Research Methodology

DEMGN832

Edited by:
Dr. Lokesh Jasrai

LOVELY PROFESSIONAL UNIVERSITY

Research Methodology

Edited By
Dr. Lokesh Jasrai

Title: RESEARCH METHODOLOGY

Author's Name: Dr. Atif Ghayas

Published By : Lovely Professional University

Publisher Address: Lovely Professional University, Jalandhar Delhi GT road, Phagwara - 144411

Printer Detail: Lovely Professional University

Edition Detail: (I)

ISBN: 978-93-94068-36-0

CONTENTS

Unit 01:Background of Research

Contents

Objectives

After studying this unit, you will be able to:

- understand the process of writing a research proposal.
- explain the significance of the research proposal.
- compare and contrastthe major paradigms of evaluation.
- outline the contribution of research towards theory.
- assess the role of ethics in research.

Introduction

What is research?

A careful consideration of study concerning a specific concern or problem using scientific methods. Research isalso defined as the creation of new knowledge and/or the use of existing knowledge in a new and creative way to generate new concepts, methodologies, and understandings. This could include creation and analysis ofprevious research to the extent that it leads to new and creative outcomes. Research also includes a careful consideration of study regarding a particular concern or problem using scientific methods. According to Earl Robert Babbie, "*Research is a systematic inquiry todescribe, explain, predict, and control the observed phenomenon. Research involves inductive and deductivemethods*."Inductive research methods are used to analyze an observed event. Deductive methods are used to verify theobserved event. Inductive approaches are associated with qualitative research and deductive methods are morecommonly associated with quantitative research.

Research is conducted with a purpose to understand:

What do organizations or businesses want to find out?

What are the processes that need to be followed to chase the idea?

What are the arguments that need to be built around a concept?

What is the evidence that will be required for people to believe in the idea or concept?

Notes

1.1 Characteristics of Research

- A systematic approach must be followed for accurate data. Rules and procedures are an integral part of theprocess that sets the objective. Researchers need to practice ethics and a code of conduct while makingobservations or drawing conclusions.
- Research is based on logical reasoning and involves both inductive and deductive methods.
- The data or knowledge that is derived is in real-time from actual observations in natural settings.
- There is an in-depth analysis of all data collected so that there are no anomalies associated with it.
- Research creates a path for generating new questions. Existing data helps create more research opportunities.
- Research is analytical. It makes use of all the available data so that there is no ambiguity ininference.
- Accuracy is one of the most important aspects of research. The information that is obtained should beaccurate and true to its nature. For example, laboratories provide a controlled environment to collectdata. Accuracy is measured in the instruments used, the calibrations of instruments or tools, and the result of the experiment.

1.2 Research Proposal

A research proposal is a formal document in which a researcher defines a topic and explains his or her plansto research on the topic. Such a proposal is used not only to create a coherent plan but also to convince ateacher or reviewer that has developed a relevant, focused, and interesting topic and that the plans toresearch that topic will work.There are several basic steps a researcher will take to develop a research plan. Each of these stepsconstitutes information that is included in the research proposal:

1. Developing and defining a topic.
2. Exploring your purpose and audience for your research.
3. Conducting preliminary research.
4. Formulating a research question (and additional questions).
5. Creating a research plan.
6. Prepare a research proposal.

Developing and Defining a Topic

Research topic requires a commitment from the researcher and hence needs to be selected with due attention.The choice will help determinewhether the researcher enjoys the lengthy process of research and writing and whether the topic fulfills the requirements. If someone choosesa topic hastily, he/she may later find it difficult to work with the topic. By taking the time andchoosing carefully, a researcher can ensure that this assignment is not only challenging but also rewarding. A researcherunderstands the importance of choosing a topic that not only fulfills the assignment requirements but alsofits his interests and priorities. Choosing a topic that interests a researcher is crucial. An instructor may provide alist of suggested topics or ask that a researcher can develop a topic on his/her own. In either case, a researcher should try to identify topics that belong to interest areas. This interest is sometimes called exigence – the personal concerns and interests thatdrive a researcher to investigate a specific topic.The outcome ofthe research should matter here and now to others. This is often called kairos – the concerns and interests beyondthat make the topic relevant now.

After identifying potential topic ideas, a researcher will need to evaluate the ideas and choose one topic to pursue.

Identifying Potential Topics

Sometimes, an instructor may provide a list of suggested topics. If so, a researcher may benefit from identifying severalpossibilities before committing to one idea. Other times, an instructor lets students decide where to beginwhen picking a topic. It is importantto know how to narrow down the ideas into a concise, manageablethesis. Discussing the ideas with the instructor will help ensure that a researcher chooses a manageable topic that fitsthe requirements of the assignment.

Below are a couple of common approaches to developing and narrowing a topic.

Starting with Brainstorming

The most common method of developing a topic is through brainstorming. In this way, the researcher can set aside time to write down or organize the thoughts or topics that are of interest. For each topic, he/she can briefly describe the need for the research (why the researcher is interested in the subject) and Kairos (why this is an appropriate and timely topic for others) and write down their current understanding of each topic. After this, a potential researcher may use his or her writing to strike out the ideas that seems less interesting, but sometimes a researcher will end up choosing the ones that he or she and others find most interesting. While this may be a good way to find a topic, he or she is interested in it, a researcher may find it difficult to narrow down that topic before starting research. If, after brainstorming, a researcher ends up with broad categoriessuch as Declining of sales, Capital Punishment, Cell Phone use, Child Abuse, Eating Disorders, etc., he/she need tocontinue to do more brainstorming to figure out specific aspects of one of these topics that he/she wants toinvestigate.

For instance, if a student tries to research the general topic of Capital Punishment, he/she will find thatthere is far too much ground to cover. But if the student starts breaking down that general topic, looking at it fromdifferent vantage points, they may find a narrower topic that is much more manageable. For instance, instead ofCapital Punishment in general, a researcher may decide to examine legal decisions about capital punishment as"cruel and unusual punishment," or racial bias in capital punishment sentences, or differences in the frequencyof executions in the states where capital punishment is legal or even the exploration of whether capitalpunishment should be legal in one's state.

Starting with Research

One effective way to choose a topic is to start with research. Instead of selecting a broader topic and trying to shorten it, the researcher should begin with a specific issue, issue or event reported in a news story or study, and then expand on that issue or set out to develop interesting and relevant research topic. This method works well for several reasons. First, when a person did research and wrote about that research, they participated in a discussion with other researchers and authors. And as with any discussion, it is helpful to find out what is being said in the previously published sources before getting an idea of your topic and decide how you want to contribute to the academic discourse on that issue. Second, it helps to make sure that the topic is small enough, and manageable, to begin with. When a broad subject is chosen and a researcher has a task to shorten it, he/she often fail to cut it short. This approach avoids the problem by looking at something specific and extending the research topic related to the content you are interested to research.

Explore Topics

In a few sentences, describe broader topics or issues the article touches on. Beyond the specific incident orevent described in the article, what larger social problems or debates does the article relate to? (EX:reading an article on a specific topic like an increase in fuel prices, topics might include international petro-prices, vehicles, government policies, taxes, etc.)

Explore Exigence

In a few sentences, explain why you are personally interested in or curious about the incidentreported in the article. If possible, connect it to your own experience. Based on this, what topics do youthink you'd like to research should be attempted to answer.

Explore Kairos

In a few sentences, identify the groups of people this incident or problem matters to (beyondyourself) and why it matters to them now, thinking not only of those involved in the incident itself but other peopleor entities or institutions in society that might have a concern

Notes

regarding this incident or incidents like it. Based onwhat matters most about this incident, what topics related to it might be worthy of research?

Explore Controversies

In a few sentences, explain what differences of opinion or debates may exist about thisincident or event and you think those differences of opinion might exist? Based on this, which of thesecontroversies might be worthy of research?

Narrowing Your Topic

Once a researcher has developed potential topics, he/she will need to choose one as the focus of the research. The researcher will also need to reduce the topic in terms of relevance and accuracy of the content. Especially when brainstorming is used, most people find the topics they have listed during brainstorming are broad given the size of the assignment. Working on a broader topic, such as online education programs or popular diets, can be confusing and needs more elaboration. Each topic has so many features that it would be impossible to combine them all. A good research proposal provides in-depth information and focused analysis. If the topic is too broad, the researcher will find it difficult to do more than just skip the top and research it. Reducing focus is important in keeping the topic manageable. To reduce stress, a researcher can check the published research on the topic, can do some initial research, and discuss the topic with others.

Exploring the Purpose and Audience for Research

Any good research proposal will also discuss the purposes and potential audiences for the research one isconducting. Often, this begins with the researcher's exigence and the topic's Kairos, and from there extends towhat the research hopes to accomplish and who he hopes to inform and persuade with this research. This isgenerally done through free writing.

Conducting Preliminary Research

To prepare for a research proposal, a researcher also needs to conduct some preliminary research. Partly, this isto ensure that there are some viable sources and possibilities available for the topic idea you've generated. Butas well, you will need some information and insights about your topic to define that topic in the researchproposal and to develop a valid research question. This preliminary research can help to understand importanthistory, concepts, and terminology about the proposed topic. It can also help a researcher o find out what people are saying aboutthe topic and the opinions that exist about it. This research can be conducted using resources available in a college library or bysearching online repositories.

Formulating a Research Question

The research question is an important aspect that the researcher should ask himself to focus on the research and improve the structure of the research. To conduct research, the researcher should seek out resources to help answer the research question. In addition, a researcher can write a study design to answer that question. In dealing with a research question, the researcher sets a research goal.The main research question should besubstantial enough to form the guiding principle of the research—but focused enough to also actually guide theresearch. A strong research question requires not only to find information but also to put together different pieces of information, interpret and analyze them, and figure out what a researcher thinks. As you consider potentialresearch questions, ask yourself whether they would be too hard or too easy to answer.To determine your research question, review the freewriting about your topic. Skim through your preliminaryresearch and list the questions you have. You can include simple, factual questions but as you continue youshould push yourself toward more complex questions that would require research, analysis, and interpretation.From there, determine your main research question—the primary focus of your research and the composition youwill write based on it.

A Plan for Research

To work with your topic successfully, you will need to determine what exactly you want to learn about it—andlater, what you want to say about it.

Take some time, then, to make a plan for your research

Plan for conducting the secondary research. What kinds of information will you need to answer your researchquestion and supporting questions? What kinds of secondary sources will you find them in? Which of thesesources should be found using online or physical library resources and

which are best found on the open web?What search words will you use to search for your sources, whether you do this using library resources or goonline?

Plan for conducting primary research

Will your research include primary research? If so, will you conductobservations, interviews, surveys, or some other kind of primary research? Who will you interview or survey orwhat will you observe? What materials will you need to gather to conduct your primary research?

Plan for time

How much time will you give yourself to complete each part of your research agenda? Which days(or time of day) will you complete each leg of your research? How much research is realistic, given yourdeadline?

Creating a Research Proposal

A research proposal can be long (several pages) or short (1-2 pages) depending on the goals and specificity ofyour research proposal assignment. A researcher should follow the specified guidelines to determine thelength and depth of the research proposal assigned. In any case, the purpose of your research proposal is todefine your topic and formalize your research plan. This proposal will help to convince your instructor of yourresearch ideas and will help your instructor give you feedback on your topic and research plans. In your researchproposal, you will define your topic, discuss the importance of this topic both to you and others, present yourmain research question and supporting questions, and offer a research plan.

Give your opinion on the importance of the research proposal.

1.3 Creating a good research proposal

Creating a good research proposal Most novice researchers ask what an outline of a proposal might look like. If exemplars of good proposals are available, it will pay you to study these before you set out developing yours.

A researcher would do well to keep the following principles in mind when developing a proposal:

1. A good proposal explains clearly three elements – what research is intended, why it is being researched, and how the researcher proposes to carry out the research.

2. A good proposal is straightforward. The first words are of vital importance. They need to get to the point directly without 'beating around the bush. There should be a succinct statement of what the study proposes to do at the start (written in the future tense), something like, 'This study will examine …' or 'This study aims to …'.

3. A good proposal uses clear and precise language. While not meant to be a literary masterpiece, all readers with a knowledge of the subject need to be able to understand exactly what is meant in the most concise language possible.

4. A good proposal should be organized. It should be written in simple, logical, prose with clear headings and subheadings to mark out major sections.

What follows is a general guide for putting together a more highly developed proposal in social sciences.

Guidelines for creating a well-developed research proposal

The following elements are important to include:

A.Research topic

- Title of the project
- Nature of the problem or issue under examination (the focus of the study)
- Proposed aims and objectives and research questions/hypotheses

B.Background and context of the study

- How did the problem or issue arise?

Notes

- Why is this an important area to study?

The significance of the study needs to be stated and comments on the practical and/or theoretical value of the research included.

- Include any underlying assumptions.
- Provide definitions of key terms or concepts used.
- Point out the limitations imposed (the boundaries set).

C.Methodology

- A tight fit between the aims of the study and the research strategy chosen must be evident. Include statements on:
- research strategy (e.g., qualitative, quantitative) and justification for approach;
- research methods (e.g. survey, case study, ethnography, experimental);
- tools of data collection (e.g., questionnaire, interviews, focus groups, documentary analysis);
- location and availability of data;
- methods of data analysis and interpretation;
- ethical implications (if relevant); and
- any problems that may be encountered in the conduct of the research.

A key part of a research application is the research proposal. Whether you are applying for self-funded research or organization-funded research, research guidelines should be followed.

D.Literature review (or a general introduction to the topic of 3-5 pages if in experimental sciences)

• Familiarity with the relevant literature needs to be demonstrated.

• A précis of relevant literature needs to include:

– what is already known on the topic;

– what gaps need to be filled;

– how the study relates to, builds on, or differs from previous work in the topic area; and

– theoretical considerations (what theory/ies from the literature would help to develop a meaningful conceptual or analytical framework?)

E.Proposed timeline/milestones

A schedule indicating plans from commencement right through to submission needs to be provided.

F.Resources needed and available

Resources needed should be listed, their availability checked and a budget proposed before beginning the project.

G.Select bibliography or references

When you have completed preparing your proposal, it would be a good idea to self-evaluate what you have produced. A checklist for doing so appears in what follows.

H. Evaluating research proposal

Finally, the proposal can be evaluated in terms of various criteria set by the authorities. The technical evaluation of a research proposal involves various points like:

Notes

TECHNICAL EVALUATION OF RESEARCH PROPOSAL

Title _____

_____

Reviewed by: _____

Date: _____

	YES	NO
A. TITLE		
Does it include the		
• subject matter/scope?	_____	_____
• type of study?	_____	_____
• location/subjects?	_____	_____
• intervention?	_____	_____
• time/duration of observation?	_____	_____
B. SIGNIFICANCE		
Is the study worth undertaking?		
1. The problem	_____	_____
• affects a large population	_____	_____
• has serious consequences	_____	_____
• related to on-going project	_____	_____
2. The answer		
• fills a gap in knowledge	_____	_____
• has practical application	_____	_____
• will improve professional practice/health service	_____	_____
Is the topic in priority list of DOST/DOH?	_____	_____
Is it within the institutional mission?	_____	_____
C. BACKGROUND AND LITERATURE REVIEW		
Are the current issues/findings discussed appropriately?	_____	_____
D. RESEARCH QUESTION		
• Is it clear and adequately formulated?	_____	_____
• Is it researchable?	_____	_____
• Is there critical mass of information?	_____	_____
• Is there current interest in the topic?	_____	_____
• Has an appropriate hypothesis been formulated?	_____	_____
• Is it based on adequate literature review and analysis?	_____	_____
Is the general objective appropriate for the research question?	_____	_____

Notes

F. OBJECTIVES

Is the general objective appropriate for the research question?

Are the specific objectives

- adequate for the general objective?
- clearly stated/incorporates the variables?
- Stated in measurable terms/indicators specified?

G. METHODS

1. Study Design

Is it correct? Appropriate for the research question?

Are confounders to be controlled in

- design phase?
- analysis phase?

2. Subjects

- Is target population suitable to the objectives?
- Is accessible population representative of target population?
- Is the sampling schema/allocation method appropriate?
- Is sample size properly computed?

3. Data Collection

- Are the relevant variables operationally defined?
- Have appropriate tools of measurement been developed?
- Are measures vs bias in place?

4. Data Processing

- Is the data processing to be done by computerization?
- Is the questionnaire/form precoded?

5. Data Analysis and Interpretation

- Is the type of analysis correct?
- Have dummy tables been prepared?
- Are the indicators to be computed suited to the objectives?
- If analytic, is the statistical test correct?
- Is the bases for acceptance/rejection of hypothesis logical?

H. SCHEDULE

- Are all important activities scheduled?
- Is the schedule feasible?

I. BUDGET

- Are all projected expenses included?
- Are th amounts reasonable?

1.4 Research Paradigms

Once the researcher finalizes theresearch topic the next thought is about the approach or methodology to follow for the research. There are threequestions that the researcher needs to ask before beginning the actual research:

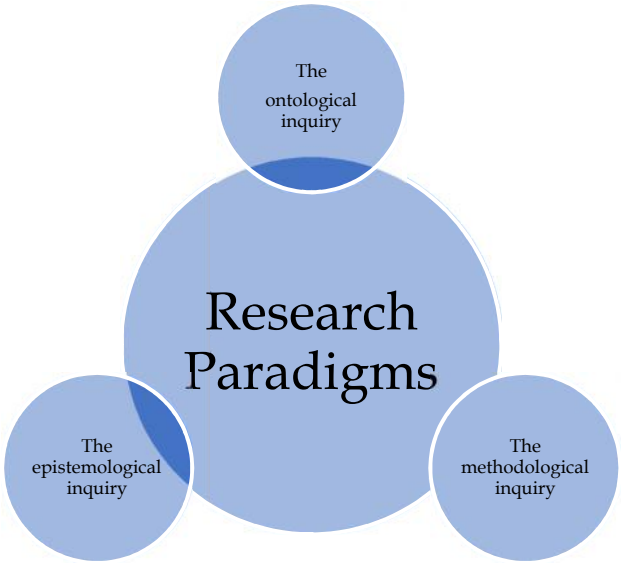


Figure 1: Research Paradigms

1. The ontological inquiry: What is the reality that the researcher wants to explore and know?

2. The epistemological inquiry: What is it (the ontology) that is available to explore and how to reach it?

3. The methodological inquiry: What are the methods and procedures that will make this inquiry possible?

All of the three above questions are part of the paradigms of research. A paradigm is a worldview about how toconduct research. Paradigm includes the methodology, approach, ontology, and epistemology to conduct theresearch. In one paradigm there can be several methodologies and the researcher can follow anyone of them.These methodologies are approaches to research that can help the researcher conduct systematic research.

For example, if ontology asks does God exist? The epistemology will ask how to know that God exists? and themethodology will focus on what procedures and methods one can use to find the existence of God. Theparadigms however are four or five that are internationally accepted, depending on whether you are researching pure sciences or in social sciences. For the new researcher, the choice of the right paradigm andresearch methodology is a difficult task. The researcher getsa better understanding of the paradigm as they workon their research project. The term paradigm was first used by Kuhn in his work The Structure of ScientificRevolution he defined research paradigm as "an integrated cluster of substantive concepts, variables, andproblems attached with corresponding methodological approaches and tools".

A researcher will be curious to know the answers to research questions. The answers to the researchquestions can be solved informally but a researcher will not be able to inform the readers how he/she conduct theresearch. A researcher needs to provide a step-by-step guide to the readers about how theresearch proceeded and how the researcher got the answers to the research question. As we know that everyresearch should have some characteristics these research characteristics give the research meaning and value.Unless the researcher follows a well-defined path to conduct the research, he/she could not justify his findings to thereaders. Additionally, other researchers cannot replicate the study nor they can learn from it. A paradigm providesthe researcher a guide to follow throughout the research. The novice researcher finds it difficult to understandresearch paradigms and their importance in research.

Notes

Ontology and research paradigms

Ontology is the reality of knowingwhat exists and that the research wants to seek. For each researchparadigm, there is an ontological view that the researcher seeks through research. Monism, pluralism, idealism,dualism, materialism are some of the ontological views that one can follow. The ontology cannot be reachedwithout knowing the epistemology of research. In pure sciences, for example, the scientist will use a real reality asan ontological view, and to know that real reality the researcher will use objectivity as an epistemological stance,quantitative methods as the methodology and hence the scientist is using positivism as the paradigm to findanswers to the research questions.

The reality of real reality is an objective way to find answers to the research questions. While interpretivism,constructivism, and pragmatism paradigms have relativism as the ontological approach.

Epistemology and research paradigms

Epistemology is the philosophical view in seeking reality. It paves the way to find the truth that is ontology.Epistemology and ontology are weaved together, and none is possible without each other. Realism, rationalism,relativism, and irrationalism are some of the epistemologies that is out there. Epistemology and ontology are likenail and hammer none can work without each other.In each research paradigm, there is some epistemology the researcher canchoose one that suits the research question.

Methodology and research paradigms

The methodology can be quantitative or qualitative and within each of these methodologies, there are severalresearch techniques. In pure sciences, quantitative research methodology is commonly used. In social sciences,qualitative research methodology is more common in use. As a combination of both quantitative and qualitativeresearch methodologies, there is the mixed-method methodology that is more adaptable and in use in both puresciences and social sciences.For a positivist, quantitative research methodology is more suitable and for an interpretivist qualitative and mixed-method approaches are more common to be used. In interpretivism and pragmatism statistical inquiry or analysisis not always required.

One very important point to be considered here is that once a researcher has decided about either oneof the ontology, epistemology, or methodology of the research he/she is bound to choose the other two from somerestricted choices. The reason being that one cannot apply subjective epistemology to positivism or qualitativemethodology to an objective inquiry. This also reveals how all three are connected and combined to form a research paradigm.Although ontology, epistemology, and methodology have a relationship with the research paradigm but as a researcher, one should know the difference between all of them. Quantitative approach cannot be called a paradigm, itis a methodology or approach to research. A research paradigm is a worldview about conducting research.

The research paradigm however provides researchers an idea to choose methods and research design. Theresearch paradigm is the one that addresses what should be the method to follow for the research and not theother way around.

Contribution of research towards theory

- Research helps to establish what is currently known, and it can be argued as a contribution totheory.
- Research goes beyond the notion of formal literature for discussion about a particular topic.
- The research contributes to knowledge by proposing a framework for reviewing already availableinformation.
- By highlighting an area in which, despite an abundance of knowledge, there is a relatively smallamount of it is known.
- Research through publications provides a piece of evidence to substantiate a claim that it has madea contribution to theory.

- Research devise a method for investigating a particular topic and its contribution is to add toexisting knowledge about the methodology.

1.5 Research Ethics

Research ethics provides guidelines for the responsible conduct of research. Besides, it educates andmonitors scientists conducting research to ensure a high ethical standard. The following is a general summary ofsome ethical principles:

1. Honesty

Honestly report data, results, methods and procedures, and publication status. Do not fabricate, falsify, ormisrepresent data.

2. Objectivity

Strive to avoid bias in experimental design, data analysis, data interpretation, peer review, personnel decisions,grant writing, expert testimony, and other aspects of research.

