	5	6	6	7
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
.56	3631	3639	3648	3656	3664	3573	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	4	4	5	6	7	8
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
.70	5022	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
.71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
.72	5248	5250	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
.75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6162	1	3	4	6	7	8	10	11	13
.79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	10	12	14
.82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
.85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	6	9	11	13	15	17
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9211	2	4	6	8	11	13	15	17	19
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20

III. Binomial Coefficients

n	$({}^nC_0)$	$({}^nC_1)$	$({}^nC_2)$	$({}^nC_3)$	$({}^nC_4)$	$({}^nC_5)$	$({}^nC_6)$	$({}^nC_7)$	$({}^nC_8)$	$({}^nC_9)$	$({}^nC_{10})$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3005	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	1184759

IV. Values of e^{-m}

m	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	0.9048	.8958	.8860	.8781	.8694	.8607	.8521	.8437	.8353	.8270
0.2	0.8178	.8106	.8025	.7945	.7866	.7788	.7711	.7634	7558	.7483
0.3	0.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	0.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	0.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	0.5488	.5434	.5379	.5326	.5278	.5220	.5160	.5117	.5066	.5016
0.7	0.4966	.4916	.4868	.4810	.4771	.4724	.4670	.4630	.4584	.4538
0.8	0.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	0.4066	.4025	.3985	.3946	.3906	.3867	..3829	.3791	.3753	.3716

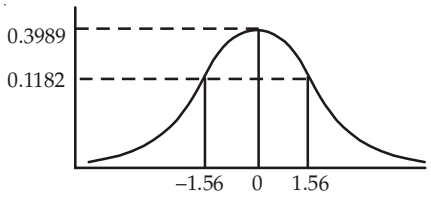
(m = 1, 2, 3,....10)

m	1	2	3	4	5	6	7	8	9	10
e^{-m}	.36788	.13534	.04979	.01832	.006738	.002479	.000912	.000335	.000123	.000045

Note: To obtain the value of $e^{-1.75}$, we write $e^{-1.75} = e^{-1} \times e^{-0.75}$ = 0.s36788 × 0.4724 = 0.17379

Notes

V. Ordinates of Normal Curve



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2802	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	0.175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003.	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

VI. Areas under the Normal Curve

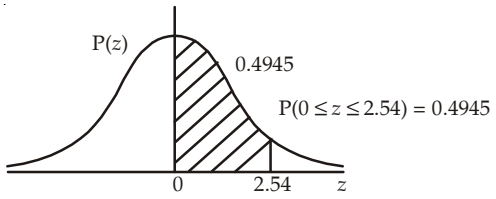
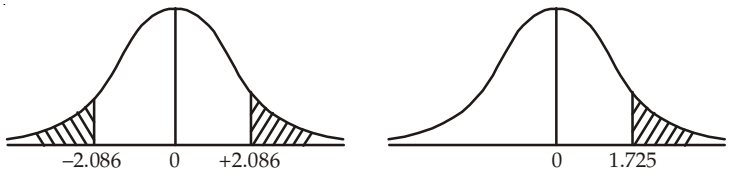


Table of Area

↓Z→	0	1	2	3	4	5	6	7	8	9
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2223
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
.7	.2580	.2612	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4849	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.49.64
2.7	.4965	.4965	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4390	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	4997	.4997	.4997	.4997	.4997	.4947	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4898	.4898
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Notes

VII. Critical Values of t

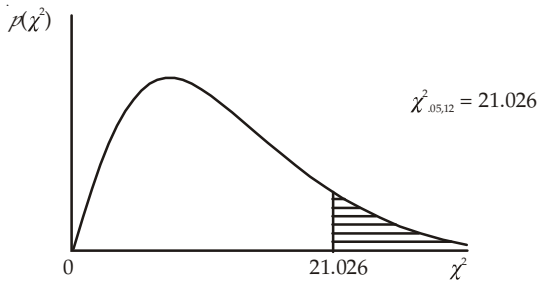


two tailed test $t_{.05,20} = \pm 2.086$; one tailed test $t_{.05,20} = 1.725$

d.f.	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2,831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
inf.	1.282	1.645	1.960	2.326	2.576

VIII. Critical Values of χ^2

Notes



Degrees of freedom	Area in right tail							
	0.99	0.975	0.95	0.90	0.10	0.05	0.25	0.01
1	0.00016	0.00098	0.00398	0.0158	2.706	3.841	5.024	6.635
2	0.0201	0.0506	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.02	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.658	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.328	37.652	40.647	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46 979	50.892

In a Chi-square (χ^2) Distribution with 10 degrees of freedom, to find the value under 0.05 of area on the right, look for cross-section of values of 10 degrees of freedom row and 0.05 column to find the value 18.307.