3. Integrity

Keep your promises and agreements; act with sincerity; strive for consistency of thought and action.

4.Carefulness

Avoid careless errors and negligence; carefully and critically examine your work and the work of your peers.Keep good records of research activities.

5.Openness

Share data, results, ideas, tools, resources. Be open to criticism and new ideas.

6.Respect for Intellectual Property

Honor patents, copyrights, and other forms of intellectual property. Do not use unpublished data, methods, orresults without permission. Give credit where credit is due. Never plagiarize.

7.Confidentiality

Protect confidential communications, such as papers or grants submitted for publication, personnel records, tradeor military secrets, and patient records.

8.Responsible Publication

Publish in order to advance research, not to advance just your career. Avoid wasteful andduplicative publication.

9.Responsible Mentoring

Help to educate, mentor, and advise students. Promote their welfare and allow them to make their own decisions.

10.Respect for Colleagues

Respect your colleagues and treat them fairly.

11.Social Responsibility

Strive to promote social good and prevent or mitigate social harms through research, public education, andadvocacy.

12.Non-Discrimination

Avoid discrimination against colleagues or students based on sex, race, ethnicity, or other factors that arenot related to their scientific competence and integrity.

13.Competence

Maintain and improve your own professional competence and expertise through lifelong education and learning;take steps to promote competence in science.

14.Legality

Notes

Know and obey relevant laws and institutional and governmental policies.

15.Animal Care

Show proper respect and care for animals when using them in research. Do not conduct unnecessary or poorlydesigned animal experiments.

16. Human Subjects Protection

When researching human subjects, minimize harms and risks and maximize benefits, respect humandignity, privacy, and autonomy.

Summary

Research is defined as the creation of new knowledge and/or the use of existing knowledge in a new and creative way so as to generate new concepts, methodologies and understandings. This could include synthesis and analysis of previous research to the extent that it leads to new and creative outcomes.

Research proposal is required to be developed by the researcher to present his/her future course of actions related to research. It includes a detailed description of the research process which a researcher wants to undertake.

A paradigm is a worldview about how to conduct research. Paradigm includes the methodology, approach, ontology, and epistemology to conduct the research. In one paradigm there can be several methodologies and the researcher can follow anyone of them. These methodologies are approaches to research that can help the researcher conduct systematic research.

Key Words

The ontological inquiry: What is the reality that the researcher wants to explore and know?

The epistemological inquiry: What is it (the ontology) that is available to explore.

The methodological inquiry: What are the methods and procedures that will make this inquiry possible?

Self-Assessment

Fill in the blanks:
1. Research is also defined as the creation of…….
2One very common approach in developing a topic is through ……
3. Rather than select a broad topic and try to narrow it down, a researcher should begin with…….
4.A good proposal uses language which is…….
5……help to keep promises and agreements; act with sincerity; strive for consistency of thought and action.

6.A research paradigm is a worldview about conducting…….

Multiple choice questions:

7. The second step in creating a research proposal is……. .
(a) Developing a topic (b) Creating a research plan (c) Exploring purpose and audience interests (d) none of these
8.A research proposal is essential to be used by.....
(a) Researchers (b) Academicians (c) Professionals (d) All of these
9. Background of the research to be undertaken is carried out before……..
(a) Methodology (b) Literature review (c) Introduction (d) Data analysis

10. Which of the following implies meaning of research.
(a) Finding Solution of the research problem (b) Search Again (c) Scientific way to search truth (d) None of these

11.Which of the following explains about methods and procedures that make an inquiry possible?

(a) Methodology (b) Epistemology (c) Ontology (d) None of these
12.Which of the following is NOT required in a research proposal while writing background and context of the study:
(a) Assumptions (b) Significance of the study (c) location and availability of data (d) problems associated with issue

Review Questions

13.What do you mean by research?

14.Write down various points to be considered while preparing a research proposal. Highlight the importance of each point in detail.

15.What are common research paradigms, elaborate.

16.What do you understand by research ethics? Why ethics are necessary in research.

Answers: Self-Assessment

1.	Knowledge	2.	Brainstorming	3.	Specific Problem
4.	Precise and Clear	5.	Integrity	6.	Research
7.	(c)	8.	(d)	9.	(a)
10.	(c)	11.	(a)	12.	(c)

Further Readings

Business Research Methods by Naval Bajpai, Pearson

Research Methodology: Methods and Techniques by Kothari, C. R. & Garg, Gaurav, New Age International.

Marketing Research by Naresh K Malhotra, Pearson

Unit 2: An introduction to Research

CONTENT

Introduction

Research can be referred to as a search for information. It can also be defined as the scientific and systematic search for relevant information on a particular subject or topic. Research is the ability to make scientific inquiries. Research is an original and systematic investigation undertaken to increase existing knowledge and understanding of the unknown to establish facts and principles. Some researchers consider research as a voyage of discovery of new knowledge.It comprises the creation of ideas and the generation of new knowledge that leads to new and improved insights and the development of new materials, devices, products, and processes. It should have the potential to produce results that are sufficiently relevant to increase and synthesize existing knowledge or correcting and integrating previous knowledge.Research is a scientific approach to answering a research question, solving a research problem, or generating new knowledge through a systematic and orderly collection, organization, and analysis of data with the goal of making the findings of research useful in decision-making. Any research endeavor is said to be scientific if:

- It is based on empirical and measurable evidence subject to specific principles of reasoning.
- It consists of systematic observations, measurement, and experimentation.
- It relies on the application of scientific methods.
- It provides scientific information and theories.
- It makes practical applications possible.
- It ensures adequate analysis of data employing rigorous statistical techniques.

Objectives of Research

The purpose of research is to explore answers to questions through the application of scientific procedures. The main aim of the research is to find out the truth which is hidden, and which has not beendiscovered as yet. Though each research study has its specific purpose, we may think ofresearch objectives as falling into several following broad groupings:

1. To gain familiarity with a phenomenon or to achieve new insights into it (studies with thisobjectin view are termed as *exploratory* or *formulative* research studies);

2. To portray accurately the characteristics of a particular individual, situation or a group(studies with this object in view are known as *descriptive* research studies);

3. To determine the frequency with which something occurs or with which it is associatedwith something else (studies with this object in view are known as *diagnostic* research studies);
4. To test a hypothesis of a causal relationship between variables (such studies are known as*hypothesis-testing* research studies).

Characteristics of Research

- Research is directed towards the solution of a problem.
- Research is based upon observable experience or empirical evidence.
- Research demands accurate observation and description.
- Research involves gathering new data from primary sources or using existing data for a new purpose.
- Research activities are characterized by carefully designed procedures.
- Research requires expertise i.e., skill necessary to carryout investigation, search the related literature and to understand and analyze the data gathered.
- Research is objective and logical – applying every possible test to validate the data collected and conclusions reached.
- Research involves the quest for answers to unsolved problems.
- Research requires courage.
- Research is characterized by the patient and unhurried activity.
- Research is carefully recorded and reported.

Criteria of Good Research

One expects scientific research to satisfy the following criteria:

1. The purpose of the research should be clearly defined, and common concepts are used.
2. The research procedure used should be described in sufficient detail to permit another researcher to repeat the research for further advancement, keeping the continuity of what has already been attained.
3. The procedural design of the research should be carefully planned to yield results that are as objective as possible.
4. The researcher should report with complete frankness, flaws in procedural design and estimate their effects upon the findings.
5. The analysis of data should be sufficiently adequate to reveal its significance and the methods of analysis used should be appropriate. The validity and reliability of the data should be checked carefully.
6. Conclusions should be confined to those justified by the data of the research and limited to those for which the data provide an adequate basis.
7. Greater confidence in research is warranted if the researcher is experienced, has a good reputation in research, and is a person of integrity.

Motivation for Research Study

What makes people undertake research? This is a question of fundamental importance. The possible motives for doing research may be either one or more of the following:

1. Desire to get a research degree along with its consequential benefits

Notes

2. Desire to face the challenge in solving the unsolved problems, i.e., concern over practicalproblems initiates research
3. Desire to get the intellectual joy of doing some creative work
4. Desire to be of service to society
5. Desire to get respectability

However, this is not an exhaustive list of factors motivating people to undertake research studies.Many more factors such as directives of the government, employment conditions, curiosity about new things, desire to understand causal relationships, social thinking and awakening, and the like may as well motivate (or at times compel) people to perform research operations.

Research Process

There are a variety of approaches to research in any field of investigation, irrespective of whether it is applied research or basic research. Each particular research study will be unique in some ways because of the particular time, setting, environment, and place in which it is being undertaken.An understanding of the research process is necessary to effectively carry out research and sequencing the stages inherent in the process.

A research design is a detailed blueprint used to guide a research study towards its objective (Aaker et al., 2000). In the introductory section, it has already been discussed that the steps in conducting a research program are interlinked and interrelated. Good research is conducted using 9 steps; they are problem or opportunity identification, decision-maker and business researcher meeting to discuss the problem and opportunity dimensions, defining the management problem and subsequently the research problem, formal research proposal.

Figure 2.1: Steps Involved in Research Process

Notes

Problem or opportunity identification

The process of business research starts with the **problem or opportunity identification**.
The management of the company identifies the problem or opportunity in the organizationor the environment. The management can identify the symptoms or the effectsof the problem, but to understand the reasons for the problems, systematic research has tobe adopted. This required research should either be executed by a business research firm ora business researcher.

Evaluate the cost of research

One of the most frequently asked questions we get at Market Connections is "how much does custom research cost?" That's like walking into a car dealership and asking, "how much does a new car cost?" Before answering either question, several factors need to be considered.

The same can be said when asking about the cost of research. Whether it's focus groups, in-depth interviews, or surveys, the price tag will depend on many factors, including what you want to achieve through the research, who you want to ask the questions of, and how you plan to act on the results. Like any major purchase, understanding your budget and priorities is important to help us ensure we can properly scope the project to best meet your needs and priorities. This is not to gauge you to the top of your range, but to maximize what we can provide, given any constraints. The cost associated with research can be divided based on:

- Qualitative Research
- Quantitative Research

QUALITATIVE RESEARCH

Consider the price of a focus group study. Prices would vary depending on the number of groups, seniority of participants, the narrowness of profession/expertise, or the location of groups. A researcher may be able to secure a single, simple group of government IT professionals for $10 thousand or an eight-group study of mid-to senior-level professionals across multiple cities for $100 thousand. More typically, two groups of business or government participants can cost between $20-35 thousand, and four groups may cost $35-75 thousand.

Here are some other price-related issues affected by the target audience:

- **The budget depends on the number of audience types a researcher is targeting and whether it makes sense to mix them into the same group or give them their group** to ensure an unbiased and more relevant discussion.

- **Where are the customers located?** If they are scattered across the country or the globe, we might very well drop the idea of an in-person group and recommend instead an online focus group as more economical for you and more convenient for the participants.
- **Are you able to provide a contact list of the people you want us to recruit,** or do you want a firm to compile that list? This can affect the price dramatically, depending on who the target is.
- **Is the target audience very senior, or a very specific and hard-to-reach segment?** The researchers intend to discuss highly complex or sensitive issues? Any of these conditions may call for a change in strategy to more private one-on-one in-depth interviews.

While these are the most commonly asked questions, there may be additional factors that could affect the cost of the project. The type of recording, analysis, reporting, participant incentives, and travel can also impact the budget.

Notes

QUANTITATIVE RESEARCH

Conversely, the price for quantitative research can range widely, from $15 thousand to over $100 thousand, with most studies in the $30-$55 thousand range.

Collect Information

The main aim of collecting information is to find out problems that are already investigated and those that need further investigation. It includes an extensive survey of all available past studies relevant to the field of investigation. Its objective is to collect background knowledge of the research topic.It also helps to identify the concepts relating to it, potential relationships between variables.

Research design decision

Research design is the framework of research methods and techniques chosen by a researcher. The design allows researchers to hone in on research methods that are suitable for the subject matter and set up their studies for success.The design of a research topic explains the type of research (experimental, survey, correlational, semi-experimental, review) and also its sub-type (experimental design, research problem, descriptive case-study).

There are three main types of research design: Data collection, measurement, and analysis.

The type of research problem an organization is facing will determine the research design and not vice-versa. The design phase of a study determines which tools to use and how they are used.

Commonly used research designs areas:

Descriptive research design: In a descriptive design, a researcher is solely interested in describing the situation or case under their research study. It is a theory-based design method that is created by gathering, analyzing, and presenting collected data. This allows a researcher to provide insights into the why and how of research. Descriptive design helps others better understand the need for the research. If the problem statement is not clear, you can conduct exploratory research.

Experimental research design: Experimental research design establishes a relationship between the cause and effect of a situation. It is a causal design where one observes the impact caused by the independent variable on the dependent variable. For example, one monitors the influence of an independent variable such as a price on a dependent variable such as customer satisfaction or brand loyalty. It is a highly practical research design method as it contributes to solving a problem at hand. The independent variables are manipulated to monitor the change it has on the dependent variable. It is often used in social sciences to observe human behavior by analyzing two groups. Researchers can have participants change their actions and study how the people around them react to gain a better understanding of social psychology.

Correlational research design: Correlational research is a non-experimental research design technique that helps researchers establish a relationship between two closely connected variables. This type of research requires two different groups. There is no assumption while evaluating a relationship between two different variables, and statistical analysis techniques calculate the relationship between them.A correlation coefficient determines the correlation between two variables, whose value ranges between -1 and +1. If the correlation coefficient is towards +1, it indicates a positive relationship between the variables and -1 means a negative relationship between the two variables.

Notes

Diagnostic research design: In diagnostic design, the researcher is looking to evaluate the underlying cause of a specific topic or phenomenon. This method helps one learn more about the factors that create troublesome situations.

Explanatory research design: Explanatory design uses a researcher's ideas and thoughts on a subject to further explore their theories. The research explains unexplored aspects of a subject and details about what, how, and why of research questions.

Fieldwork and Data Collection

As a next step, fieldwork and data collection activities are planned. An analysis of secondary data sources is also executed to have supporting ideas. It is used in various stages. of the research execution. These are also useful in presenting the findings. The researcher has to also decide whether he or she has to go for a survey or has to adopt the observation methods and decide whether the research will be based on the field data collection or will be a laboratory experiment.

Data preparation and data entry

After fieldwork, the collected data are in raw format. Before performing data analysis, a researcher needs to structure the data. There is a specific scientific procedure to deal with the missing data and other problems related to the data collection process. After preparing the data, a researcher has to feed it into a computer spreadsheet in a pre-determined manner to execute the data analysis exercise. Preparing this data matrix through the spreadsheet is also a scientific exercise and requires a lot of expertise and experience.

Data Analysis

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.A simple example of Data analysis is whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision. This is nothing but analyzing our past or future and making decisions based on it.

Interpretation of result and presentation of findings

It has been already discussed that after applying data analysis techniques, a statistical result is obtained. There is a need to interpret the result and present the non-statistical findings derived from the statistical result. A meaningful interpretation of the result is a skillful activity and is an important aspect of research. The researcher has to determine whether the result of the study is in line with the existing literature. It is also important to present the findings scientifically. The results obtained from the analysis are statistical.

Management Decision and Its Implementation

As the last step of conducting a research program, the findings are conveyed to the decision-maker after consultation with the research programmer. The decision-maker analyses the findings and takes an appropriate decision in the light of the statistical findings presented by the researcher. This is not a formal part of the research process. Here, it is included as a step of the research process because it is the decision-maker who will ultimately take the decision and is the managerial implication of the research programme.

Notes

Write down the steps needed to research on "Attitude of Bank customers towards online banking.

Research Problem

Defining a research problem is the fuel that drives the scientific process and is the foundation of any research method and experimental design, from true experiment to case study.It is one of the first statements made in any research paper and, as well as defining the research area, should include a quick synopsis of how the hypothesis was arrived at.Operationalization is then used to give some indication of the exact definitions of the variables, and the type of scientific measurements used.

This will lead to the proposal of a viable hypothesis. As an aside, when scientists are putting forward proposals for research funds, the quality of their research problem often makes the difference between success and failure.

Defining a Research Problem

Formulating the research problem begins during the first steps of the scientific process.

As an example, a literature review and a study of previous experiments, and research, might throw up some vague areas of interest.Many scientific researchers look at an area where a previous researcher generated some interesting results but never followed up. It could be an interesting area of research, which nobody else has fully explored.A scientist may even review a successful experiment, disagree with the results, the tests used, or the methodology, and decide to refine the research process, retesting the hypothesis.

This is called the conceptual definition and is an overall view of the problem. A science report will generally begin with an overview of the previous research and real-world observations. The researcher will then state how this led to defining a research problem.

A research problem is a statement about an area of concern, a condition to be improved, a difficulty to be eliminated, or a troubling question that exists in scholarly literature, in theory, or in practice that points to the need for meaningful understanding and deliberate investigation.A problem statement is a clear description of the issue(s), it includes a vision, issue statement, and method used to solve the problem.

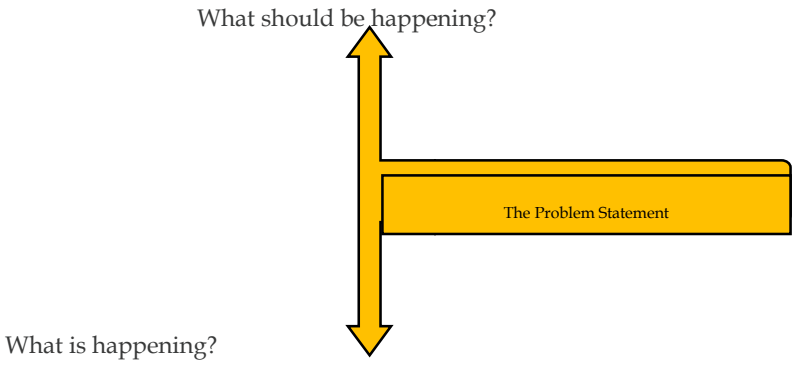


Figure 2.1: Description of Problem Statement

Writing a problem statement

It is used at the beginning to keep research on track during theeffort and is used to validate that the effort delivered an outcome that solves the problem statement.

Notes

Basic characteristics of research problem:

- Reflecting on important issues or needs
- Based on factual evidence (it's non-hypothetical)
- It should be manageable and relevant.
- Suggest a testable and meaningful hypothesis (avoiding useless answers)
- Formulating a research problem is the first and most important step of the research process.
- Identify the destination before taking the journey.
- Research question = Foundation of building

Research problem determines:

- type of study design
- type of sampling strategy
- research instrument
- type of analysis

Sources of research problems:

- Journal, article, etc.
- Personal interest and experiences
- Deduction from theory
- Experts
- Conversation with colleagues or at professional conferences.
- Observation
- Literature reviews
- Replication of studies

A **problem statement** is a short, clear explanation of the issue to be researched. It sets up the context, relevance, and aims of the project.

Research Design

The word 'design' has various meanings. But, about the subject concern, it is a pattern or an outline of the research project's workings. It is the statement of essential elements of a study that provides basic guidelines for conducting the project. It is the same as the blueprint of an architect's work.

The research design is similar toa broad plan or model that states how the entire research project would be conducted. It is desirable that it must be in written form and must be simple and clearly stated. The real project is carried out as per the research design laid down in advance.

Contents of Research Design

1. Statement of research objectives, i.e., why the research project is to be conducted

2. Type of data needed

3. Definition of population and sampling procedures to be followed

4. Time, costs, and responsibility specification

5. Methods, ways, and procedures used for collection of data

6. Data analysis – tools or methods used to analyze data

7. Probable output or research outcomes and possible actions to be taken based on those outcomes

Types of Research Designs

The research design is a broad framework that describes how the entire research project is carried out. Basically, there can be three types of research designs – exploratory research design, descriptive research design, and experimental (or causal) research design. Use of particular research design depends upon the type of problem under study.

Exploratory Research Design

This design is followed to discover ideas and insights to generate possible explanations. It helps in exploring the problem or situation. It is, particularly, emphasized to break a broad vague problem statement into smaller pieces or sub-problem statements that help to forma specific hypothesis.

The hypothesis is a conjectural (imaginary, speculative, or abstract) statement about the relationship between two or more variables. Naturally, in the initial stage of the study, we lack sufficient understanding ofthe problem to formulate a specific hypothesis. Similarly, we have several competing explanations of the marketing phenomenon. Exploratory research design is used to establish priorities among those competitive explanations.

The exploratory research design is used to increase the familiarity of the analyst with a problem under investigation. This is particularly true when the researcher is new in the area, or when a problem is of a different type.

This design is followed to realize the following purposes:

1. Clarifying concepts and defining the problem

2. Formulating problem for more precise investigation

3. Increasing the researcher's familiarity with the problem

4. Developing hypotheses

5. Establishing priorities for further investigation

Exploratory research design is characterized by the flexibility to gain insights and develop hypotheses. It does not follow a planned questionnaire or sampling. It is based on a literature survey, experimental survey, and analysis of selected cases. Unstructured interviews are used to offer respondents a great deal of freedom. No research project is purely and solely based on this design. It is used as complementary to descriptive design and causal design.

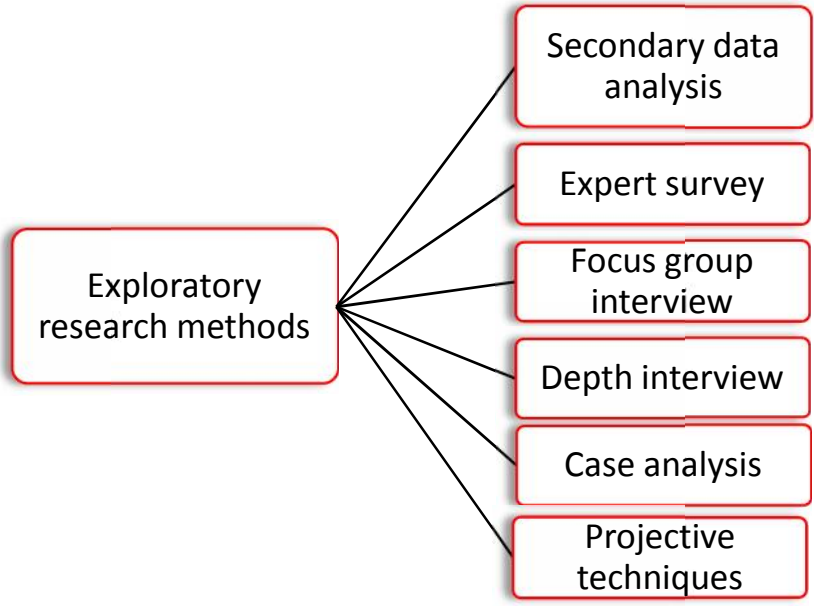


Figure 2.2: Exploratory Research Methods

Types of Exploratory research design

While it may sound a little difficult to research something that has very little information about it, there are several methods which can help a researcher figure out the best research design, data collection methods and choice of subjects. There are two ways in which research can be conducted namely primary and secondary. Under these two types, there are multiple methods which can used by a researcher. The data gathered from these researches can be qualitative or quantitative. Some of the most widely used research designs include the following:

1. Primary research methods

Primary research is information gathered directly from the subject. It can be through a group of people or even an individual. Such research can be carried out directly by the researcher himself or can employ a third party to conduct it on their behalf. Primary research is specifically carried out to explore a certain problem that requires an in-depth study.

- **Expert Surveys/polls**: Surveys/polls are used to gather information from a predefined group of respondents. It is one of the most important quantitative methods. Various types of surveys or polls can be used to explore opinions, trends, etc. With the advancement in technology, surveys can now be sent online and can be very easy to access. For instance, the use of a survey app through tablets, laptops, or even mobile phones. This information is also available to the researcher in real-time as well. Nowadays, most organizations offer short length surveys and rewards to respondents, to achieve higher response rates.
  For example, A survey is sent to a given set of audience to understand their opinions about the size of mobile phones when they purchase one. Based on such information organizations can dig deeper into the topic and make the business-related decision.

- **Depth Interviews:** While you may get a lot of information from public sources, but sometimes an in-person interview can give in-depth information on the subject being

studied. Such research is a qualitative research method. An interview with a subject matter expert can give you meaningful insights that a generalized public source won't be able to provide. Interviews are carried out in person or on the telephone which has open-ended questions to get meaningful information about the topic.

For example,An interview with an employee can give you more insights to find out the degree of job satisfaction, or an interview with a subject matter expert of quantum theory can give you in-depth information on that topic.

- **Focus group Interviews:** Focus group is yet another widely used method in exploratory research. In such a method a group of people is chosen and are allowed to express their insights on the topic that is being studied. Although, it is important to make sure that while choosing the individuals in a focus group they should have a common background and have comparable experiences.

For example, A focus group helps researchers identify the opinions of consumers if they were to buy a phone. Such research can help the researcher understand what the consumer value while buying a phone. It may be screen size, brand value, or even the dimensions. Based on which the organization can understand what are consumer buying attitudes, consumer opinions, etc.

- **Case Analysis**: A case study is an in-depth study of a particular situation rather than a sweeping statistical survey. It is a method used to narrow down a very broad field of research into one easily researchable topic. Whilst it will not answer a question completely, it will give some indications and allow further elaboration and hypothesis creation on a subject. The case study research design is also useful for testing whether scientific theories and models work in the real world.

- **Projective Techniques:** Projective techniques allow respondents to project their subjective or true opinions and beliefs onto other people or even objects. The respondent's real feelings are then inferred from what s/he says about others.Projective techniques are normally used during individual or small group interviews. They incorporate several different research methods. Among the most commonly used are:

  Word association test

  Sentence completion test

  Thematic apperception test (TAT)

  Third-person techniques

Secondary data analysis

Secondary research is gathering information from previously published primary research. In such research, you gather information from sources like case studies, magazines, newspapers, books, etc.

- **Online research:** In today's world, this is one of the fastest ways to gather information on any topic. A lot of data is readily available on the internet and the researcher can download it whenever he needs it. An important aspect to be noted for such a research is the genuineness and authenticity of the source websites that the researcher is gathering the information from.

For example: A researcher needs to find out what is the percentage of people that prefer a specific brand phone. The researcher just enters the information he needs in a search engine and gets multiple links with related information and statistics.

- **Literature research**: Literature research is one of the most inexpensive method used for discovering a hypothesis. There is tremendous amount of information available in libraries, online sources, or even commercial databases. Sources can include newspapers, magazines, books from library, documents from government agencies, specific topic related articles, literature, Annual reports, published statistics from research organizations, and so on.

However, a few things have to be kept in mind while researching from these sources. Government agencies have authentic information but sometimes may come with a nominal cost. Also, research from educational institutions is generally overlooked, but in fact, educational institutions carry out more research than any other entities.Furthermore, commercial sources provide information on major topics like political agendas, demographics, financial information, market trends, and information, etc.

For example, A company has low sales. It can be easily explored from available statistics and market literature if the problem is market-related or organization-related or if the topic being studied is regarding the financial situation of the country, then research data can be accessed through government documents or commercial sources.

2. *Descriptive Research Design*

Descriptive research design is typically concerned with describing the problem and its solution. It is a more specific and purposive study. Before rigorous attempts are made for descriptive study, the well-defined problem must be on hand. The descriptive study rests on one or more hypotheses.

For example, "our brand is not much familiar," "sales volume is stable," etc. It is more precise and specific. Unlike exploratory research, it is not flexible. Descriptive research requires clear specification of who, why, what, when, where, and how of the research. Descriptive design is directed to answer these problems. It is further classified into Cross-Sectional and Longitudinal Study types.

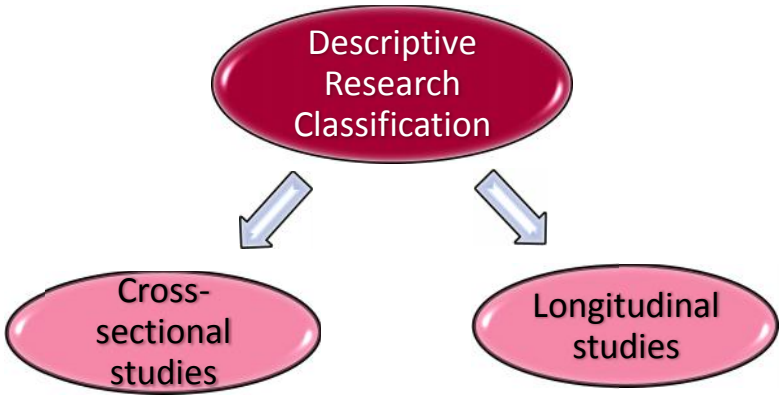


Figure 2.3: Descriptive Research Classification

Cross-Sectional Study: Cross-sectional study is defined as an observational study where data is collected to study a population at a single point in time to examine the relationship between variables of interest.

1. In an observational study, a researcher records information about the participants without changing anything or manipulating the natural environment in which they exist.
2. The most important feature of a cross-sectional study is that it can compare different samples at one given point in time. For example, a researcher wants to understand the relationship between joggers and level of cholesterol, he/she might want to choose two age groups of daily joggers, one group is below 30 but more than 20 and the other, above 30 but below 40 and compare these to cholesterol levels amongst non-joggers in the same age categories.
3. The researcher at this point in time can create subsets for gender but cannot consider past cholesterol levels as this would be outside the given parameters for cross-sectional studies.
4. Cross-sectional studies allow the study of many variables at a given time. Researchers can look at age, gender, income etc. in relation to jogging and cholesterol at a very little or no additional cost involved.
5. However, there is one downside to cross-sectional study, this type of study is not able to provide a definitive relation between cause and effect relation (a cause and effect relationship is one where one action (cause) makes another event happen (effect), for example, without an alarm, you might oversleep.)
6. This is majorly because cross-sectional study offers a snapshot of a single moment in time, this study doesn't consider what happens before or after. Therefore in this example stated above it is difficult to know if the daily joggers had low cholesterol levels before taking up jogging or if the activity helped them to reduce cholesterol levels that were previously high.

- **Longitudinal Study:** Longitudinal study, like the cross-sectional study, is also an observational study, in which data is gathered from the same sample repeatedly over an extended period of time. Longitudinal study can last from a few years to even decades depending on what kind of information needs to be obtained.
  1. The benefit of conducting longitudinal study is that researchers can make notes of the changes, make observations and detect any changes in the characteristics of their participants. One of the important aspects here is that longitudinal study extends beyond a single frame in time. As a result, they can establish a proper sequence of the events occurred.
  2. Continuing with the example, in longitudinal study a researcher wishes to look at the changes in cholesterol level in women above the age of 30 but below 40 years who have jogged regularly over the last 10 years. In longitudinal study setup, it would be possible to account for cholesterol levels at the start of the jogging regime, therefore longitudinal studies are more likely to suggest a cause-and-effect relationship.
  3. Overall, research should drive the design, however, sometimes as the research progresses it helps determine which of the design is more appropriate. Cross-sectional studies can be done more quickly as compared to longitudinal studies. That's why a researcher may start off with cross-sectional study and if needed follow it up with longitudinal studies.

3. Causal or Experimental Research Design

Causal research design deals with determining cause and effect relationship. It is typically in form of experiment. In causal research design, attempt is made to measure impact of manipulation on

independent variables (like price, products, advertising and selling efforts or marketing strategies in general) on dependent variables (like sales volume, profits, and brand image and brand loyalty). It has more practical value in resolving marketing problems. We can set and test hypotheses by conducting experiments.Test marketing is the most suitable example of experimental marketing in which the independent variable like price, product, promotional efforts, etc., are manipulated (changed) to measure its impact on the dependent variables, such as sales, profits, brand loyalty, competitive strengths product differentiation and so on.

Review Questions

1 What are the steps in business research processdesign?

2 What is the difference between a management problemand a research problem?

3 What are the different types of research?

4 For what purposes, exploratory research is used?

5 What is descriptive research and when do researchers conduct it.

Multiple choice questions:

6. What is a research design?
(a) A way of conducting research that is not grounded in theory
(b) The choice between using qualitative or quantitative methods
(c) The style in which you present your research findings, e.g. a graph
(d) A framework for every stage of the collection and analysis of data

7. Which of the following should be included in a research proposal?
(a) Your academic status and experience
(b) The difficulties you encountered with your previous reading on the topic
(c) Your choice of research methods and reasons for choosing them
(d) All of the above

8. Advance plan of research is called as
(a) Research process
(b) Research design
(c) Research proposal
(d) None of the above

9.Research design consist of following things except…….
(a) Hypothesis
(b) Expenditure
(c) Research problem
(d) None of the above

10.Source of research problem include:
(a) Researcher's experience
(b) Practical issue that require solutions
(c) Theory and past research
(d) All of the above

11…… is snapshot of some aspect of the market environment.
(a) Causal
(b) Exploratory
(c) Descriptive
(d) None of the above

12.Following are techniques of Qualitative Research ?
(a) Depth interview
(b) Focus group
(c) Projective technique
(d) All of the above

Fill in the blanks:

13. ……are the two types of research data.
14…… is the form of data below can usually be obtained more quickly and at a lower cost than the others
15.……design deals with determining cause and effect relationship

Answers: Self-Assessment:

6. (d) 7. (c) 8. (b) 9. (a) 10. (d) 11. (c) 12(d) 13. Qualitative and Quantitative 14. Secondary 15. Causal Research

Further Readings:

1. Business Research Methods by Naval Bajpai, Pearson
2. Research Methodology: Methods and Techniques by Kothari, C. R. & Garg, Gaurav, New Age International.
3. Marketing Research by Naresh K Malhotra, Pearson

1. https://www.iedunote.com/research-process
2. www.wisdomjobs.com/e-university/research-methodology-tutorial-355/motivation-in-research
3. www. muet-crp.yolasite.com
4. www. callygood.medium.com
5. www. smallbusiness.chron.com
6. www.courses.lumenlearning.com
7. www. www.iedunote.com

Unit 3:Reviewing Literature

CONTENTS

Objectives

After studying this unit, you will be able to

- identify the appropriate content of selected material for research.
- highlight requirements for academic writing.
- overview referencing practices in research.
- gain an understanding of the existing literature related to research.
- overview steps involved in literature review process.

Introduction

Research creates the need to draw boundaries around an idea, topic or subject area. It is helpful to think about how and where information for the area under research is generated. For this, the researcher needs to identify the branches of knowledge creation in a subject area.

Information does not exist in an environment like raw materials, instead it is created by individuals working in a specific field of knowledge who use specialized methods to generate new knowldege. Disciplines use, produce and disseminate knowledge. Looking at the list of university courses reveals the key to a disciplined structure. Areas such as political science, biology, history and mathematics are as unique disciplines as social work. Everyone has their own logic for how new knowledge is introduced and made accessible. The researcher needs to be comfortable in identifying the branches that can provide information in any search.For example, think of disciplines that can provide information on a topic such as the role of sport in society. A researcher tries to anticipate what kind of perspective each discipline has on the subject. The following types of questions can examine what different branches can contribute.

- What is important about the topic to the people in that discipline?
- What is most likely to be the focus of their study about the topic?
- What perspective would they be likely to have on the topic?

3.1 Review of Literature

Literature review is one of the most important steps in the research process. It is an account of what is already known about a particular topic. It is a critical view of existing information, especially studies that are significant for continuous research. The main purpose of the literature review is to convey to the readers about the work already done, the established knowledge and ideas on a particular subject of research. Although the review of the literature largely concerns existing knowledge, it equally emphasizes its potential use in the future. Therefore, a strong literature review draws insights from current or current knowledge and shows a direction to ongoing study that will be significant for the future.

In general, literature review embraces: conceptual review, theoretical review, and empirical review. In the first part of the review, the researcher generally needs to clarify the research topic by providing an interesting terminological explanation. Because of this, researchers define terminology and describe research problems. In the theoretical review section, researchers mention some related theories for backing up a proposed study. By reviewing empirical studies, researchers briefly summarize previous research and show the research gap (s) through its critique. In doing so, researchers "highlight the agreement and differences between the authors / theories and identify unanswered questions or gaps" (Kumar, 1996, p.30). Through any literature review, the goal of the researchers is to reach the existing gap. Therefore, literature review is a daunting task, but it is essential if the research process is to be successful. Moreover, it "gives credibility and legitimacy to research" (Cohen, Manion and Morrison, 2011, p. 112).

Purpose ofLiterature Review

The purpose of literature review is to convey to the reader previous knowledge, facts established on an issue and their strengths and weaknesses. Literature review allows the reader to update with the status of the research by allowing them to summarize, evaluate and compare original research in the field. Some of these purposes may be as follows:

- It helps in identifying research problems and developing and improving research questions. It also helps to develop hypotheses for testing in research studies.
- It projects what is known and not known about the field of research to find out how research can best contribute to that knowledge. Literature review also attempts to provide a description of the strengths and weaknesses of the design / investigation methods and devices used in previous research work and also shows any gaps or inconsistencies in the body of knowledge.
- It also reveals unanswered questions about topics, concepts or problems. Similarly, it stimulates the need to replicate previous studies in different study settings or different samples or sizes or different study populations.
- It provides a relevant theoretical or conceptual framework for research problems, so it helps the researcher identify or develop new or refine the existing theories through empirical research.
- It helps to plan the methodology of existing research studies.
- It also helps in the development of research devices, identification of suitable designs for research studies and assistance in data collection methods.
- It is important in interpreting the findings of the study, developing effects and recommendations. It also points the way forward for further research.

Give your views on the importance of literature review for a research.

Literature Review: Its Types

It is important to include three important aspects of knowledge in each field. First, there are primary studies that researchers do and publish. Second, it is the study reviews that summarize and offer new interpretations and often extend beyond the original study. Third, there are

assumptions, conclusions, opinions and interpretations that are informally shared that form part of current understanding in the field.

When designing a literature review, it is important to note that this third level of knowledge is always cited as "true", although it often has a loose relationship with primary studies and secondary literature reviews. Given this, while literature reviews are designed to provide an overview and synthesis of the relevant sources a researcher has discovered, there are many approaches to how they can be done, depending on the type of analysis a researcher is considering.

Listed below are definitions of types of literature reviews:

Argumentative Review

This form selects the literature to support or refute the argument, deeply understood assumptions or philosophical problems already established in the literature. The aim is to develop a body of literature that establishes contradictory perspectives. Given the value-filled nature of some social science research [e.g. Educational reform; Immigration control], an argumentative approach to analyse literature can be genuine and important form of discourse. However, they can also present bias problems when used to make good claims of the sort found in systematic reviews.

Integrative Review

Integrative review refers to a type of research that seamlessly reviews, critiques, and produces typical literature on a topic, generating new frameworks and perspectives. The body of literature includes all studies that address related or similar hypotheses. A well-conducted integrated review meets the same standards as primary research on clarity, rigidity, and replication.

Historical Review

Historical reviews are focused on examining research throughout a period, often starting with the first time an issue, concept, theory, phenomena emerged in the literature, then tracing its evolution within the scope of a discipline. The purpose is to place research in a historical context to show familiarity with state-of-the-art developments and to identify the likely directions for future research.

Methodological Review

Reviews do not always focus on what someone said [the content], but how they said it [method of analysis]. This approach provides a framework of understanding at different levels (e.g. theory, significant fields, research approaches and data collection and analysis techniques), enabling researchers to use a wide range of knowledge from the conceptual level to practical documents. It can find its uses in the areas of inquiries related to ontological and epistemological consideration, quantitative and qualitative integration, sampling, interviewing, data collection and data analysis, and helps highlight many ethical issues which a researcher should be aware of and consider conducting the study.

Systematic Review

This form of literature review contains an overview of the existing evidence related to clearly devised research question, which uses pre-determined and standardized methods to identify and critically evaluate the relevant research. It is further characterised by collection, reporting, and analysis of data from previous studies. Typically, it focuses on a very specific empirical question, which often arises in the form of cause-effect, such as "How much does A contribute to B?"

Theoretical Review

The purpose of this form of review is to examine essence of theory that is stored about an issue, concept, theory, event. Theoretical literature review helps to establish the theories that already exist, the relationships between them, the degree to which existing theories have been examined, and to develop new hypotheses to test. This form of literature review is often used reveal findings that current theories are inadequate for explaining new or is used in emerging research problems. The unit of analysis can focus on a theoretical concept or a whole theory or framework.

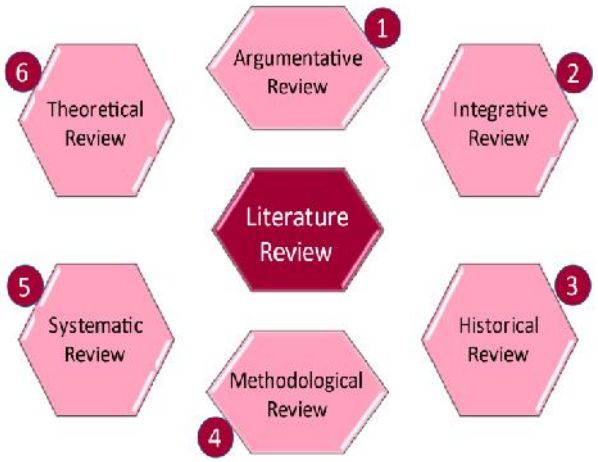


Figure 3.1: Types of Literature Review

Sources of Literature

The Literature refers to the collection of scholarly writings on a topic. This includes peer-reviewed articles, books, dissertations and conference papers. Sources are considered primary, secondary, or tertiary depending on the **originality** of the information presented and their **proximity** or how close they are to the source of information. This distinction can differ between subjects and disciplines.

In the current scenario, research findings may be communicated informally between researchers through email, presented at conferences (primary source), and then, possibly, published as a journal article or technical report (primary source). Once published, the information may be commented on by other researchers (secondary sources), and/or professionally indexed in a database (secondary sources). Later the information may be summarized into an encyclopaedic or reference book format (tertiary sources).

The sources of literature can be classified under two broad headings:

- Primary Sources
- Secondary Sources

Primary Sources

A primary source is a document or record that reports on a study, experiment, trial or research project. Primary sources are usually written by the person(s) who did the research, conducted the study, or ran the experiment, and include hypothesis, methodology, and results.

Primary Sources include:

- Pilot/prospective studies
- Cohort studies
- Survey research
- Case studies
- Lab notebooks
- Clinical trials
- Dissertations

Secondary Sources

Secondary sources of research contain descriptions of studies prepared by another person rather than the original researcher. Secondary sources list the primary information and studies. Itsummaries, compares and evaluatesto present the current state of knowledge in the subject.

Sources may include a bibliography that may lead to the primary research. There are various types of secondary sources, these are:

- Electronic database
- Printed Sources
- Conference Papers
- Theses
- Encyclopaedia/Dictionary
- Research Reports

Electronic database

Electronic literature search is very useful, but sometimes it can be time consuming & unpredictable because there are many website & web pages that can lead to information overload & confusion. But there are available some online databases that make it easy to find the research published in online journals.

Examples: Scopus, Pubmed, Web of Science

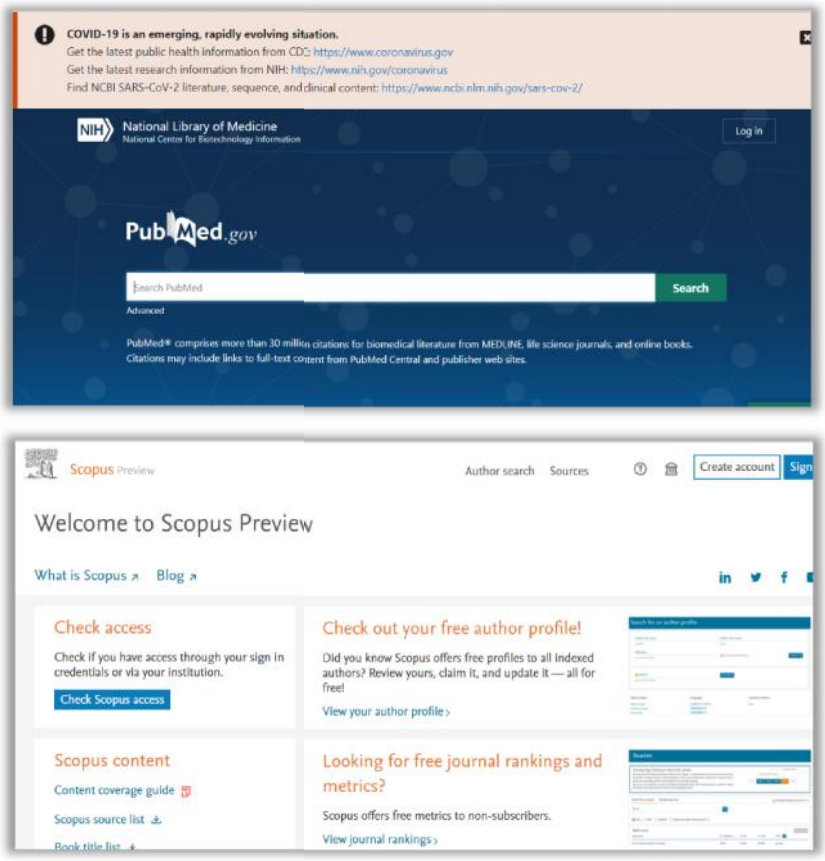


Figure 3.2: a) Pubmed b) Scopus (Electronic databases)

Printed Sources

Printed sources incudes journals, trade publications, and magazines. Printed sources find their way as a source of literature in various forms which have been listed below:

Magazines: A magazine is a collection of articles and images about diverse topics of popular interest and current events.

Features of magazines:

- articles are usually written by journalists.
- articles are written for the average adult
- articles tend to be short
- articles rarely provides a list of reference sources at the end of the article
- lots of color images and advertisements
- the decision about what goes into the magazine is made by an editor or publisher
- magazines can have broad appeal, like *Time* and *Newsweek*, or a narrow focus, like *Sports Illustrated* and *Mother Earth News*.

Trade Publications: Trade publications or trade journals are periodicals directed to members of a specific profession. They often have information about industry trends and practical information for people working in the field.