Notes

IX. Critical Values of F

Degrees of freedom for denominator	Degrees of freedom for numerator																			
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.24	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Notes

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	5 4	9 8	13 12	17 16	21 20	26 24	30 28	3 3
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4 4	8 7	12 11	16 15	20 18	23 22	27 26	3 2
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1186	3 3	7 7	11 10	14 14	18 17	21 20	25 24	2 2
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3 3	6 7	10 10	13 13	16 16	19 19	23 22	2 2
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3 3	6 6	9 9	12 12	15 14	19 17	22 20	2 2
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3 3	6 6	9 8	11 11	14 14	17 17	20 19	2 2
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3 3	6 5	8 8	11 10	14 13	16 16	19 18	2 2
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	3 3	5 5	8 8	10 10	13 12	15 15	18 17	2 2
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2 2	5 4	7 7	9 9	12 12	14 14	17 17	1 1
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2 2	4 4	7 6	9 8	11 11	13 13	16 15	1 1
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	1
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	1
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	1
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	1
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	1
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	1
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	1
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	1
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	1
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	1
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	1
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	1
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	1
33	5187	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	1
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	1
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	1
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	1
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	6	6	
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	

Notes

X. Quality Control Charts

$\bar{X}$-chart				σ –Chart					R-chart				
Sample size	Factors for Control-limits			Factors for central line	Factors for control limits				Factors for central line		Factors for control limits		
n	A	A_1	A_2	c_2	B_1	B_2	B_3	B_4	d_2	D_1	D_2	D_3	D_4
2	2·121	3·760	1·880	0·5642	2	1·843	0	3·267	1·128	0	3 686	0	3·267
3	1·732	2·394	1·023	0·7236	0	1·858	0	2·568	1·693	0	4 358	0	2·575
4	1·500	1·880	0·729	0·7979	0	1·808	0	2·266	2·059	0	4 698	0	2·282
5	1·342	1·596	0·577	0·8407	0	1·756	0	2·089	2·326	0	4·981	0	2 115
6	1·225	1·410	0·483	0·8686	0·026	1·711	0·030	1·970	2·534	0	5·076	0	2·004
7	1·134	1·277	0·419	0·8882	0·105	1·672	0·118	1·882	2 704	0·205	5 203	0·078	1 924
8	1·061	1·175	0·373	0·9027	0·167	1·638	0 185	1·815	2·847	0 387	5·307	0·136	1·864
9	1·000	1·094	0·337	0·9139	0 219	1·609	0 236	1·761	2·970	0 546	5 394	0 184	1·816
10	0·949	1·028	0·308	0·9227	0·262	1 584	0·284	1·716	3·078	0·687	5 469	0 223	1 777
11	0·905	0·973	0·285	0·9300	0 299	1·561	0·321	1·679	3·173	0·812	5·534	0 256	1 744
12	0·866	0·925	0·266	0·9359	0 331	1·541	0 354	1·648	3 258	0 924	5 592	0 284	1 71(
13	0·832	0·884	0·249	0·9410	0 359	1 523	0 382	1·618	3 336	1 026	5 646	0·308	1 622
14	0·802	0·848	0·235	0·9453	0 384	1·507	0·406	1·594	3·407	1·121	5 693	0 329	1·671
15	0·775	0·816	0·223	0 9490	0 406	1·492	0·428	1·572	3·472	1·207	5·737	0 348	1·652
16	0·750	0·788	0·212	0·9523	0 427	1·478	0·448	1·552	3·532	1·285	5·779	0·364	1·636
17	0·728	0·762	0·203	0 9551	0·445	1 465	0·466	1·534	3·588	1·359	5·817	0·379	1·621
18	0·707	0·738	0·194	0 9576	0·461	1 454	0·482	1·518	3·640	1·426	5 854	0·392	1·608
19	0·688	0· 717	0·184	0·9599	0·477	1 443	0·497	1·503	3·689	1·490	5 888	0·404	1·596
20	0·671	0·697	0·110	0 9619	0·491	1·433	0·510	1 490	3 735	1·544	5 922	0·418	1·580
21	0·655	0·679	0·173	0·9638	0·504	1·424	0·523	1·477	3 788	1·606	5·950	0 425	1·575
22	0·640	0·662	0·167	0·9655	0·516	1·415	0·534	1·466	3·819	1·659	5·979	9·434	1·566
23	0·626	0·647	0·162	0·9670	0·527	1·407	0 545	1·455	3·858	1·710	5·006	0·443	1·557
24	0·612	0·632	0·157	0 9684	0 538	1·339	0·555	1·445	3·859	1·759	6·031	0·452	1·548
25	0·600	0·619	0·153	0·9696	0 548	1·392	0·565	1·435	3·931	1·804	6·058	0·059	1·541

LOVELY PROFESSIONAL UNIVERSITY
Jalandhar-Delhi G.T. Road (NH-1)
Phagwara, Punjab (India)-144411
For Enquiry: +91-1824-300360
Fax.: +91-1824-506111
Email: odl@lpu.co.in