Features of trade publications:

- Authors are specialists in their fields.
- Focused on members of a specific industry or profession.
- No peer review process
- Include photographs, illustrations, charts, and graphs, often in color.
- Technical vocabulary

Scholarly, Academic, and Scientific Publications : Scholarly, academic, and scientific publications are a collections of articles written by scholars in an academic or professional field. Most journals are peer-reviewed or refereed, which means a panel of scholars reviews articles to decide if they should be accepted into a specific publication. Journal articles are the main source of information for researchers and for literature reviews.

Features of journals:

- written by scholars and subject experts.
- author' credentials and institution will be identified.
- written for other scholars.
- dedicated to a specific discipline that it covers in depth.
- often report on original or innovative research
- long articles, often 5-15 pages or more
- articles almost always include a list of sources at the end (Works Cited, References, Sources, or Bibliography) that point back to where the information was derived.
- no or very few advertisements.
- published by organizations or associations to advance their specialized body of knowledge.

Primary, Secondary, and Tertiary Sources: Primary sources of information are those types of information that come first. Example includes: original research, like data from an experiment. It also includes diaries, journals, photographs data from the census bureau or a survey

There are different types of primary sources for different disciplines. In the discipline of history, for example, a diary or transcript of a speech is a primary source. In education and nursing, primary sources will generally be original research, including data sets.

Secondary sources are written about primary sources to interpret or analyze them. They are a step or more removed from the primary event or item. Some examples of secondary sources are: commentaries on speeches critiques of plays, journalism, or books a journal article that talks about a primary source eg: text book, biographies etc

Conference Papers : Conference papers refer to articles that are written with the goal of being accepted to a conference: typically, an annual (or biannual) venue with a specific scope where you

can present your results to the community, usually as an oral presentation, a poster presentation, or a tabled discussion.

Thesis: A thesis is a theory which is based on own ideas and research and represent in a logical way. It is one of the most important concepts of college expository writing. It usually consists of a several original research that has already been carried out and seeks to find a particular framework for a strong opinion.

Thesis can be a source of literature and can be written at various levels as mentioned below:

a. Undergraduate Thesis

b. Masters Thesis

c. Doctoral Thesis

Research Reports: Research reports are recorded data prepared by researchers or statisticians after analysing information gathered by conducting organized research, typically in the form of surveys or qualitative methods.

A Quality journal approves the article for publication only after getting it reviewed from various subject experts.

3.2 Academic Writing

Academic writing attempts to fulfil the objectives which are non-commercial and is limited to education purpose. Academic writing is always clear, concise, focussed, structured and backed up by evidence. Its purpose is to aid the reader's/researcher's understanding.It has a formal tone and style, but it is not complex and does not use long sentences and complicated vocabulary.Each subject discipline will have certain writing conventions, vocabulary and types of discourse that a researcher will become familiar with over the course of your degree. However, there are some general characteristics of academic writing that are relevant across all disciplines.

Characteristics of academic writing

Academic writing is:

- **Planned and focused:** answers the question and demonstrates an understanding of the subject.
- **Structured:** is coherent, written in a logical order, and brings together related points and material.
- **Evidenced:** demonstrates knowledge of the subject area, supports opinions and arguments with evidence, and is referenced accurately.
- **Formal in tone and style:** uses appropriate language and tenses, and is clear, concise and balanced.

Modeling Customers' Intention to Use E-Wallet in a Developing Nation:
Extending UTAUT2 With Security, Privacy and Savings

Vishal Sindhu, Lovely Professional University, Phagwara, India
https://orcid.org/0000-0002-9650-4649
Avinash Rana, Lovely Professional University, Phagwara, India

ABSTRACT

Intention to use e-wallets is affected by a number of factors which are related to consumer perception about privacy, security, price value, benefits etc. A sample of 613 customers of e-wallets in Punjab state in India was validated through mall intercept method. The results indicate that hedonic motivation, perceived security, general privacy, facilitating conditions, performance expectancy, perceived savings and social influence, and price value in this order, influence the intention to adopt e-wallets. Habit and effort expectancy are the hindrances that have a negative impact on the e-wallet adoption. Factors such as hedonic motivations, security, and privacy have larger roles. The service providers should maintain the privacy and security of users and engage customers by modifying the existing services' range and features. The study endorses reduction in the efforts of using e-wallets, and the conversion of habit into more willingly performed behavior. The resulting model can draw meaningful insights about adoption of this emerging payment platform.

KEYWORDS

Adoption, E-Wallets, Mobile, Privacy, Savings, Security, Technology, UTAUT2

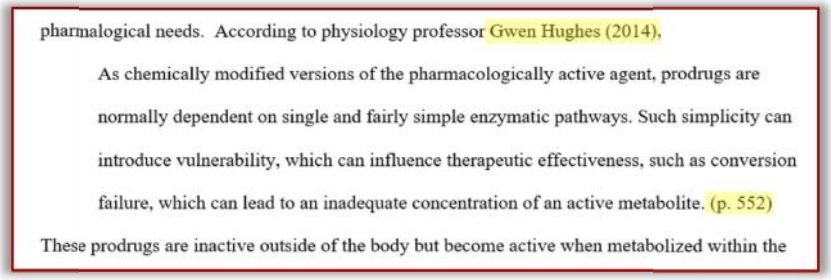
1. INTRODUCTION

Past several years have marked the popularity of mobile commerce applications, such as mobile shopping, mobile travel service, online to offline (O2O) consumption. (Yang, Lin, Li, & Yu, 2015). This has created plenty of opportunities for organizations by making versatile innovation based business models (Akturan & Tescan, 2012). The innovative payment methods and services take up diverse names such as mobile wallets, mobile payments, mobile commerce, and mobile banking. Amongst all forms of online payments, an e-wallet is a prepaid account in which the user can place his/her money, and its underlying components are software (stores personal information, grant security and encryption) and information (database of user details) (Sharma, 2017). Mobile commerce accounts

Figure 3.3: Example of Academic Writing (Research Paper)

Citation

Broadly, a citation is a reference to a published or unpublished source (not always the original source).More precisely, a citation is an abbreviated alphanumeric expression embedded in the body of an intellectual work that denotes an entry in the bibliographic references section of the work for the purpose of acknowledging its relevance.Generally, the combination of both the in-body citation and the bibliographic entry constitutes what is commonly thought of as a citation.



pharmalogical needs. According to physiology professor Gwen Hughes (2014),

> As chemically modified versions of the pharmacologically active agent, prodrugs are normally dependent on single and fairly simple enzymatic pathways. Such simplicity can introduce vulnerability, which can influence therapeutic effectiveness, such as conversion failure, which can lead to an inadequate concentration of an active metabolite. (p. 552)

These prodrugs are inactive outside of the body but become active when metabolized within the

Figure 3.4: Citation

Referencing

Academic writing relies on more than just the ideas and experience of one author. It also uses the ideas and research of other sources: books, journal articles, websites, and so forth.Referencing is used to tell the reader where ideas from other sources have been used in an assignment.It shows the reader that you can find and use sources to create a solid argument. It properly credits the originators of ideas, theories, and research findings. It shows the reader how your argument relates to the big picture. Different academic disciplines have priorities of what is important to the subsequent reader of an academic paper, and different journals have differing rules about the citation of sources.

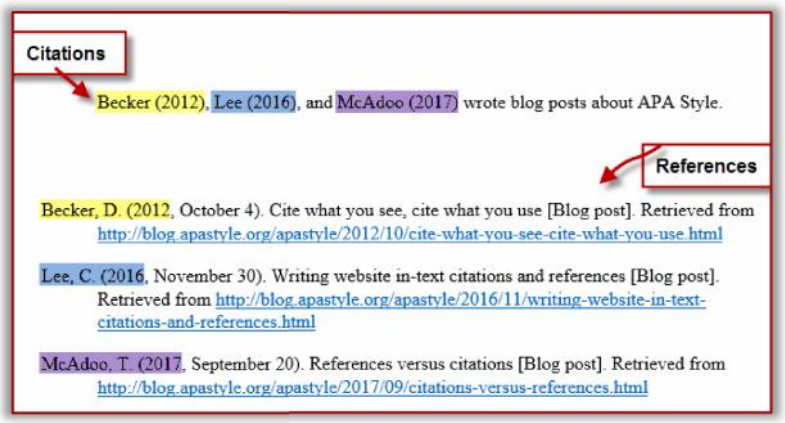


Figure 3.5: Difference between Citations and References

Referencing Styles

APA

APA stands for "American Psychological Association" and comes from the association of the same name.Although originally drawn up for use in psychological journals, the APA style is now widely used in the social sciences, in education, in business, and numerous other disciplines.

Example: - Pinker, S. (1999). Words and rules: The ingredients of language. London: Phoenix

Chicago

Chicago is sometimes referred to as Turabian or Chicago/TurabianChicago is used mainly in the social sciences, including history, political studies, and theology.

Example: - Grazer, Brian, and Charles Fishman. A Curious Mind: The Secret to a Bigger Life. New York: Simon & Schuster, 2015.

Vancouver

Originally came from The International Committee of Medical Journal Editors which produced the "Uniform Requirements for Manuscripts Submitted to Biomedical Journals" following a meeting that was held in Vancouver in 1978 [Source: Jönköping University Library].

The Vancouver style is used mainly in the medical sciences.

Example: - Ramalho R, Helffrich G, Schmidt DN, Vance D. Tracers of uplift and subsidence in the Cape Verde archipelago. Journal of the Geological Society. 2010;167(3): 519–538.

Harvard

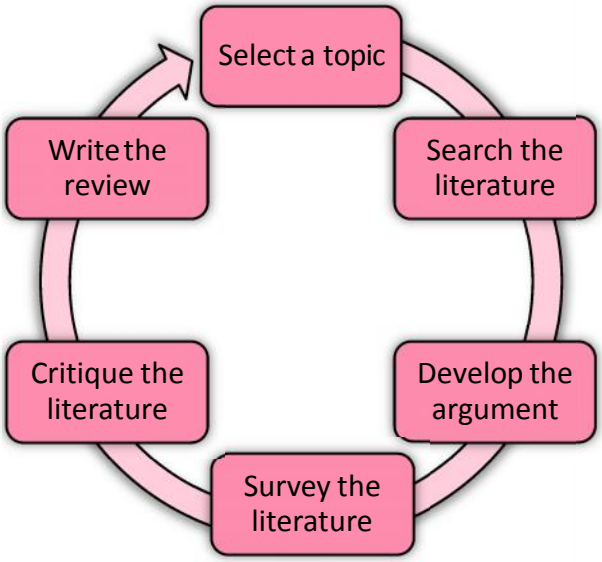
Harvard came originally from "The Bluebook: A Uniform System of Citation" published by the Harvard Law Review Association.The Harvard style and its many variations are used in law, natural sciences, social and behavioural sciences, and medicine.

Example: - Neville, C 2010, The complete guide to referencing and avoiding plagiarism, Open University Press, New York.

The researcher should select the referencing style with care as failing to give credits can prove fatal for a researcher and it is unethical as well.

Literature Review: Process

In course of literature review, researchers need to follow a long procedure. For example, researchers have to work out on the key terminologies that help to make the concept clear to the researchers themselves and the other readers. For this, they may have to use primary or secondary data sources. Similarly, they may have to strive hard to locate and evaluate them critically. Some of the sources may be useful and valid and other may not be. On top of this, writing them following certain format is a tedious procedure. For this, researchers need to follow numerous steps and sub-steps. These steps are the following:



Selection of Topic

- Read the research carefully.
- Talk over the idea with someone
- Scan academic journals
- Read professional blogs and listservs
- Look for Research Agendas on professional association websites
- Focus on the research topic

Search the Literature

- Types of sources that can be included: Books, Articles, Abstracts, Reviews, Dissertations and theses, Research reports, Websites.
- Identify the most important / useful databases for specificdiscipline.
- Develop an understanding of the academic terminology for your field of study and determine time frame.
- Look for empirical and theoretical literatureand also include primary and secondary sources.
- Identify important authors who are contributing to the development of the topic under research and use a system to organize and manage material.Example: Mendley, Refworks

Develop the Argument

A researcher will develop these arguments in the next two steps: surveying and critiquing the literature

Develop two types of arguments:

- Argument of discovery – develop findings that present the current state of knowledge about your research interest.
- Argument of advocacy – analyse and critique the knowledge gained from developing the argument of discovery to answer the research question.
- Analyse the claims within the literature to develop thearguments.
- Claim – the argument's declaration or assertion
- Evidence – data that define and support the claim

Types of Claims

- Fact
- Worth
- Policy
- Concept
- Interpretation

Survey and Critique the Literature

- Survey Develops the discovery argument and the advocacy argument.

Answers the questions:

- "What do we know about the subject of our study?"
- "Based on what we know, what conclusions can we draw about the research question?"
- A researcher should critically assess each piece of literature to analyse its content.
- A researcher need to be:
    - Methodical
    - Systematic
    - Rigorous
    - Consistent

Stage 1: Skim and Read

- Skim first – note topic, structure, general reasoning, data, and bibliographical references

- Go back and skim the preface and introduction, try to identify main ideas contained in the work
- Identify key parts of the article or key chapters in books

Stage 2: Highlight and Extract Key Elements

- Try to understand historical context and current state
- Identify themes, trends, patterns
- Also look for gaps and anomalies

Key questions related to literature:

- What are the origins and definitions of the topic?
- What are the key theories, concepts, and ideas?
- What are the major debates, arguments, and issues?

The key elements that all research studies should include:

- Problem
- Purpose
- Research questions
- Sample
- Methodology
- Key findings
- Conclusions
- Recommendations

Write the Review

- Use the results of analysis and critique of the literature to develop the organization of review
- Develop a detailed outline:
  - Identify the themes and/or patterns that have emerged
  - Translate these into headings and subheading
  - Reorganize and reassemble all of the separate pieces and details to create an integrated review.
  - Make connections between and among ideas and concepts.

3.3 Summary

A review of scholarly literature provides information that can be used to investigate a topic of importance to learn what is known about that topic for its own sake (i.e., to improve teaching or therapeutic practices) or as a basis for designing a research study. The formulation of a research topic is enabled by reading about research that has already been conducted because the reader can figure out what is already known as well as become acquainted with the strengths and weaknesses of methods used in prior research. Multiple sources exist for the conduct of literature reviews, including secondary sources that provide an overview of past research and primary sources that report original research. Primary sources can be identified through several different electronic means. A literature review is used to develop research questions of different types, such as descriptive, correlational, or interventionist. Researchers can also benefit by looking outside of published scholarly research to community members to provide a different perspective on what needs to be studied and how it should be studied.

3.4 Key Words

Literature Review: A literature review surveys books, scholarly articles, and any other sources relevant to a particular issue, area of research, or theory, and by so doing, provides a

description, summary, and critical evaluation of these works in relation to the research problem being investigated.

The purpose of a literature review: Place each work in the context of its contribution to understanding the research problem being studied and describe the relationship of each work to the others under consideration.

Types of Literature Reviews: Argumentative Review, Integrative Review, Historical Review, Methodological Review.

APA style of referencing: This style of referencing is known as "American Psychological Association" referencing.

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

3.5 Review Questions

1. What do you understand by Literature Review? Elaborate its importance.
2. Discuss the scope of literature review in context of business research.
3. As per your analysis, what are the advantages and disadvantages of primary data?
4. Is it necessary for the researcher to mention about the bibliographies and appendices?
5. Why/why not?
6. Illustrate advantages of online literature search with the help of examples
7. When planning to do social research, it is better to:

(a) Approach the topic with an open mind
(b) Do a pilot study before getting stuck into it
(c) Be familiar with the literature on the topic
(d) Forget about theory because this is a very practical undertaking can't have one without the other

8. The purpose of a literature review is to.

(a) Help you find out what is the research problem
(b) Identify the literature to collect data.
(c) Demonstrate an awareness of the theoretical context in which the current. study can be located
(d) Help to find out what tools can be applied for analysis

9. A literature review is

(a) Conducted after you have decided upon your research question
(b) Helps in the formulation of your research aim and research question
(c) Is the last thing to be written in your research report
(d) Is not part of a research proposal

10. Which is the most reliable source of information for your literature review?

(a) A TV documentary
(b) A newspaper article
(c) A peer reviewed research article
(d) A relevant chapter from a textbook

Fill in the blanks:

11. ……scholarly literature provides information that can be used to investigate a topic of importance to learn.
12. The discovery argument and the advocacy argument can be developed through……..

13. Chicago style of referencing is used mainly in the……
14. Scopus is an example of …….. database
15. Academic writing is backed up by ……..

Answers: Self-Assessment
7. (d) 8. (c) 9. (b) 10. (c) 11.Review 12.Survey 13.Social Sciences 14.Electronic 15. evidence

3.6 Further Readings

1. Cooper and Schinder, *Business Research Methods*, TMH.
2. CR Kotari, *Research Methodology*, Vishwa Prakashan.
3. David Luck and Ronald Rubin, *Marketing Research*, PHI.

1. Cohen, L. Manion, L. & Morrison, K (2011).Research method in education. London: Routledge.
2. Creswell, J.W. (2009). Research design. London: Sage.
3. Creswell, J.W. (2015). EducationalResearch design. New Jersey: Pearson.
4. Flick, U. (2012).Introducing research methodology. New Delhi: Sage.
5. Henn, M., Weinstein, M. & Foard, N. (2008).A short introduction to social research. New Delhi: Vistaar Publications.
6. https://uscupstate.libguides.com/c.php?g=627058&p=6601225.
7. https://libraryguides.nau.edu/c.php?g=665927&p=5074952#:~:text=Primary%20Sources,hypothesis%2C%20methodology%2C%20and%20results.&text=Survey%20research
8. https://www.igi-global.com/article/modeling-customers-intention-to-use-e-wallet-in-a-developing-nation/241249
9. https://library.leeds.ac.uk/info/14011/writing/106/academic_writing#:~:text=Show%20all%20contents-,Contents,long%20sentences%20and%20complicated%20vocabulary.

Unit 04: Types of Data in Research

CONTENTS

Objectives

After studying this unit, you will be able to:

- outline data types used in research.
- understand the meaning of primary and secondary data.
- explore the benefits and limitations of using primary and secondary data.
- formulate a research problem for an in-depth and more precise investigation.
- apply methods materials and techniques relevant to the solution of the problem.
- explain variables and their types in the context of research.
- infer research designs used in qualitative research.

Introduction

A business researcher has to tackle the problem of converting the management question into a research question. To do this, the researcher must have some information readily available before formally starting an experiment or research. This information is also important to understand different dimensions of a management problem. The readily available data sources also provide an opportunity to access other researcher's work that had similar kind of problems. This provides an opportunity to the researchers to develop their research problems in a more comprehensive manner. The available data sources are also important to identify the relevant variables to be included in the study and to frame the research questions properly. In the modern era, when computer and Internet facility are available everywhere, it is important for a researcher to be focused on the right source of data. It will help him or her to be concentrated on the concerned source and the research energy will not be devoured in searching an available unlimited source or more specifically web source. The quantity of the data will never be a problem for a researcher, but its added features of time and cost efficiencies will be a matter of concern. The chapter begins with the discussion on the difference between the primary and secondary data.

4.1 Meaning of Primary and Secondary Data

Primary data are mainly collected by a researcher to address the research problem. In other words, these are not readily available from various sources, rather the researcher has to systematically

collect the data relevant to a pre-specified research problem. **Secondary data** are the data that have already been collected by someone else before the current needs of a researcher. The present researcher only uses these data with related reference and never collects it from the field. When compared with the primary data, secondary data can be collected easily with time and cost efficiency. Both the primary and the secondary data have its own relative advantages and disadvantages. Although the secondary data are readily available and provide a base to tackle the research problem, the importance of the primary data is unquestionable. The arrangement of chapters in this book is mainly based on the primary data. The next section discusses some of the advantages and disadvantages of using secondary data.

Primary Data Collection Methods

Primary data collection methods are different ways in which primary data can be collected. It explains the tools used in collecting primary data, some of which are highlighted below:

Interviews

An interview is a method of data collection that involves two groups of people, where the first group is the interviewer (the researcher(s) asking questions and collecting data) and the interviewee (the subject or respondent that is being asked questions). The questions and responses during an interview may be oral or verbal as the case may be. Interviews can be carried out in 2 ways, namely; in-person interviews and telephonic interviews. An in-person interview requires an interviewer or a group of interviewers to ask questions from the interviewee in a face-to-face fashion. It can be direct or indirect, structured or structure, focused or unfocused, etc. Some of the tools used in carrying out in-person interviews include a notepad or recording device to take note of the conversation—very important due to human forgetful nature.

Telephonic interviews, on the other hand, are carried out over the phone through ordinary voice calls or video calls. The 2 parties involved may decide to use video calls like Skype to carry out interviews.A mobile phone, Laptop, Tablet, or desktop computer with an internet connection is required for this.

Pros

- In-depth information can be collected.
- Non-response and response bias can be detected.
- The samples can be controlled.

Cons

- It is more time-consuming.
- It is expensive.
- The interviewer may be biased.

Surveys & Questionnaires

Surveys and questionnaires are 2 similar tools used in collecting primary data. They are a group of questions typed or written down and sent to the sample of study to give responses. After giving the required responses, the survey is given back to the researcher to record. It is advisable to conduct a pilot study where the questionnaires are filled by experts and meant to assess the weakness of the questions or techniques used.

There are 2 main types of surveys used for data collection, namely; online and offline surveys. Online surveys are carried out using internet-enabled devices like mobile phones, PCs, Tablets, etc.They can be shared with respondents through email, websites, or social media. Offline surveys, on the other hand, do not require an internet connection for them to be carried out.The most common type of offline survey is paper-based surveys. However, there are also offline surveys that can be filled with a mobile device without access to an internet connection.

This kind of survey is called online-offline surveys because they can be filled offline but require an internet connection to be submitted.

Pros

- Respondents have adequate time to give responses.
- It is free from the bias of the interviewer.
- They are cheaper compared to interviews.

Cons

- A high rate of non-response bias.
- It is inflexible and can't be changed once sent.
- It is a slow process.

Observation

The observation method is mostly used in studies related to behavioral science. The researcher uses observation as a scientific tool and method of data collection. Observation as a data collection tool is usually systematically planned and subjected to checks and controls.There are different approaches to the observation method—structured or unstructured, controlled or uncontrolled, and participant, non-participant, or disguised approach.

The structured and unstructured approach is characterized by careful definition of subjects of observation, style of observer, conditions, and selection of data. An observation process that satisfies this is said to be structured and vice versa.A controlled and uncontrolled approach signifies whether the research took place in a natural setting or according to some pre-arranged plans. If an observation is done in a natural setting, it is uncontrolled but becomes controlled if done in a laboratory.Before employing a new teacher, academic institutions sometimes ask for a sample teaching class to test the teacher's ability. The evaluator joins the class and observes the teaching, making him or her a participant.

The evaluation may also decide to observe from outside the class, becoming a non-participant. An evaluator may also be asked to stay in class and disguise as a student, to carry out a disguised observation.

Pros

- The data is usually objective.
- Data is not affected by past or future events.

Cons

- The information is limited.
- It is expensive.

Focus Groups

Focus Groups are a gathering of 2 or more people with similar characteristics or who possess common traits. They seek open-ended thoughts and contributions from participants. A focus group is a primary source of data collection because the data is collected directly from the participant. It is commonly used for market research, where a group of market consumers engages in a discussion with a research moderator.

It is slightly similar to interviews, but this involves discussions and interactions rather than questions and answers. Focus groups are less formal and the participants are the ones who do most of the talking, with moderators there to oversee the process.

Pros

- It incurs a low cost compared to interviews. This is because the interviewer does not have to discuss with each participant individually.
- It takes lesser time too.

Cons

- Response bias is a problem in this case because a participant might be subjective to what people will think about sharing a sincere opinion.
- Group thinking does not clearly mirror individual opinions.

Experiments

An experiment is a structured study where the researchers attempt to understand the causes, effects, and processes involved in a particular process. This data collectionmethod is usually controlled by the researcher, who determines which subject is used, how they are grouped, and the treatment they receive.

During the first stage of the experiment, the researcher selects the subject which will be considered. Some actions are therefore carried out on these subjects, while the primary data consisting of the actions and reactions are recorded by the researcher.After which they will be analyzed, and a conclusion will be drawn from the result of the analysis. Although experiments can be used to collect different types of primary data, it is mostly used for data collection in the laboratory.

Pros

- It is usually objective since the data recorded are the results of a process.
- Non-response bias is eliminated.

Cons

- Incorrect data may be recorded due to human error.
- It is expensive.

Secondary data is the data that has already been collected through primary sources and made readily available for researchers to use for their own research. It is a type of data that has already been collected in the past.A researcher may have collected the data for a particular project, then made it available to be used by another researcher. The data may also have been collected for general use with no specific research purpose like in the case of the national census.

Data classified as secondary for particular research may be said to be primary for another research. This is the case when data is being reused, making it primary data for the first research and secondary data for the second research it is being used for.

Sources of Secondary Data

Sources of secondary data include books, personal sources, journals, newspapers, websites, government records, etc. Secondary data are known to be readily available compared to of primary data. It requires very little research and needs for manpower to use these sources. With the advent of electronic media and the internet, secondary data sources have become more easily accessible. Some of these sources are highlighted below.

Books

Books are one of the most traditional ways of collecting data. Today, there are books available for all topics you can think of. When carrying out research, all you have to do is look for a book on the topic being researched on, then select from the available repository of books in that area. Books,

when carefully chosen are an authentic source of authentic data and can be useful in preparing a literature review.

Published Sources

There are a variety of published sources available for different research topics. The authenticity of the data generated from these sources depends majorly on the writer and publishing company. Published sources may be printed or electronic as the case may be. They may be paid or free depending on the writer and publishing company's decision.

Unpublished Personal Sources

This may not be readily available and easily accessible compared to the published sources. They only become accessible if the researcher shares with another researcher who is not allowed to share it with a third party. For example, the product management team of an organization may need data on customer feedback to assess what customers think about their product and improvement suggestions. They will need to collect the data from the customer service department, which primarily collected the data to improve customer service.

Journal

Journals are gradually becoming more important than books these days when data collection is concerned. This is because journals are updated regularly with new publications periodically, therefore giving to date information. Also, journals are usually more specific when it comes to research. For example, we can have a journal on, "Secondary data collection for quantitative data" while a book will simply be titled, "Secondary data collection".

Newspapers

In most cases, the information passed through a newspaper is usually very reliable. Hence, making it one of the most authentic sources of collecting secondary data. The kind of data commonly shared in newspapers is usually more political, economic, and educational than scientific. Therefore, newspapers may not be the best source for scientific data collection.

Websites

The information shared on websites is mostly not regulated and as such may not be trusted compared to other sources. However, some regulated websites only share authentic data and can be trusted by researchers. Most of these websites are usually government websites or private organizations that are paid, data collectors.

Blogs

Blogs are one of the most common online sources for data and may even be less authentic than websites. These days, practically everyone owns a blog, and a lot of people use these blogs to drive traffic to their website or make money through paid ads. Therefore, they cannot always be trusted. For example, a blogger may write good things about a product because he or she was paid to do so by the manufacturer even though these things are not true.

Diaries

They are personal records and as such rarely used for data collection by researchers. Also, diaries are usually personal, except for these days when people now share public diaries containing specific events in their life. A common example of this is Anne Frank's diary which contained an accurate record of the Nazi wars.

Government Records

Government records are a very important and authentic source of secondary data. They contain information useful in marketing, management, humanities, and social science research. Some of these records include; census data, health records, education institute records, etc. They are usually collected to aid proper planning, allocation of funds, and prioritizing of projects.

Podcasts

Podcasts are gradually becoming very common these days, and a lot of people listen to them as an alternative to radio. They are more or less like online radio stations and are generating increasing popularity. Information is usually shared during podcasts, and listeners can use it as a source of data collection. Some other sources of data collection include:

- Letters
- Radio stations
- Public sector records.

4.2 Benefits and Limitations of Using Secondary Data

The main **advantage of using secondary data** sources is that they already exist; therefore, the time spent on the study is considerably less than that on studies that use the primary data collection. The **disadvantages of using secondary data** are related to the fact that their selection and quality, and the methods of their collection, are not under the control of the researcher and that they are sometimes impossible to validate (Sorensen et al., 1996). In some cases, the researchers find great difficulty in collecting the primary data. In such situations, the secondary data provide a base to tackle the problem. It is suggested that the secondary data not only offer advantages in terms of cost and effort, as conventionally described in the research method books, but in certain cases, their use may also overcome some of the difficulties that particularly afflict business researchers in gathering the primary data (Cowtown, 1998). There may be cases when the problem is general, such as the demographic structure of a population in a particular region, in such cases, there is no meaning in collecting the primary data. The various available secondary data sources such as the indiastat.com, the Centre for Monitoring Indian Economy (CMIE) products, and so on are capable of providing this information and are easily accessible. Similarly, sales, net sales, profit after tax, and much other information related to any company are easily accessible through a well-known data source of CMIE, commonly known as PROWESS. Hence, collecting the primary data for this purpose is meaningless.

Regarding disadvantages, the accuracy of secondary data is most of the time questionable as the researcher is unaware of the pattern of data collection. Besides, the researcher has no control over the data collection pattern. The researcher may try to use the secondary data that are developed for some other purpose in some other time frame in some other circumstances. This poses a great question mark on the currency and relevance of the data in terms of its use in the current problem. Moreover, the secondary data become outdated quickly. It is a big restriction on the frequent use of secondary data. For example, a consumer attitude study conducted 3 years ago may not be useful today because a lot of developments have taken place in these 3 years, which are not incorporated into the study conducted 3 years ago. The secondary data are not free from the limitations of the original research. Once a researcher decides to use a specific secondary database, he or she is subjected to the methods and limitations chosen by the original researchers, and therefore, the researcher who considers using a secondary database must know its limitations and potentials (Best, 1999).

Discuss the benefits of using primary data

4.3 Nature of Qualitative and Quantitative Research

Research methods are specific procedures for collecting and analyzing data. Developing your research methods is an integral part of your research design. When planning your methods, there

are two key decisions a researcher needs to take. First, aspect is to decide how to **collect data**. There are two types of research:

Qualitative research methods involve observations to gather non-numerical data. It is primarily explorative research. It is used to gain an understanding of underlying reasons, opinions, and motivations. It provides insights into the problem or helps to develop ideas or hypotheses for potential quantitative research. It is also used to uncover trends in thought and opinions, and dive deeper into the problem.

Quantitative methods emphasize objective measurements and the statistical, mathematical, or numerical analysis of data collected through polls, questionnaires, and surveys, or by manipulating pre-existing statistical data using computational techniques. Quantitative research focuses on gathering numerical data and generalizing it across groups of people or to predict or explain a particular phenomenon.

Qualitative Research

Qualitative research is used to understand how people experience the world. While there are many approaches to qualitative research, they tend to be flexible and focus on retaining rich meaning when interpreting data. Common approaches include grounded theory, ethnography, action research, phenomenological research, and narrative research. They share some similarities but emphasize different aims and perspectives.

Qualitative Research Methods

Each of the research approaches involves using one or more data collection methods. These are some of the most common qualitative methods:

- **Observations:** recording what you have seen, heard, or encountered in detailed field notes.
- **Interviews:** personally, asking people questions in one-on-one conversations.
- **Focus groups:** asking questions and generating discussion among a group of people.
- **Surveys:** distributing questionnaires with open-ended questions.
- **Secondary research:** collecting existing data in the form of texts, images, audio or video recordings, etc.

Characteristics of Qualitative Research

- Emergent - acceptance of adapting inquiry as understanding deepens and/or situations change; the researcher avoids rigid designs that eliminate responding to opportunities to pursue new paths of discovery as they emerge.
- Purposeful - study cases are selected because they are "information-rich" and illuminative. That is, they offer useful manifestations of the phenomenon of interest; sampling is aimed at insight about the phenomenon, not empirical generalization derived from a sample and applied to a population.
- Data -observations yield a detailed, quotations about people's perspectives and lived experiences; often derived from carefully conducted case studies and review of material culture.
- Personal experience and engagement - researcher has direct contact with and gets close to the people, situation, and phenomenon under investigation; the researcher's personal experiences and insights are an important part of the inquiry and critical to understanding the phenomenon.

Qualitative data analysis

Qualitative data can take the form of texts, photos, videos and audio. For example, you might be working with interview transcripts, survey responses, fieldnotes, or recordings from natural settings.

Most types of qualitative data analysis share the same five steps:

1. **Prepare and organize your data.** This may mean transcribing interviews or typing up fieldnotes.
2. **Review and explore your data.** Examine the data for patterns or repeated ideas that emerge.
3. **Develop a data coding system.** Based on your initial ideas, establish a set of codes that you can apply to categorize your data.
4. **Assign codes to the data.** For example, in qualitative survey analysis, this may mean going through each participant's responses and tagging them with codes in a spreadsheet. As you go through your data, you can create new codes to add to your system if necessary.
5. **Identify recurring themes.** Link codes together into cohesive, overarching themes.

Advantages of qualitative research

Qualitative research often tries to preserve the voice and perspective of participants and can be adjusted as new research questions arise. Qualitative research is good for:

Flexibility

The data collection and analysis process can be adapted as new ideas or patterns emerge. They are not rigidly decided beforehand.

Natural settings

Data collection occurs in real-world contexts or in naturalistic ways.

Meaningful insights

Detailed descriptions of people's experiences, feelings and perceptions can be used in designing, testing or improving systems or products.

Generation of new ideas

Open-ended responses mean that researchers can uncover novel problems or opportunities that they wouldn't have thought of otherwise.

Disadvantages of qualitative research

Researchers must consider practical and theoretical limitations in analyzing and interpreting their data. Qualitative research suffers from:

Unreliability

The real-world setting often makes qualitative research unreliable because of uncontrolled factors that affect the data.

Subjectivity

Due to the researcher's primary role in analyzing and interpreting data, qualitative research cannot be replicated. The researcher decides what is important and what is irrelevant in data analysis, so interpretations of the same data can vary greatly.

Limited generalizability

Small samples are often used to gather detailed data about specific contexts. Despite rigorous analysis procedures, it is difficult to draw generalizable conclusions because the data may be biased and unrepresentative of the wider population.

Labor-intensive

Although software can be used to manage and record large amounts of text, data analysis often has to be checked or performed manually.

Quantitative Research

Quantitative research deals in numbers, logic, and an objective stance. Quantitative research focuses on numeric and unchanging data and detailed, convergent reasoning rather than divergent reasoning [i.e., the generation of a variety of ideas about a research problem in a spontaneous, free-flowing manner].

Quantitative Research Methods

A researcher can use quantitative research methods for descriptive, correlational or experimental research**.**

- Descriptive research seeks an overall summary of the study variables.
- Correlational researchinvestigates relationships between study variables.
- Experimental research, a researcher systematically examines whether there is a cause-and-effect relationship between variables.

Correlational and experimental research can both be used to formally test hypotheses, or predictions, using statistics. The results may be generalized to broader populations based on the sampling method used. To collect quantitative data, you will often need to use operational definitions that translate abstract concepts (e.g., mood) into observable and quantifiable measures (e.g., self-ratings of feelings and energy levels).

Quantitative Data Analysis

Once data is collected, a researcher may need to process it before it can be analyzed. For example, survey and test data may need to be transformed from words to numbers. Then, statistical analysis can be used to answer your research questions.

Descriptive statistics will give you a summary of your data and include measures of averages and variability. You can also use graphs, scatter plots and frequency tables to visualize your data and check for any trends or outliers.

Using **inferential statistics**, you can make predictions or generalizations based on your data. You can test your hypothesis or use your sample data to estimate the population parameter.

Characteristics of Quantitative Research:

- The data is usually gathered using structured research instruments.
- The results are based on larger sample sizes that are representative of the population.
- The research study can usually be replicated or repeated, given its high reliability.
- Researcher has a clearly defined research question to which objective answers are sought.
- All aspects of the study are carefully designed before data is collected.
- Data are in the form of numbers and statistics, often arranged in tables, charts, figures, or other non-textual forms.
- Project can be used to generalize concepts more widely, predict future results, or investigate causal relationships.
- Researcher uses tools, such as questionnaires or computer software, to collect numerical data.

The goal of **quantitative research** methods is to collect numerical data from a group of people, then generalize those results to a larger group

Advantages of Quantitative Research

Quantitative research is often used to standardize data collection and generalize findings.
Strengths of this approach include:

Replication

Repeating the study is possible because of standardized data collection protocols and tangible definitions of abstract concepts.

Direct comparisons of results

The study can be reproduced in other cultural settings, times or with different groups of participants. Results can be compared statistically.

Large samples

Data from large samples can be processed and analyzed using reliable and consistent procedures through quantitative data analysis.

Hypothesis testing

Using formalized and established hypothesis testing procedures means that you have to carefully consider and report your research variables, predictions, data collection and testing methods before coming to a conclusion.

4.4 Disadvantages of Quantitative Research

Despite the benefits of quantitative research, it is sometimes inadequate in explaining complex research topics. Its limitations include:

Superficiality

Using precise and restrictive operational definitions may inadequately represent complex concepts. For example, the concept of mood may be represented with just a number in quantitative research but explained with elaboration in qualitative research.

Narrow focus

Predetermined variables and measurement procedures can mean that you ignore other relevant observations.

Structural bias

Despite standardized procedures, structural biases can still affect quantitative research. Missing data, imprecise measurements or inappropriate sampling methods are biases that can lead to the wrong conclusions.

Lack of context

Quantitative research often uses unnatural settings like laboratories or fails to consider historical and cultural contexts that may affect data collection and results.

4.5 Data and Variables Used in Qualitative and Quantitative Methods

Data refers to distinct pieces of information, usually formatted and stored in a way that is in accordance with a specific purpose.A researcher needs to understand and correctly apply statistical measurements to data and should therefore correctly conclude certain assumptions about it. The diagram mentioned below reflects data types.

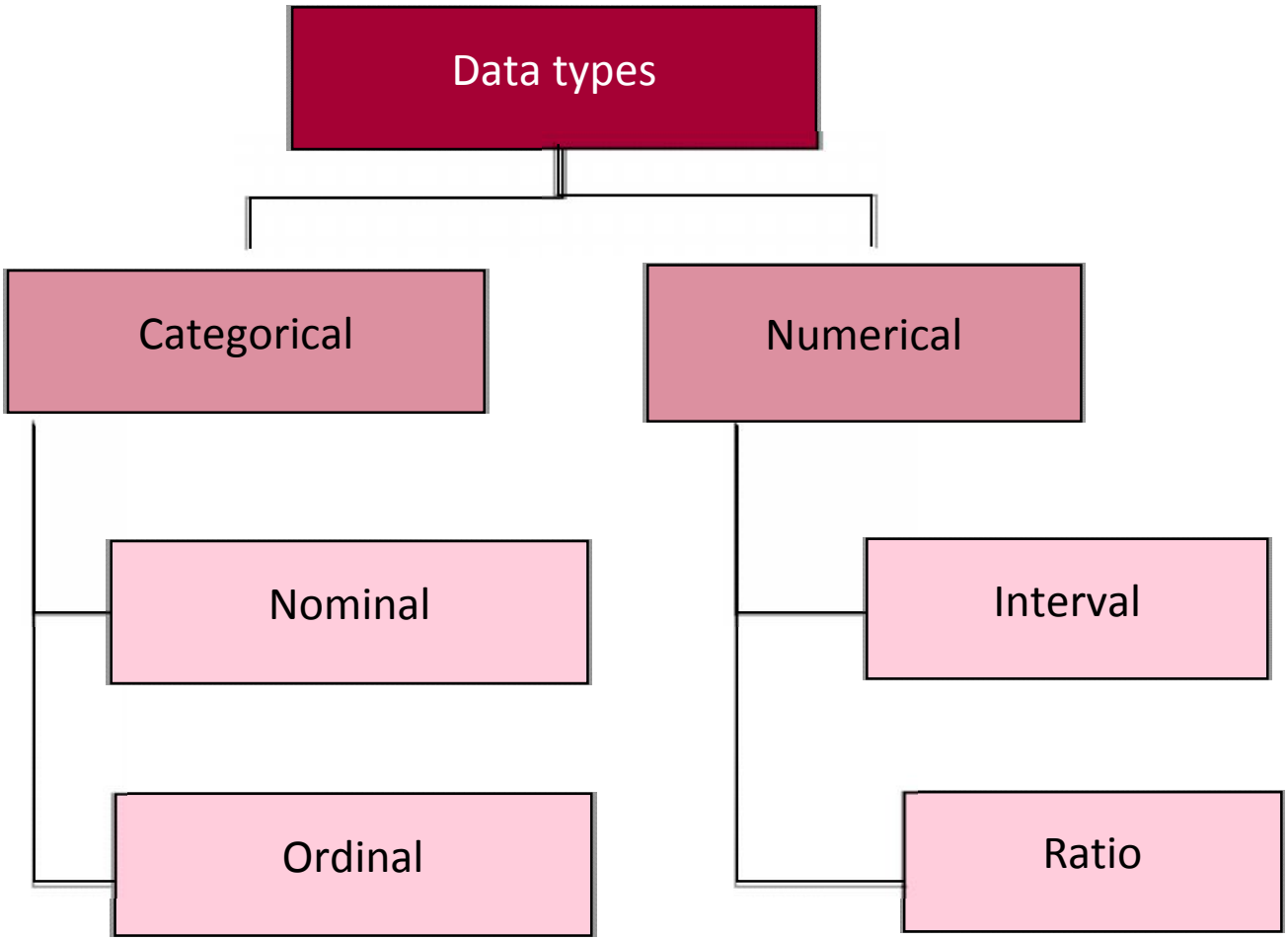


Figure 4.1: Data Types

Categorical Data

Categorical data represents characteristics. Therefore, it can represent things like a person's gender, language etc. Categorical data can also take on numerical values (Example: 1 for female and 0 for male).

Nominal Data

Nominal values represent discrete units and are used to label variables, that have no quantitative value. Nominal data that has no order. If a researcher changes the order of its values, the meaning will not change.

Ordinal Data

Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters.The main limitation of ordinal data is the differences between the values is not really known. Ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction etc.

Numerical Data

This type of data can't be measured but it can be counted. It basically represents information that can be categorized into a classification. Example: Number of heads in 100-coin flips.

Continuous Data

Continuous Data represents measurements and therefore their values can't be counted but they can be measured. Example: Height of a person

Interval Data

Interval values represent ordered units that have the same difference. Therefore, interval data represents numeric values that are ordered and where differences between the values are known exactly.Ratio values are also ordered units that have the same difference. Ratio values are the same as interval values, with the difference that they do have an absolute zero. Good examples are height, weight, length etc.

Ratio Data

Ratio values are also ordered units that have the same difference. Ratio values are the same as interval values, with the difference that they do have an absolute zero. Good examples are height, weight, length etc.

Variable

A variable is any property, a characteristic, a number, or a quantity that increases or decreases over time or can take on different values (as opposed to constants, such as n, that do not vary) in different situations.

Independent Variable

The variable that is used to describe or measure the factor that is assumed to cause or at least to influence the problem or outcome is called an independent variable. The definition implies that the experimenter uses the independent variable to describe or explain the influence or effect of it on the dependent variable. Variability in the dependent variable is presumed to depend on variability in the independent variable.

Dependent Variable

The variable that is used to describe or measure the problem or outcome under study is called a dependent variable.In a causal relationship, the cause is the independent variable, and the effect is the dependent variable. If we hypothesize that smoking causes lung cancer, 'smoking' is the independent variable and cancer the dependent variable.

Background Variable

In almost every study, researchers collect information such as age, sex, educational attainment, socioeconomic status, marital status, religion, place of birth, and the like. These variables are referred to as background variables.

Moderating Variable

In any statement of relationships of variables, it is normally hypothesized that in some way, the independent variable 'causes' the dependent variable to occur. In simple relationships, all other variables are extraneous and are ignored. In actual study situations, such a simple one-to-one relationship needs to be revised to take other variables into account to better explain the relationship. This emphasizes the need to consider a second independent variable that is expected to have a significant contributory or contingent effect on the originally stated dependent-independent relationship. Such a variable is termed as a moderating variable.

Extraneous Variable

Several variables might conceivably affect the hypothesized independent-dependent variable relationship, thereby distorting the study. These variables are referred to as extraneous variables.

4.6 Writing up Qualitative Research

Qualitative Research

Researchers more comfortable with quantitative research. Quantitative methods deal with the collection and processing numerical data.

Qualitative research answer questions:

- How often? To what extent?
- How much? How many … but cannot answer questions on: Why? how? In what way?

Qualitative Research Designs

Four major types of qualitative research design include:

1. Phenomenology
2. Ethnography
3. Grounded theory
4. Case study

Phenomenology is used to identify phenomena and focus on subjective experiences and understanding the structure of those lived experiences. It was founded in the early 20th century by Edmund Husserl and Martin Heidegger and originated from philosophy. Phenomenology is used to describe, in depth, the common characteristics of the phenomena that has occurred. The primary data collection method is through in-depth interviews.

Ethnographic studies are qualitative procedures utilized to describe, analyze and interpret a culture's characteristics. Ethnography was developed in the 19thand 20th centuries and used by anthropologists to explore primitive cultures different from their own; it originated from Anthropology. Ethnography is used when a researcher wants to study a group of people to gain a larger understanding of their lives or specific aspects of their lives. The primary data collection method is through observation over an extended period of time. It would also be appropriate to interview others who have studied the same cultures.

Grounded theory is a systematic procedure of data analysis, typically associated with qualitative research, that allows researchers to develop a theory that explains a specific phenomenon. Grounded theory was developed by Glaser and Strauss and is used to conceptualize phenomenon using research; grounded theory is not seen as a descriptive method and originates from sociology. The unit of analysis in grounded theory is a specific phenomenon or incident, not individual behaviors. The primary data collection method is through interviews of approximately 20 – 30 participants or until data achieves saturation.

Case studies are believed to have originated in 1829 by Frederic Le Play. Case studies are rooted in several disciplines, including science, education, medicine, and law. Case studies are to be used when (1) the researcher wants to focus on how and why, (2) the behavior is to be observed, not manipulated, (3) to further understand a given phenomenon, and (4) if the boundaries between the context and phenomena are not clear. Multiple methods can be used to gather data, including interviews, observation, and historical documentation.

Summary

The researcher who decides to use such a scale in their study has to make another set of judgments: how well does the scale measure the intended concept; how reliable or consistent is it; how appropriate is it for the research context and intended respondents; and on and on. Believe it or not, even the respondents make many judgments when filling out such a scale: what is meant by various terms and phrases; why is the researcher giving this scale to them; how much energy and effort do they want to expend to complete it, and so on. Even the consumers and readers of the research will make lots of judgments about the self-esteem measure and its appropriateness in that research context. What may look like a simple, straightforward, cut-and-dried quantitative measure is based on lots of qualitative judgments made by lots of different people.

Keywords

Quantitative data can be counted, measured, and expressed using numbers. *Qualitative data* is descriptive and conceptual. Qualitative data can be categorized based on traits and characteristics.

An experiment is a structured study where the researchers attempt to understand the causes, effects, and processes involved in a particular process.

Data refers to distinct pieces of information, usually formatted and stored in a way that is in accordance with a specific purpose.

Secondary research involves collecting existing data in the form of texts, images, audio or video recordings, etc

SelfAssessment

1. Secondary/existing data may include which of the following?

A. Official documents
B. Personal documents
C. Archived research data
D. All of the above

2. Which of the following terms best describes data that were originally collected at an earlier time by a different person for a different purpose?

A. Primary data
B. Secondary data
C. Experimental data
D. Field notes

3. Which of the following is not a major method of data collection?

A. Questionnaires
B. Focus groups
C. Correlational method
D. Secondary data

4. Questionnaire is a _____
A. Research method
B. Measurement technique
C. Tool for data collection
D. Data analysis technique

5. Website is an example of ----------
A. Primary data source
B. Secondary data source
C. Hybrid data source
D. None of these

6. _____ are the basic building blocks of qualitative data.

A. Categories
B. Units
C. Individuals
D. None of the above

7. In qualitative research you talk to more people than in quantitative research is this statement:

A. True
B. False

8. Quantitative studies emphasize the measurement and analysis of causal relationships between variables, but not on…….

A. Results
B. Process
C. Introduction
D. Context

9. Which of the following is NOT one of Research Designs for Qualitative Studies?

A. Goals
B. Conceptual Framework
C. Survey
D. Conclusion

10. Which of the following is a qualitative research technique?

A. Observation
B. Experimentation
C. Postal questionnaire
D. Focus group

11. _____ are the basic building blocks of qualitative data.

A. Categories
B. Units
C. Individuals
D. None of the above

12. In qualitative research you talk to more people than in quantitative research is this statement:

A. True
B. False

13. Quantitative studies emphasize the measurement and analysis of causal relationships between variables, but not on…….

A. Results
B. Process
C. Introduction
D. Context

14. Which of the following is NOT one of Research Designs for Qualitative Studies?

A. Goals
B. Conceptual Framework
C. Survey
D. Conclusion

15. Which of the following is a qualitative research technique?

A. Observation
B. Experimentation
C. Postal questionnaire
D. Focus group

16. _____ are the basic building blocks of qualitative data.

A. Categories
B. Units
C. Individuals
D. None of the above

17. In qualitative research you talk to more people than in quantitative research is this statement:
A. True
B. False

18. Quantitative studies emphasize the measurement and analysis of causal relationships between variables, but not on…….
A. Results
B. Process
C. Introduction
D. Context

19. Which of the following is NOT one of Research Designs for Qualitative Studies?

A. Goals
B. Conceptual Framework
C. Survey
D. Conclusion

20. Which of the following is a qualitative research technique?

A. Observation
B. Experimentation
C. Postal questionnaire
D. Focus group

Answer for Self Assessment

1.	D	2.	B	3.	C	4.	C	5.	B
6.	A	7.	B	8.	B	9.	C	10.	D
11.	A	12.	B	13.	B	14.	C	15.	D

16. A 17. B 18. B 19. C 20. D

Review Questions

1. What is Qualitative research? How it is different from Quantitative research
2. Explain the data types used in Qualitative and Quantitative research
3. What is secondary data? Highlight the advantages of using secondary data in research.
4. What are advantages of Qualitative research? Explain in detail.

Further Reading

- Business Research Methods by Naval Bajpai, Pearson
- Research Methodology: Methods and Techniques by Kothari, C. R. & Garg, Gaurav, New Age International.
- Marketing Research by Naresh K Malhotra, Pearson

Online Link

https://library.fiu.edu/researchmethods/datatypes
https://www.djsresearch.co.uk/glossary/item/Qualitative-Research-Design
https://neltaeltforum.wordpress.com/2018/02/06/preview-of-literature-review-as-a-corner-stonresearch/.

Unit 05: Sampling Design

CONTENTS

Objectives

After studying this unit, you will be able to:

- Decide how to collect sample data
- Gain insights into terminologies used in sampling
- Overview steps involved in sampling design process
- 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

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.

5.1 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

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.

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

Example: Reactions to global advertising by a company.

5.2 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.

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 you 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.

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

5.3 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.

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 beingunderstood 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.

Task: Carefully balance the conflicting criteria to select a good sample design.

5.4 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.

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.

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.

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

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.

Non-probability Sampling Techniques

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

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.

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

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.

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.

5.5 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 you 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.

5.6 Errors in Sampling

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 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.

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.

Sampling Frame Error

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

Example:

1. A 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.

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:

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.

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.

5.7 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.

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 dispatch 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.

5.8 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.

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.
- 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.

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.

Self Assessment

1. Sample is regarded as a subset of?

A. Data
B. Set
C. Distribution
D. Population

2. A statistical investigation in which the data are collected for each and every element/unit of the population, it is termed as

A. Census
B. Distribution
C. Population
D. Subset

3. A population includes ……from a specified group, all possible outcomes or measurements that are of interest.

A. All members
B. Few members
C. Proportionate members
D. None of these

4. Element represents

A. No Unit
B. One Unit
C. Multiple Units
D. None of these

5. ……..involves unequal chance of being included in the sample

A. Non-probability sampling
B. Probability sampling
C. Moving sampling
D. Unequal sampling

6. Which of the following is NOT an advantage of sampling?

A. Accuracy & quality control
B. More data collection
C. Economy in terms of cost
D. Economy in terms of time

7. A sample should be of ……..which fulfils the research requirements.

A. Optimum Size
B. Larger Size
C. Small Size
D. Medium Size

8. Factors affecting size of sample includes:

A. Method of data used
B. Ability of researcher
C. Knowledge of research

D. Previous research

9. Characteristics of a Good Sample can be determined through

A. Non-random selection
B. Random selection
C. Data
D. Researcher

10. Selection Error is an example of:

A. Non sampling error
B. Common sampling error
C. Predictive error
D. Poor sampling error

11. Which of the following is not a type of non-probability sampling?

A. Quota sampling
B. Convenience sampling
C. Snowball sampling
D. Stratified random sampling

12. Among these, which sampling is based on equal probability?

A. Simple random sampling
B. Stratified random sampling
C. Systematic sampling
D. Probability sampling

13. In random sampling, the probability of selecting an item from the population is:

A. Unknown
B. Known
C. Un-decided
D. One

14. Sample is a sub-set of:

A. Population
B. Data
C. Set
D. Distribution

15. Increasing the sample size has the following effect upon the sampling error?

A. It increases the sampling error
B. It reduces the sampling error
C. It has no effect on the sampling error
D. All of the above

Answers for Self Assessment

1.	D	2.	A	3.	A	4.	B	5.	A
6.	B	7.	A	8.	A	9.	B	10.	B
11.	D	12.	A	13.	B	14.	A	15.	B

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. Quota sampling does not require prior knowledge about the cell to which each populationunit belongs. Does this attribute serve as an advantage or disadvantage for QuotaSampling?
4. What suggestions would you give to reduce non sampling error?
5. One mobile phone user is asked to recruit another mobile phone user. What samplingmethod is this known as and why?
6. Sampling is a part of the population. True/False? Why/why not?
7. What do see as the reason behind purposive sampling being known as judgementsampling?

Further Readings

1. Cooper and Schinder, *Business Research Methods*, TMH.
2. CR Kotari, *Research Methodology*, Vishwa Prakashan.
3. David Luck and Ronald Rubin, *Marketing Research*, PHI.
4. Naresh Amphora, *Marketing Research*, Pearson Education.
5. S.N. Murthy & U. Bhojanna, *Business Research Methods*, 3rd Edition, Excel Books.
6. William Zikmund, *Business Research Methods*, Thomson.

Unit 06:Measurement and Scaling Technique

CONTENTS

Objectives

After studying this unit, you will be able to:

- To know various sound measurement tools
- Explain the techniques of developing measurement tools
- Describe the meaning of scaling and its techniques
- Distinguish among Comparative and non-comparative scales
- Describe the strategies of multi-dimensional scaling

Introduction

According to given principles, measurement is the process of assigning numbers or other symbols to the properties of the things being measured. A concept (or construct) is a broad concept that describes a group of things, qualities, events, or processes. Age, gender, number of children, education, and income are examples of rather definite constructs. Brand loyalty, personality, channel power, and satisfaction are all factors that are considered in relatively abstract formulations.

The creation of a continuum on which measured things are placed is known as scaling. A scale is a measuring device that consists of a collection of elements that are arranged in ascending order of value or magnitude.

6.1 Scales of Measurement: Tools for Sound Measurement

These are of four kinds of scales, such as :

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

Nominal Scale

Numbers are used to identify the things on this scale. Students' university registration numbers, for example, are issued to them, as are the numbers on their jerseys. In this kind of scaling, the objective of marking numbers, symbols, labels, and other symbols is not to establish an order, but rather to simply place labels to identify activities and count the objects and subjects. Individuals, corporations, products, brands, and other entities are classified into categories using this measurement scale, which has no suggested order. It's referred to as a categorical scale rather frequently. It is a classification system, not a continuum, in which the entity is placed. It requires a basic count of the number of cases given to each category, and if desired, numbers may be nominally assigned to each category to label it.

Characteristics

1. It does not have an arithmetic origin.
2. It shows no relationship in terms of order or distance.
3. It categorises objects and groups them accordingly.

Use: This scale is commonly used to perform surveys and ex-post-facto research.

Have you ever been to Bangalore?

Yes-1 No-2

'Yes' is coded as 'One' and 'No' is coded as 'Two'. The numeric value assigned to the responses has no relevance and serves just as a means of identification. The answers supplied by respondents will not be affected if the numbers are changed to one for 'No' and two for 'Yes.' The numbers used in nominal scales are solely for counting purposes. The telephone numbers are an example of nominal scale, as each number corresponds to a single subscriber. The purpose of employing a nominal scale is to ensure that no two people or objects receive the same numerical value. Bus route numbers, for instance, are an example of nominal scale.

"How old are you"? It is an example of the scale called nominal.

we use nominal scale in the cases like "What is the ID number of your Card? books arrangement in the library- subject wise, author wise.

It is not to be forgotten that nominal scale has definite limitation, namely.

1) There is no ranking system.

2) There are no mathematical operations that can be performed.

3) Statistical implications - The standard deviation and mean cannot be calculated. However,the mode can be used in several ways.

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. In market research, we often

ask the respondents to rank the items, like for example, "A soft drink, based upon flavour or color". 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 Mobile Phone:

Attributes	Item
i) Company Image	3
ii) Functions	2
iii) Price	1
iv) Comfort	4
v) Design	5

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

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.

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.

Interval Scale

Nominal and ordinal scales are less powerful than interval scales. On the object being measured, the distance shown on the scale indicates an equal distance. The interval scale can inform us "how far apart the items are in relation to an attribute." This indicates that the differences are comparable. The difference between the numbers "1" and "2" is the same as the difference between the numbers "2" and "3."

The idea of "equality of interval" is employed in the interval scale, which means that the intervals are used as the basis for making the units equal, assuming that the intervals are equal.

Researchers can only justify using the arithmetic mean as a measure of average when the data is interval scaled. The interval or cardinal scale uses the same units of measurement as the ordinal scale, allowing you to understand not only the order of the scale scores but also the distance between them. The zero point on an interval scale, however, must be understood as arbitrary and not a true zero. Of course, this has consequences for the kind of data manipulation and analysis we may perform on data obtained in this manner. A constant can be added or subtracted from all of the scale values without changing the scale's shape, but the values cannot be multiplied or divided. Two respondents with scale positions 1 and 2 are as far off as two respondents with scale positions 4 and 5, however a person with a score of 10 does not feel twice as intensely as someone with a score of 5. Temperature is measured in Centigrade or Fahrenheit on an interval scale. Because the respective temperatures on the centigrade scale, 100°C and -3.9°C, are not in the ratio 2:1, we cannot say that 50°F is twice as hot as 25°F.

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: The majority of common statistical methods of analysis just require interval scales in order to be applied. These aren't discussed here because they're so widespread and can be found in almost every introductory statistics textbook.

In case, we would like to measure the refrigerator rating by using interval scale, It would look as follows:

(a) Brand Name	Poor …… Good
(b) Price	High …….. Low
(c) Service after-sales	Poor …… Good
(d) Utility	Poor ……..Good

The researcher cannot assume that a respondent who gives a rating of 6 is three times more positive about a product under investigation than a responder who gives a rating of 2.

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

Examine the distinctions between ordinal and interval scales..

Ratio Scale

A ratio scale is a type of interval scale with a significant zero point. This scale can be used to measure length, weight, or distance. It is possible to describe how many times larger or smaller one object is when compared to another using this scale. Actual variables are measured using these scales. A ratio scale is the highest degree of measurement. This scale combines the characteristics of an interval scale with the addition of a fixed origin or zero point. Weights, lengths, and times are all examples of ratio scaled variables. Ratio scales allow researchers to compare difference in scores as well as the relative magnitude of those discrepancies. The difference between 5 and 10 minutes, for example, is the same as the difference between 10 and 15 minutes, because 10 minutes is twice as long as 5. Given that sociology and managerial research rarely goes beyond the interval level of measurement, it is not recommended that this level of analysis be given extra attention. To summarize, ratio scales can be used to execute almost all statistical operations.

Characteristics

1. This scale has a measurement of absolute zero.
2. Geometric and harmonic techniques are utilised to determine central tendency.

Use: All statistical approaches can make use of the ratio scale.

For Instance, this year's sales of product A are twice as high as last year's sales of the identical product.

Statistical implications: This scale allows for the execution of all statistical operations.

6.2 Techniques for Developing Measurement Tools

Scale constructing procedures are used to assess a group's or an individual's attitude. In other words, the scale building technique aids in estimating an individual's or a group's interest or behaviour toward another or another's environment rather than one's own. When using a scale building technique, you must think about a number of things, including the individual's or group's mentality. Several of these choices are difficult to make.:

establishing the level of related data;

- Determining whether the given value is nominal, ordinal, interval, or ratio.
- Identifying the statistical analysis that will be useful in the scale designing.
- Choosing the appropriate scale design approach.
- Choosing the physical layout of the scales.
- Choosing the appropriate scale categories to apply.

The following are examples of different types of comparison techniques:

1. Pair wise Comparison Scale: This is an ordinal level scale design technique in which a respondent is given two options and asked to choose one.

2. Rasch Model Scale: In this technique, numerous respondents are involved with several things at the same time, and comparisons are drawn from their responses to determine the scale values. Rate-order scale: This is another ordinal level scale construction technique in which a respondent is given many items to rank.

3. Constant Sum Scale: A respondent is frequently given a fixed sum of money, credits, or points to allocate to various objects in order to determine the scale values of the items in this scale building technique.

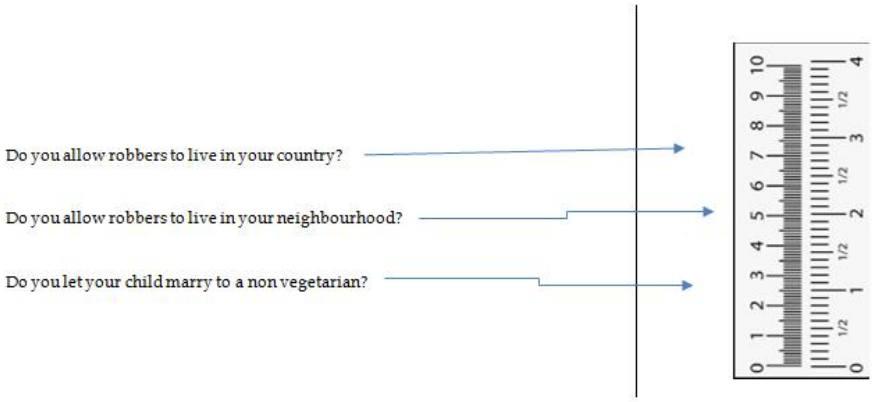
The various types of non-comparative techniques are:

1. **Continuous rating scale**: In this technique, respondents rate an item using a sequence of numbers called scale points. Graphic rating scaling is another name for this method.
2. **Likert scale:** This method allows respondents to rate items on a five-to-seven-point scale based on their level of agreement or disagreement with the item.
3. **Semantic differential scale:** Respondents are asked to rate the distinct qualities of an item on a seven-point scale in this technique.

6.3 Scaling – What does it mean

Scaling is a procedure or series of procedures for determining an individual's attitude. The assigning of objects to numbers according to a rule is known as scaling. Text statements, which can be expressions of attitude or principle, are the objects in the definition. Scaling does not directly measure an individual's attitude. It is initially migrated to statements, after which the numbers are assigned. The diagram below illustrates how to rate people's attitudes on a scale of one to ten.

Figure: 5.1



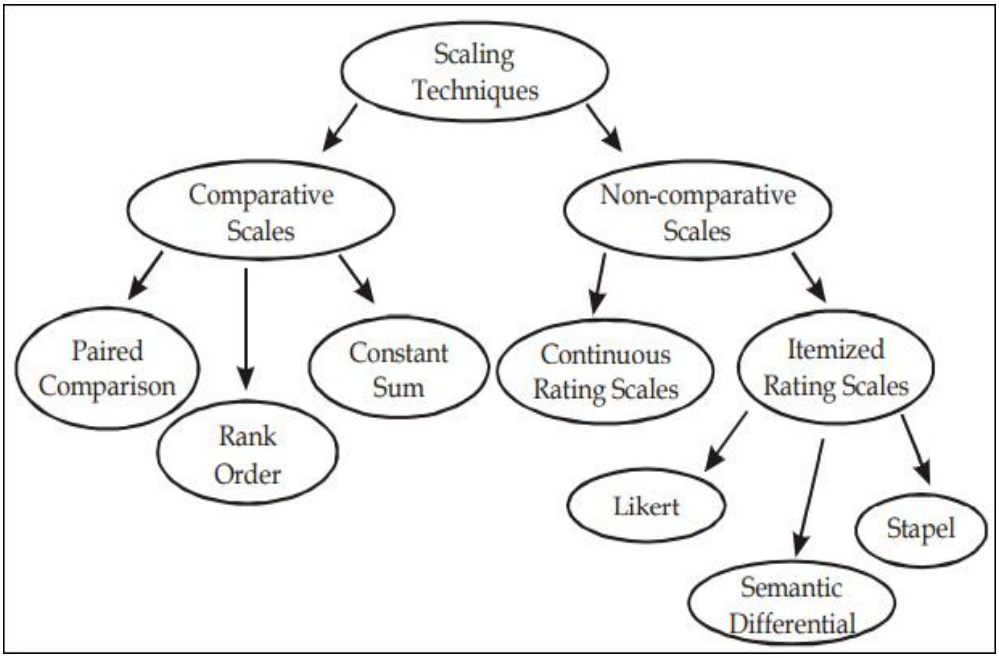
We will measure an individual's attitude by analysing his ideas, in the figure above, about drinkers. As you get further down, you'll see that people's attitudes and behaviours toward drinkers get increasingly erratic. If a person agrees with one of the statements in the list, he is more likely to agree with all of the assertions that follow it. As a result, in this case, the rule is growing one. This is referred to as scaling. Scaling is used to test a hypothesis during the research process. Scaling can be used as part of probing inquiry in some cases.

6.4 Comparative and Non-comparative Scaling Techniques

1. Comparative Scales: It directly compares two or more objects.

2. Non-comparative Scales: Objects are scaled independent of each other.

Figure 6.2: Classifying Scaling Techniques



6.5 Comparative Scaling Techniques

Paired Comparison

In this , a respondent is asked to choose between five different brands of coffee – A, B, C, D, and E – in terms of flavour. In pairs, he must convey his preference. The following formula is used to calculate the number of pairings. The brands to be scored are shown two at a time, allowing each brand in the category to be compared to every other brand in the category once. The respondents were asked to distribute 100 points between each pair based on how much they liked one over the other. Each brand receives a separate score.

Number of Pairs $=N(N-1)/2$

In this case, it is $= 5(5-1)/2$

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

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

	Ram	Shyam	Mohan	Ishwar
Ram		0	0	0
Shyam	1		1	0
Mohan	1	0		0
Ishwar	1	1	1	0
# Number of time preferrred	3	1	2	0

Rank Order Scaling

- Respondents are shown many objects at the same time and
- then Respondents are shown many objects at the same time
- The information gathered is ordinal in nature.
- In order of magnitude, they are arranged or ordered.
- Frequently used to assess brand preferences and brand features.

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
Ram	1
Shyam	3
Mohan	2
Ishawar	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 Units allocated represent the importance attached to the objects
2. Data obtained are interval in nature
3. Units allocated represent the importance attached to the objects
4. Allows for fine discrimination among alternatives

Four marketing professors are listed below, along with three features that students often value. Please assign a number to each component that indicates how well you believe each instructor performs in that area. The higher the number, the better the score. The sum of all of the teachers' scores on a certain aspect should be 100.

Instructor	Availability	Fairness	Easy Tests
Ram	25	35	25
Shyam	35	25	25
Mohan	25	25	25
Ishwar	15	15	25
Sum Total	100	100	100

Non-comparative Scale

Continuous Rating Scale VERY POOR …….....VERY GOOD 10 20 30 40 50 60 70 80 90 100

Likert Scale

It is also known as summated rating scale. This is a collection of statements about an attitude object. On the scale of '5 points, Agree, and Disagree,' each statement has 5 points, Agree, and Disagree. Because the scores of different items are added together to generate a total score for the respondent, they are also known as summated scales. The Likert Scale is divided into two parts: item and evaluation. A comment about a specific product, event, or attitude is frequently included in the item component. The evaluation section consists of a series of replies ranging from "strongly agree" to "strongly disagree." In this case, a five-point scale is used. Numerals such as +2, +1, 0, –1, –2 are utilized. Let's look at how a customer's attitude toward a shopping mall is measured using an example.

SL 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	-	-	-	-	-

The overall attitude of the respondents is assessed by adding up his or her numerical ratings on the statements that make up the scale. Because some sentences are positive and others are negative, this is the most important task to complete before calculating the ratings. In other words, a "strongly agree" category is associated with a positive statement, while a "strongly disagree" category is associated with a negative statement. The statement must be given the same amount every time, such as +2 or –2. "How successfully are the statements generated?" determines the Likert Scale's success. The more favourable the attitude, the higher the respondent's score. For example, if there are two shopping malls, ABC and XYZ, and the Likert Scale scores are 30 and 60, we can conclude that people prefer XYZ to ABC.

There must be an equal number of positive and negative statements on the Likert Scale.

Semantic Differential Scale

This is very much like the Likert Scale. It also includes a number of items for respondents to rate. The following is the key distinction between Likert and Semantic Differential Scale: It employs the term "bipolar" in its adjectives and phrases. The Semantic Differential Scale has no statements. A seven-point scale separates each pair of adjectives.

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

Respondents were asked to select one of seven categories that best described their attitudes. The calculation is carried out in the same manner as in the Likert Scale. Assume we're attempting to assess the packing of a specific product. The following is a seven-point scale:

"I feel ……..

1. Delighted
2. Pleased
3. Mostly satisfied
4. Equally satisfied and dissatisfied
5. Mostly dissatisfied
6. Unhappy
7. Terrible.

Thurstone Scale

This is also called as an equal appearing interval scale. The following are the steps to construct a Thurstone Scale:

Step 1: Generate a big number of statements about the attitude to be assessed.

Step 2: A group of judges, say 20 to 30, is given these statements (75 to 100) and asked to classify them according to the degree of favorability and unfavorability.

Step 3: The judges must create 11 piles. The statements in the heaps range from "most unfavourable" in pile 1 to "neutral" in pile 6 and "most favourable" in pile 11.

Step 4: Analyze the frequency distribution of ratings for each statement and delete any statements with highly disparate ratings from various judges.

Step 5: For the final scale, choose one or two statements from each of the 11 piles. To make the scale, arrange the statements in a random order.

Step 6: Those whose attitudes were to be scaled were given a set of items and asked to indicate whether they agreed or disagreed with each one. Some people may agree with just one assertion, while others may agree with multiple statements.

Suppose we're interested in the attitudes of respondents from a specific socioeconomic group on savings and investments. The following would be the final set of statements:

1) It is more important to live in the present than in the future. As a result, there is no need to savings.

2) There are numerous attractions where you can spend the money you have saved.

3) It is preferable to spend savings rather than invest them.

4) Investments are risky because the funds are also frozen.

5) You earn money to spend, not to save.

6) There is no way to save these days.

7) A portion of one's salary should be set aside and invested.

8) (The future is uncertain, and we shall be protected by our investments.)

9) Every person should have a certain amount of savings and investments.

10) One should make an effort to save more money so that the majority of it can be invested.

11) All savings should be put into a long-term investment.

Conclusion: A respondent who agrees with points 8, 9, and 11 is said to have a positive attitude about saving and investing. The person who agrees with statements 2, 3, and 4 is someone who has a negative attitude. Furthermore, a respondent's attitude is not considered consistent if he chooses statements 1, 3, 7, or 9.

Multidimensional Scaling

This is used to research customer attitudes, namely perceptions and preferences. These methods aid in determining which product qualities are most significant to customers and

determining their relative importance. Multi-Dimensional Scaling can be used to investigate the following topics:

1. What are the most important characteristics to consider when selecting a product (soft drinks, modes of transportation)? (a) What characteristics do customers compare while evaluating different product brands? Is it a matter of cost, quality, or availability, for example?

2. According to the customer, what is the appropriate combination of attributes? (That is, which two or more attributes will a consumer consider before making a purchase decision.)

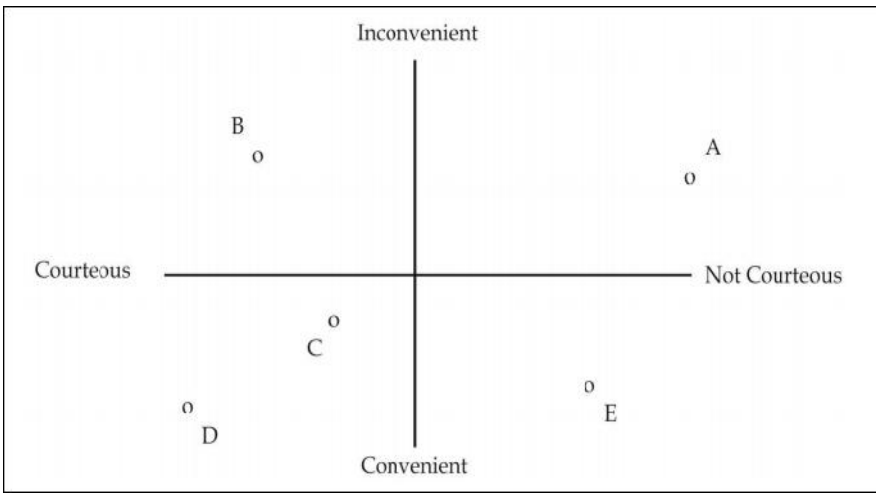
3. Which advertising messages are in line with the consumer's impressions of the brand?

Multidimensional scaling is a method of describing brand resemblance and preference. Respondents were asked to rate the similarity of numerous items (products, brands, etc.) as well as their preference for different objects. Perceptual mapping is another name for this scaling.

There are two methods for gathering input data for perceptual mapping:

1. Non-attribute method: In this method, the respondent is asked to make a direct judgement about the objects. The respondent determines the criteria for comparing the objects in this method.
2. Attribute method: In this method, Instead of choosing the criterion, respondents were asked to compare the objects based on the criteria supplied by the researcher in this approach.

Figure 6.3



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.

Stapel Scales

1. While creating pairs of bipolar adjectives is problematic, modern versions of the Stapel scale use a single adjective to replace the semantic difference.

2. A Stapel scale's benefits and drawbacks, as well as the outcomes, are remarkably comparable to those of a semantic differential. The stapel scale, on the other hand, is generally easy to conduct and administer.

Statistical Properties of Different Scales

Scale	Basic characteristics	Common examples	Other example	Suitable Statistics	
				Descriptive	Inferential
Nominal	Numbers identify and classify objects	Student registration numbers, numbers on football players' shirts	Gender classification of retail outlet types	Percentages, mode	Chi-square, binomial test
Ordinal	Numbers indicate the relative positions of the objects but not the magnitude of differences between them	Rankings of the top four teams in the football World Cup	Ranking of service quality delivered by a number of shops; rank order of favourite TV programmes	Percentile, median	Rank order correlation, Friedman, ANOVA
Interval	Numbers indicate the relative positions of the objects but not the magnitude of differences between them	Temperature (Fahrenheit, Celsius)	Attitudes, opinions, index numbers	Range, mean, standard deviation	Product moment correlations, t tests, ANOVA, regression, factor analysis
Ratio	Zero point is fixed; ratios of scale values can be computed	Length, weight	Age, income, costs, sales, market shares	Geometric mean, harmonic mean	Coefficient of variation

Summary

A nominal, ordinal, interval, or ratio scale can be used to measure something. The scales describe the degree of liking/disliking, agreement/dissent, or belief in an object. Each scale has its own set of statistical implications. In market research, there are four types of scales: paired comparison, Likert, semantic differential, and Thurstone scale. The semantic differential scale is a seven-point scale, whereas the Likert scale is a five-point scale. In the semantic difference scale, bipolar adjectives are utilised. The Thurstone scale is intended to examine the respondents' attitudes about any issue of public concern. Before the scale is used for measurement, its validity and reliability are checked. "Does the scale measure what it claims to measure?" is the question. The sort of validity required depends on "What is being measured." There are three methods for checking validity.

Keywords

Scaling: referred to as the assignment of objects to numbers according to a set of rules.

Measurement: assigning numbers or other symbols to characteristics of objects being measured, according to predetermined rules.

Interval Scale: Interval scale may tell us "How far the objects are apart with respect to an attribute?"

Likert Scale: This is a combination of statements about an attitude object. On the scale of '5 points, Agree, and Disagree,' each statement has 5 points, Agree, and Disagree.

Ordinal Scale: In most market research studies, the ordinal scale is applied for ranking.

Ratio Scale: Ratio scale is a special kind of internal scale that has a meaningful zero point.

Self Assessment

1..... scale may tell us "How far the objects are apart with respect to an attribute?" Which of the following is suitable to be filled-in in the given statement?

A. Interval
B. Nominal
C. Ratio
D. Ordinal

2.Ratio scale is a special kind of internal scale that has a meaningful ..... Which of the following is suitable to be filled-in in the given statement?

A. Zero point
B. Mid Point
C. Fraction
D. None of these

3.The salary of Ram is twice as much as the salary of Shyam – this is an example of:

A. Nominal scale measurement
B. Ordinal scale measurement
C. Interval scale measurement
D. Ratio scale measurement

4.The constant sum rating scale would result in which type of measurement?

A. Nominal scale
B. Ordinal scale
C. Interval scale
D. Ratio scale

5.Social class is an example of:

A. Nominal scale
B. Ordinal scale
C. Interval scale
D. Ratio scale

6.Which of the following scales possess an absolute zero?

A. Nominal scale

B. Ordinal scale

C. Interval scale

D. Ratio scale

7.In which of the following scales the objects are arranged according to their magnitude in an ordered relationship?

A. Nominal scale

B. Ordinal scale

C. Interval scale

D. Ratio scale

8.In which of the following scales does difference in scores have meaningful interpretation?

A. Nominal

B. Ratio

C. Interval

D. Both b and c

9.'Priyanka Chopra is more beautiful than Madhuri Dixit' – this is an example of

A. Nominal

B. Ordinal

C. Ratio

D. Interval

10.In which of the following scales can all possible statistical techniques be applied?

A. Nominal

B. Ordinal

C. Ratio

D. Interval

11.The numbers assigned to the members of Team India is an example of

A. Nominal

B. Ordinal

C. Ratio

D. Interval

12.In comparative rating scales

A. Judgments are arranged in a particular order.

B. Relative judgments are involved.

C. Halo effect plays an important role in getting a correct response.

D. Response categories are numbered.

13.Measurement involves creating a continuum upon which measured objects are located.

A. False

B. True

14.we are not measuring the object but some characteristic of it when we measure the perceptions, attitudes, and preferences of consumers.

A. False

B. True

15.Only a limited number of statistics, all of which are based on frequency counts, are permissible on the

numbers in a nominal scale.

A. False

B. True

Answers for Self Assessment

1.	A	2.	A	3.	D	4.	D	5.	B
6.	D	7.	B	8.	D	9.	B	10.	C
11.	A	12.	B	13.	A	14.	B	15.	B

Review Questions

1. What is measurement and scaling?

2. Discuss measurement scales as tools of sound measurement?

3. Explain the characterisitics of nominal, ordinal, interval and ratio scales?

4. Define multi dimensional scaling? What are the possible uses of multi dimensional scaling, in your opinion?

5. Explain the construction of

(a) Likert scale

(b) Semantic differential scale

(c) Thurstone scale

6. Justify your answer by identifying the type of scale you will use in ordinal, nominal, internal, ratio scales?

Further Readings

C.R.Kotari, Research Methodology, VishwaPrakashan. David Luck and Ronald Rubin, Marketing Research, PHI.

G.C.Beri, Marketing Research, TMH.

Paneerselvam, R, Research Methods, PHI.

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

Tull and Donalds, Marketing Research, MMIL.

Web Links

https://www.coursera.org/lecture/applying-data-analytics-business-in-marketing/lesson-1-3-2- measurements-and-scaling-techniques-primary-scales-

of-measurement-nhWnx
https://www.uoguelph.ca/hftm/book/export/html/2106
https://conjointly.com/kb/scaling-in-measurement/

Unit 07: Data Collection Methods

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

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.

7.1 Methodology for Collection of Primary Data

Many times due to inadequacy of data or stale information, the need arises for collecting a fresh firsthand information. In marketing research, there are broadly two ways by which primary information can be gathered namely, observation and communication.

Benefits 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. Firsthand 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.

Limitation of Primary data

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.

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.

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.

Behaviour or attitude of the children, and also of those who are inarticulate.

7.2 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.

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

The observation method is the only method applicable to study the growth of plants and crops.

To use a more structured approach, it would be necessary to decide precisely whatis 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 understand 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.

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.

It is easier to record structured observation than non-structured observation.

Merits 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.

Demerits 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.

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
3. TV?
4. A study to find out the potential location for a snack bar in a city.

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.

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.

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?

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.

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.

Advantages of Survey

1. Access to wide range of participants.
2. Collection of large amount of data.
3. May be more ethical than experimental designs.

Disadvantages of Survey

1. Lack of control.
2. Data may be superficial.

3. Costly to obtain representative data.

7.3 Survey Methods

The Survey method is the technique of gathering data by asking questions to people who are thought to have desired information. Broadly survey methods can be put like: Survey research tool and time involved to conduct research.

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.

Advantages

1. The ability to let the Interviewee see, feel and/or taste a product.
2. 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.
3. 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

1. Personal interviews usually cost more per interview than other methods.
2. Change in the characteristics of the population might make sample non-representative.

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 limit 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.

7.4 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.
- 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.

7.5 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 setup 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 generalized 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.

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.

Software selection is especially important in internet survey so it should be selected with proper care and after analyzing through feasibility studies

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 setup 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. 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.

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 thisreason, 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

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.

Disadvantages

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.

"Tell me spontaneously, what comes to your mind if I ask you about cigarettesmoking".

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.

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.

Questionnaire

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.

Importance and Limitations of Questionnaire in Market Research

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 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.

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.

Which television programme did you see last Saturday? This requires a reasonably goodmemory, 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.

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 to 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.

"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 endeavourinthis 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?"

"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 much 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 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: "Whatfactor 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.

"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:

"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 measurementerror.

Close-Ended Questions

There are two basic formats in this type:

- Make one or more choices among the alternatives.
- Rate the alternatives.

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 bebuying 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.

One basic criticism of closed-ended questionnaires is that middle alternatives arenot 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.

"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.

"Do you think that, Brazil played poorly in the FIFA cup?" This is a straightforwardquestion. The answer could be 'yes', 'no' or 'don't know' depending on the knowledge the respondents have about the game.

"Do you think anything should be done to make it easier for people to pay theirphone bill, electricity bill and water bill under one roof"?

"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?

Instead of using the word 'reasonably', 'usually', 'occasionally', 'generally', 'on the whole'.

"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.

"Do you feel that firms today are employee-oriented and customer-oriented?"

There are two separate issues here - [yes] [no]

"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.

"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) Rainbow4) Radio-One.

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 you're 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.

2. Loaded Questions: A leading question is also known as a loaded question. In a loaded Notes question, special emphasis is given to a word or a phrase, which acts as a lead to respondent.

"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 the meaning of the word effective distribution is.

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 valueprice 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.

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.

"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

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 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 respondentto comprehend, leave alone answer.

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.

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"?

"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".

"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".

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.

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?

From the above example, it is clear that layout - 2 is better. This is because likely respondent error due to confusion is minimized.

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.

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 favorable responses are equal to the number of unfavorable responses. If the researcher knows that there is a possibility of a favorable 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 he would 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 cangive 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 ofrespondents. 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.
iv. To add or remove questions arise in the process of pretest, if any.
v. To purge the words with double meaning, etc.

Summary

- Primary data may pertain to lifestyle, 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.

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.

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.

Self Assessment

1. The----------is the most frequently used primary method of data collection in any area of business research. It involves a predetermined set of queries in a structured format.

A. In-depth interviews
B. Focus group discussions
C. Questionnaire
D. None of the above

2. The----------is the most frequently used primary method of data collection in any area of business research. It involves a predetermined set of queries in a structured format.

A. In-depth interviews
B. Focus group discussions
C. Questionnaire
D. None of the above

3. The great weakness of questionnaire design is _____
A. Precision
B. Accuracy
C. Theory
D. Consensus

4. Which of the following is *not* a part of the questionnaire design process?
A. Specify the type of questioning method.
B. Arrange questions in proper order.
C. Reproduce the questionnaire.
D. Develop the sampling plan.

5. The first step in the questionnaire design process is _____
A. Specify the type of interview method
B. Identify the form and layout
C. Specify the information needed
D. Determine the content of individual questions

6. The reason for the respondent's inability to answer the questions in a questionnaire could be because of:
A. The person might not have the required information
B. The person might not remember the answer
C. The person might not be able to articulate the answer
D. All of the above

7. The information that is of the most importance to the research project and should be obtained first is _____
A. Qualifying information
B. Identification information
C. Basic information
D. Classification information

8. Consider the following question: Don't you think the current government has an excellent poverty alleviation programme? Yes/no
A. Is a leading question
B. Is a loaded question
C. Is a double-barrelled question
D. Is an interval scaled question

9. Do you think taking dowry is the right of every Indian male? Is an example of a
A. Forced choice
B. Open-ended
C. Dichotomous
D. Loaded question

10. A questionnaire is an informal set of questions for obtaininginformation from respondents.
A. True
B. False

11. A well-designed questionnaire can motivate the respondents andincrease the response rate.

A. True

B. False

12. To conduct an e-mail survey, the survey is written within the body of the e-mail message.

A. True

B. False

13. Observing children playing with new toys is an example of unstructured observation

A. True

B. False

14. A major disadvantage of the questionnaire method is

A. It requires a lot of skill for administration.

B. It does not provide for spontaneity of response.

C. There is pressure for immediate response.

D. It can be simultaneously conducted on a large sample.

15. Which of the following is a disadvantage of the survey method of data collection?

A. The questionnaire is simple to administer.

B. The data obtained are reliable because the responses are limited to the alternatives stated.

C. Wording questions properly is not easy.

D. Coding, analysis, and interpretation of data are relatively simple.

Answer for Self Assessment

1.	C	2.	C	3.	C	4.	D	5.	C
6.	D	7.	C	8.	A	9.	D	10.	B
11.	A	12.	A	13.	A	14.	B	15.	C

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?

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.

Web Links

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC420179/

http://www.fao.org/3/w3241e/w3241e05.htm

https://www.youtube.com/watch?v=ywCDy7IWwHw

Unit 08: Descriptive Statistics and Time Series

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
- Outline the concept of time series and its various factors.
- Identify components of time series and explain the time series forecasting method.
- Define an index number and explain its use.
- Interpret and use a range of index numbers commonly used in business and other sectors.

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.

8.1 Measure of Central Tendency

A necessary function of any statistical study is data summarization. The massive mass of cumbersome data is summarised in the form of tables and frequency distributions as a first step in this approach. These tables and frequency distributions must be summarised further in order to put the data's qualities into sharp focus. In any statistical investigation, a measure of central tendency, also known as an average, is a critical and crucial summary measure.

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 quite difficult for the human mind to comprehend a large body of numerical facts. To summarise such data into a single figure that is easier to understand and remember, an average measure is used.
2. **To facilitate comparison:** Averages can be used to compare different sets of data. Workers' earnings in two factories, for example, can be compared using the mean (or average) wages of each factory's workers.
3. **To help in decision-making:**The majority of research, planning, and other decisions are based on the average value of particular variables. For example, if a company's average monthly sales are declining, the sales manager may need to make some steps to rectify the situation.

Characteristics of a Good Average

The following criteria must be present in a good average way of measuring:

1. It should be precisely specified, preferably by an algebraic formula, so that different people get the same result for the same set of facts.
2. It should be simple to calculate.
3. It should be simple to comprehend.
4. It should be based on all of the findings.
5. It should be able to be subjected to further algebraic manipulation.
6. Extreme observations should not have an undue impact.
7. It should be relatively unaffected by sampling changes.
8.

8.2 Various Measures of Average

The concept "central tendency" relates to average measurements, specifically the middle, or normal, value of a set of data, which is typically calculated using the three m's: mean, median, and mode. The measurements of central tendency are the mean, median, and mode:

Meanor 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 X1, X2… Xn , respectively. The sum of these observations X1 + X2 +..... + Xn will be denoted in abbreviated form as,

$$\sum_{i=1}^{n} X_i$$

where ∑ (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 ∑, as written earlier.

When there is no ambiguity in range of summation, this indication can be skipped and we may simply write $X_1 + X_2 + ... + X_n = SX_i$.

Mean or Arithmetic Mean can be worked out using ungrouped and grouped data.

Mean: Ungrouped Data

For Ungrouped Data or raw data, the mean has following formula

$$\bar{x} = \frac{\sum x}{n} \text{ Where } \bar{X} = \text{Mean } \sum x = \textit{mean of the measurement or values}$$

$$n = Number\ of\ measurements$$

Problem:

- Mr. Sharma operates a Web site service that employs 8 people. Find the mean age of his workers if the ages of the employees are as follows:
  55, 63, 34, 59, 29, 46, 51, 41

Solution:

- Add the numbers to determine the total age of the workers.

  55 + 63 + 34 + 59 + 29 + 46 + 51 + 41 = 378
- Divide the total by the number of months.
- $\frac{378}{8} = 47.25$

Mean: Grouped Data

A mean can also be determined for data that is grouped, or placed in intervals. Unlike listed data, the individual values for grouped data are not available, and it is not possible to calculate their sum. To calculate the mean of grouped data, the first step is to determine the midpoint of each interval or class.These midpoints must then be multiplied by the frequencies of the corresponding classes. The sum of the products divided by the total number of values will be the value of the mean.

Problem:

- In an office, there are 25 employees. Each employee travels to work every morning in his or her own car. The distribution of the driving times (in minutes) from home to work for the employees is shown in the table below.

Driving Times (minutes)	Number of Employees
0 to less than 10	3
10 to less than 20	10
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Solution:

Step 1: Determine the midpoint for each interval.

- For 0 to less than 10, the midpoint is 5.
- For 10 to less than 20, the midpoint is 15.
- For 20 to less than 30, the midpoint is 25.

- For 30 to less than 40, the midpoint is 35.
- For 40 to less than 50, the midpoint is 45.

Step 2: Multiply each midpoint by the frequency for the class.

- For 0 to less than 10, (5)(3) = 15
- For 10 to less than 20, (15)(10) = 150
- For 20 to less than 30, (25)(6) = 150
- For 30 to less than 40, (35)(4) = 140
- For 40 to less than 50, (45)(2) = 90

Step 3: Add the results from Step 2 and divide the sum by 25

- 15 + 150 + 150 + 140 + 90 = 545

  Mean=545/25 =21.8

Outliers

As a summary statistic, the mean is frequently employed. Extreme values, on the other hand, have an impact (outliers). Outliers are numbers that are extraordinarily high or low. The mean isn't an appropriate summary statistic when there are extreme values at one end of a data set.

Median

The value of the variate that splits it into two equal pieces is the median of the distribution. The ordinate drawn at the median divides the area under the curve into two equal portions in terms of a frequency curve. The median is a positional average because its value is determined by the item's position rather than its size.

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 $\frac{n}{2}+1$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.

Since n = 7, i.e., odd, the median is the size of 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.

Since n = 7, i.e., odd, the median is the size of (7+1/2) , i.e., 4th observation.

Hence, median, denoted by Md = 20. i.e., 4th observation.

Hence, median, denoted by Md = 20.

The Mode

There will only be one mean and one median for the same collection of data. When describing the mode of a data set, the term modal is frequently employed. The term "unimodal" refers to a data collection that contains only one value that occurs most frequently. Bimodal data is defined as a set of data with two values that occur with the same maximum frequency. The term "multimodal" refers to a set of data that contains more than two values that occur with the same highest frequency.

Example1: Find the mode of the following data:

76, 81, 79, 80, 78, 83, 77, 79, 82, 75

In the above data set, the number 79 appears twice, but all the other numbers appear only once. Since 79 appears with the greatest frequency, it is the mode of the data values.

Example 2: The ages of 12 randomly selected customers at a local Best Buy are listed below:

23, 21, 29, 24, 31, 21, 27, 23, 24, 32, 33, 19

What is the mode of the above ages?

The above data set has three values that each occur with a frequency of 2. These values are 21, 23, and 24. All other values occur only once. Therefore, this set of data has three modes.

Example 3: You begin to observe to the color of clothing your employees wear. Your goal is to find out what color is worn most frequently so that you can offer company shirts to your employees.

Monday: Red, Blue, Black, Pink, Green, and Blue

Tuesday: Green, Blue, Pink, White, Blue, and Blue

Wednesday: Orange, White, White, Blue, Blue, and Red

Thursday: Brown, Black, Brown, Blue, White, and Blue

Friday: Blue, Black, Blue, Red, Red, and Pink

The color blue was worn 11 times during the week. All other colors were worn with much less frequency in comparison to the color blue

8.3 Dispersion and Distribution

In statistics, the measure of central tendency gives a single value that represents the whole value; however, the central tendency cannot describe the observation fully. Dispersion refers to the extent to which a set of data is spread out, or dispersed from the 'average'.The measures of dispersion are also called averages of the second order because they are based on the deviations of the different values from the mean or other measures of central tendency which are called averages of the first order.

The measure of dispersion helps a researcher to study the variability of the items.

In a statistical sense, dispersion has two meanings:

1. It measures the variation of the items among themselves.

2. Measures the variation around the average.

Simply, if the difference between the value and average is high, then dispersion will be high. A low dispersion will be low if difference between calculated value and average is less.

Characteristics of an Ideal Measure of Dispersion

- It should be rigidly defined.
- It should be easy to understand and easy to calculate.
- It should be based on all the observations of the data.
- It should be easily subjected to further mathematical
- It should be least affected by the sampling fluctuation.
- It should not be unduly affected by the extreme values

In order to measure dispersion, following are the measures that are quite often used

1. Range
2. Quartile deviation
3. Mean Deviation
4. Standard Deviation

Range

Range is the simple measure of dispersion, which is defined as the difference between the largest value and the smallest value. Mathematically, the absolute and the relative measure of range can be written as the following:

$$R = L - S$$

Coefficient of Range= L-S/L+S, Where R= Range, L= largest value, S=smallest value

So, it can be said that the Range is the difference between the lowest and highest values.

Example: In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9.

So the range is 9 − 3 = 6.

The range can sometimes be misleading when there are extremely high or low values.

Example: In {8, 11, 5, 9, 7, 6, 3616}:

The lowest value is 5, and the highest is 3616,

So the range is 3616 − 5 = 3611.

The single value of 3616 makes the range large, but most values are around 10.

Quartile Deviation

The Quartile Deviation is a simple way to estimate the spread of a distribution about a measure of its central tendency (usually the mean).

It gives an idea about the range within which the central 50% of sample data lies.

The first quartile or the lower quartile or the 25th percentile, also denoted by Q1, corresponds to the value that lies halfway between the median and the lowest value in the distribution (when it is already sorted in the ascending order). Hence, it marks the region which encloses 25% of the initial data.

Similarly, the third quartile or the upper quartile or 75th percentile, also denoted by Q3, corresponds to the value that lies halfway between the median and the highest value in the distribution (when it is already sorted in the ascending order). It, therefore, marks the region which encloses the 75% of the initial data or 25% of the end data.

The Coefficient of Quartile Deviation

Coefficient of Quartile Deviation = Q3-Q1/Q3+Q1×100

Since it involves a ratio of two quantities of the same dimensions, it is unit-less. Thus, it can act as a suitable parameter for comparing two or more different datasets which may or may not involve quantities with the same dimensions.

Example1:The number of vehicles sold by a Toyota Showroom in a day was recorded for 10 working days. The data is given as:

Day	Frequency
1	20
2	15
3	18
4	5
5	10
6	17
7	21
8	19
9	25
10	28

We first need to sort the frequency data before proceeding with the quartiles calculation

Sorted Data – 5, 10, 15, 17, 18, 19, 20, 21, 25, 28 n(number of data points) = 10

Now, to find the quartiles, we use the logic that the first quartile lies halfway between the lowest value and the median; and the third quartile lies halfway between the median and the largest value.

$$\text{First Quartile } Q_1 = \frac{n+1}{4} \text{ th term.}$$
$$= \frac{10+1}{4} \text{ th term } = 2.75 \text{ th term}$$
$$2\text{nd term } + 0.75 \times (3 \text{ rd term } - 2 \text{ nd term })$$
$$= 10 + \quad 0.75 \times (15 - 10)$$
$$= \quad 10 + 3.75$$
$$= 13.75$$

$$\text{Third Quartile } Q_3 = \frac{3(n+1)}{4} \text{ th term.}$$
$$= \frac{3(10+1)}{4} \text{ th term } = 8.25 \text{ th term}$$
$$= 8 \text{ th term } + 0.25 \times (9 \text{ th term } - 8 \text{ th term })$$
$$= 21 + 0.25 \times (25 - 21)$$
$$= \quad 21 + 1$$
$$= 22$$

Using the values for Q1 and Q3, now we can calculate the Quartile Deviation and its coefficient

$$= \frac{Q_3 - Q_1}{2}$$
$$= \frac{22 - 13.75}{2}$$
$$= \frac{8.25}{2}$$
$$= 4.125$$

$$= \frac{\bar{Q}^3 - \bar{Q}^1}{\bar{Q}^3 + \bar{Q}^1} \times 100$$
$$= \frac{22 - 13.75}{22 + 13.75} \times 100$$
$$= \frac{8.25}{35.75} \times 100$$
$$\approx 23.08$$

Mean Deviation

The mean deviation is defined as a statistical measure which is used to calculate the average deviation from the mean value of the given data set.

Step 1: Find the mean value for the given data values

Step 2: Now, subtract mean value form each of the data value given (Note: Ignore the minus symbol)

Step 3: Now, find the mean of those values obtained in step 2.

Mean Deviation = $[\Sigma \ |X - \mu|]/N$

Σ represents the addition of values

X represents each value in the data set

M represents the mean value of the data set

N represents the number of data values

Example:

Problem: Determine the mean deviation for the data values

5, 3, 7, 8, 4, 9.

Given data values are 5, 3, 7, 8, 4, 9.

We know that the procedure to calculate the mean deviation.

First, find the mean for the given data:

Mean, $\mu = (5+3+7+8+4+9)/6$

$\mu = 36/6$

Now, subtract each mean from the data value, and ignore the minus symbol if any (Ignore"-")

5 – 6 = 1

3 – 6 = 3

7 – 6 = 1

8 – 6 = 2

4 – 6 = 2

9 – 6 = 3

Now, the obtained data set is 1, 3, 1, 2, 2, 3.

Finally, find the mean value for the obtained data set. Therefore, the mean deviation is = (1+3 + 1+ 2+ 2+3) /6 = 12/6 = 2

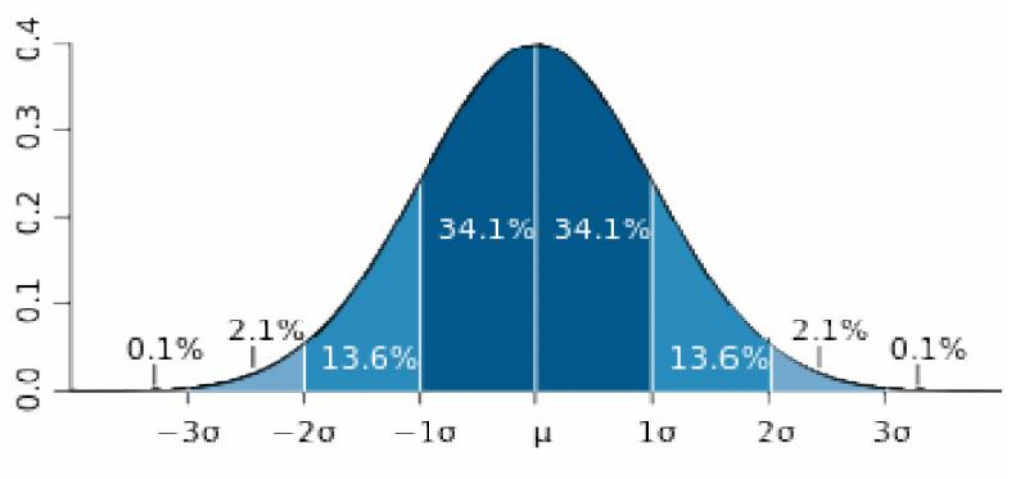
Hence, the mean deviation for 5, 3, 7, 8, 4, 9 is 2.

Standard Deviation

The Standard Deviation is a measure of how spread out numbers are.

Its symbol is σ (the greek letter sigma)

Normal Distribution Curve



The bell curve (also known as a "normal distribution" by statisticians) is a prominent tool used in statistics to understand standard deviation. In actual life, the graph of a normal distribution below reflects a large amount of data. The Greek letter, at the centre, represents the mean, or average. Each segment (dark blue to light blue in colour) reflects a standard deviation from the mean.

The formula to find the standard deviation (s) when working with samples is:

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

8.4 Index Numbers

A statistical instrument for evaluating relative changes in magnitude of a group of related variables across time is an index number. The percentages are used to express the index numbers. The base period is the first of the two periods with which the comparison is to be made. The index number for the base period is always 100. As a result, studying index numbers allows us to determine the percentage change in the values of various variables through time with respect to the base year.

An index number is an economic data figure reflecting price or quantity compared with a standard or base value.Index numbers are named after the activity they measure.

Price Index:

Measure changes in price over a specified period of time. It is basically the ratio of the price of a certain number of commodities at the present year as against base year

Quantity Index:

As the name suggest, these indices pertain to measuring changes in volumes of commodities like goods produced or goods consumedetc.

Value Index:

These are used to compare changes in the monetary worth of items imported, exported, produced, and consumed.

The base value is normally 100, and the index is 100 times the base value's ratio. If a commodity costs twice as much in 2019 as it did in 2018, the index number for 2019 would be 200.

Business activity, the cost of living, and employment are all measured by index numbers.

It allows economists to simplify complex business data.

Simple index number

A simple index number is a number that expresses the relative change in price, quantity, or value from one period to another. Let $p0$ be the base period price, and $p1$ be the price at the selected or given period. Thus, the simple price index is given by:

$P = p1 / p0 (100)$

Characteristics of index numbers

- Expressed in percentage
- Relative measures or measures of net changes
- Measure change over a period of time or in two or more places
- Specialized average
- Measuring changes which are not directly measurable

The construction of the price index numbers

1. Selection of Base Year
2. Selection of Commodities
3. Collection of Prices
4. Selection of Average
5. Selection of Weights
6. Purpose of Index Numbers
7. Selection of Method

1. Selection of Base Year

The choice of the base year is the first stage or challenge in preparing the index numbers. The base year is the year against which price changes from subsequent years are assessed and expressed as percentages. A regular year should be used as the basis year.

In other words, it should be devoid of unusual circumstances such as wars, famines, floods, political unrest, and so on. The base year can be chosen in one of two ways: (a) using the fixed base technique, in which the base year remains constant; or (b) using the chain base method, in which the base year changes over time, for example, the base year for 1980 will be 1979, 1979 will be 1978, and so on.

2. Selection of Commodities

The selection of commodities is the second issue in the development of index numbers. Because all commodities cannot be included, only representative commodities should be chosen, keeping in mind the index number's purpose and nature.

3. Collection of Prices

Points to be kept in mind:

(a) From where the prices to be collected;

(b) Whether to choose wholesale prices or retail prices;

(c) Whether to include taxes in the prices or not etc.

4. Selection of Average

The fourth problem is to find a reasonable average because the index numbers are a specialized average. In theory, the geometric mean is the best option for this. In practice, however, the arithmetic mean is chosen since it is easier to understand.

5. Selection of Weights

The commodities should be given appropriate weights based on their relative importance.

When calculating the cost-of-living index for instructors, for example, book costs will be given more weight than when calculating the cost-of-living index for workers. Weights should be neutral and chosen logically rather than arbitrarily.

6. Purpose of Index Numbers

Different index numbers are created for specific purposes, and no one index number can be considered an "all-purpose" index number. It is critical to understand the function of an index number before it is created.

7. Selection of Method

There are two methods of computing the index numbers:

(a) Simple index number:

Simple index number can be constructed either by -

(i) Simple aggregate method

OR

(ii) Simple average of price relative's method.

(b) Weighted index number:

Weighted index number can be constructed either by:

(i) Weighted aggregative method

OR by

(ii) Weighted average of price relative's method - The choice of method depends upon the availability of data, degree of accuracy required and the purpose of the study.

Limitations of index numbers

- Limitedcoverage
- Indexnumbersarebasedonsample items.
- Qualitativechangesareignored
- Ignoreschangesin theconsumptionpattern
- Limitedapplicability
- Misleading results -index numbers may not perfect it.
- Wrongbaseyearhasbeentaken; wrong formulaeor wrongweightageis takenetc.
- Basedonaverages

8.5 Time Series

An ordered sequence of a variable's values at evenly spaced time periods.

Data collected at several points in time is referred to as time series data. This is in contrast to cross-sectional data, which looks at persons, businesses, and other entities at a particular point in time.

There is the possibility of correlation between observations since data points in time series are collected at neighbouring time periods. One of the characteristics that sets time series data apart from cross-sectional data is this.

Example of Time Series Data

Field	Example topics
Economics	Gross Domestic Product (GDP), Consumer Price Index (CPI), S&P 500 Index, and unemployment rates
Social sciences	Birth rates, population, migration data, political indicators
Epidemiology	Disease rates, mortality rates, mosquito populations
Medicine	Blood pressure tracking, weight tracking, cholesterol measurements, heart rate monitoring

What Is Time Series Data?

Time series data is a collection of quantities that are assembled over even intervals in time and ordered chronologically. The time interval at which data is collection is generally referred to as the time series frequency.

Components of a Time Series

Different types of variables can be affected by different types of factors, e.g., factors affecting the agricultural output may be entirely different from the factors affecting industrial output.

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

1. Secular/Simply Trend

The general tendency of data to increase, diminish, or stagnate over a long period of time is known as a secular trend or simply trend. The majority of business and economic time series would show a trend to increase or decrease over time. Data on industrial production, agricultural production, population, bank deposits, deficit financing, and so on, for example, reveal that these magnitudes have been rising for a long time. In contrast to this, a time series may exhibit a diminishing trend, for example, when one product is substituted with another, the demand for the substituted commodity, such as cotton clothing, coarse grains like bajra, jowar, and so on, will show a declining trend. The death rate is projected to show a downward trend as medical services improve, etc. In any scenario, basic forces such as changes in population, technology, production composition, and so on are responsible for the shift in trend.

A long period isn't defined by a definite length of time. Long lengths of time vary depending on the situation. For example, in the case of population or output trends, the long period could be ten years, whereas the daily demand trend for vegetables could be a month. It should be observed, however, that the longer the period, the more important the trend. Furthermore, the growth or drop in values does not have to continue in the same direction throughout the duration. The statistics could show a rising (or falling) trend at first, followed by a decreasing (or rising) trend. etc.

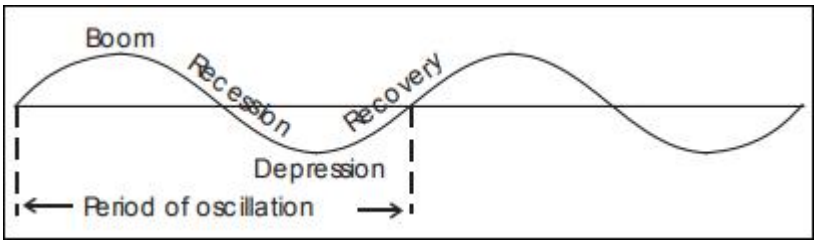
Objectives of Measuring Trend

Measuring the trend of a time series data has four basic goals:

- Researching the series' previous growth or decline. When short-term fluctuations are ignored, trend describes the data's basic growth or fall tendency.
- The trend curve can be projected into the future for predicting if the same behavior is assumed to continue in the future.
- To investigate the impact of other factors, the trend can be assessed first and then subtracted from the observed values.
- Two or more time series' trend values can be utilized to compare them.

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:



If the period of oscillation is less than one year, the oscillatory movements are called Seasonal Variations, and if the period is larger than one year, they are called Cyclical Variations. Seasonal and cyclical changes may be present in a time series where the time gap between subsequent measurements is less than or equal to one year. Seasonal changes, on the other hand, are nonexistent if the time period between successive measurements is more than a year. Although periodic variations are more or less regular, they are not always uniformly periodic, meaning that the pattern of their variations in different periods may or may not be the same in terms of time period and amount of periodic variations. For example, if a cycle takes five years to complete, the next cycle may take more or shorter than five years to complete.

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.

The sale of garments and jewelry may be highest during the wedding season, the sale of sweets during Diwali, and so on, are examples of differences that are the outcome of people's rituals and traditions. It's worth noting that both of the aforementioned causes occur on a regular basis and are frequently 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: Most economic and business time series display cyclical changes, which are commonly referred to as trade (or business) cycles. Any trade cycle has four parts, which are known as the boom, recession, depression, and recovery phases, respectively. Various phases in the above sequence repeat themselves on a regular basis. The period of cyclical variations is the time gap between two identical phases. The time span is always longer than a year. The duration of cyclical oscillations is usually between 3 and 10 years.

Objectives of Measuring Cyclical Variations: The main objectives of measuring cyclical variations are:

A. To analyse the behavior of cyclical variations in the past.
B. To predict the effect of cyclical variations so as to provide guidelines for future business policies.

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 are 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.

Forecasting Approaches

- Quantitative Forecasts uses one or more mathematical models that rely on historical data and/or causal variable to forecast demand.
- Qualitative Forecasts uses such factors like decision makers' intuition, emotions, personal experiences, and value system.

QUALITATIVE Forecasting Approaches

• JURY OF EXECUTIVE OPINION: The opinions of a group of high-level experts or managers, often in combination with statistical models, are pooled to arrive at an estimate of demand.

• DELPHI METHOD: Three different kinds of participants are included:

Decision makersconsists of a group of 5 to 10 experts who will be making the actual forecast.

Staff personnel assist decision makers by preparing, distributing collecting, and summarizing a series of questionnaires and survey results.

The respondents are a group of people, often located in different places, whose judgments are valued. They provide inputs to the decision makers before the forecast is made.

Consumer Market Survey: This method uses input from customers or potential customers regarding future purchasing plans. It can help not only in preparing a forecast but also in improving product design and planning for new products.

Sales Force Composite: Each salesperson estimates what sales will be in his or her region. The forecasts are then reviewed to ensure that they are realistic. Then they are combined at district and national levels to reach an overall forecast.

Quantitative Forecasting Approaches

The various forecasting methods are for quantitative data forecasting:

1. Naive Approach 2. Moving Averages 3. Exponential Smoothing 4. Trend Projection	TIME SERIES MODELS
5.Linear Regression	ASSOCIATIVE MODEL

TIME SERIES MODELS:

A time series is a set of observations of a variable at regular intervals over time. In decomposition analysis, the components of a time series are generally classified as trend T, cyclical C, seasonal S, and random or irregular R.

Time series are tabulated or graphed to show the nature of the time dependence. The forecast value (Ye) is commonly expressed as a multiplicative or additive function of its components; examples here will be based upon the commonly used multiplicative model.

YC =T. S. C. R multiplicative model

YC =T + S+ C+ R additive model

Where T is Trend, S is Seasonal, C is Cyclical, and R is Random components of a series.

Trend is a gradual long-term directional movement in the data (growth or decline).

Seasonal effects are similar variations occurring during corresponding periods, e.g., December retail sales. Seasonal can be quarterly, monthly, weekly, daily, or even hourly indexes.

Cyclical factors are the long-term swings about the trend line. They are often associated with business cycles and may extend out to several years in length.

Random component are sporadic (unpredictable) effects due to chance and unusual occurrences. They are the residual after the trend, cyclical, and seasonal variations are removed.

Summary

- Descriptive statistics are used to characterise the fundamental characteristics of a study's data.
- They give quick summaries of the sample and the measurements.
- They are the foundation of practically every quantitative data analysis, along with simple graphical analysis.
- Descriptive statistics are used to display quantitative data in a logical and understandable manner. We may have several measures in a research investigation.
- When summarising a quantity such as length, weight, or age, it is typical to use the arithmetic mean, median, or, in the case of a unimodal distribution, the mode to answer the first question.
- We use quantiles to select specific values from the cumulative distribution function.

- The variance, its square root, the standard deviation, the range, the interquartile range, and the average absolute deviation are the most popular metrics of variability for quantitative data (average deviation)
- A time series is a collection of data points taken at different times and separated by time intervals.

- Time series analysis refers to approaches for attempting to comprehend time series, usually to comprehend the underlying context of the data points or to create predictions.
- Time series forecasting is when a model is used to predict future events based on known previous events: anticipating future data points before they are measured.
- Data collected on a quarterly, monthly, weekly, daily, or hourly basis is likely to show seasonal fluctuations.
- In two or more scenarios, an index number is a technique for comparing the general magnitude of a group of distinct but linked variables.

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.

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.

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.

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.

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. The heights of 15 students of a class were noted as shown below. Compute arithmetic mean.

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

3. 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

4. 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

5. Define index number and discuss the characteristics of index numbers.
6. Examine various steps and problems involved in the construction of an index number.
7. What Is Time Series and discuss its components?
8. Discuss the various forecasting approaches in time series?

Self-Assessment

1. ..... is defined as the sum of observations divided by the number of observations.

A. Arithmetic Mean
B. Mean
C. Median
D. None

2. Median and mode are also known as the..... averages.

A. Central value
B. Positional
C. Modal
D. None

3. Median of distribution is that value of the variate which divides it into..... Parts.

A. two equal
B. three equal
C. no parts
D. none

4. Find the median of the following data: 160, 180, 200, 280, 300, 320, 400

A. 140
B. 180
C. 300
D. 280

5. In a frequency distribution the last cumulative frequency is 300, Median shall lie in:

A. 130^{th}

B. 140th
C. 160th
D. 150th

6. In a frequency distribution, the last cumulative frequency is 500. Q3 (Third Quartile) must lie in.
A. 275th
B. 375th
C. 150th
D. 175th
7. The average monthly production of a factory for the first 8 months is 2,500 units, and for the next 4 months, the production was 1,200 units. The average monthly production of the year will be
A. 2066.55 units
B. 2085.55 units
C. 2075.55 units
D. none

8. In a week the prices of a bag of rice were 350, 280, 340, 290, 320, 310, 300. The range is
A. 100
B. 70
C. 60
D. 90

9. The most frequently occurring number in a set of values is called the ____.
A. Mean
B. Median
C. Mode
D. Range

10. When a set of numbers is heterogeneous, you can place more trust in the measure of central tendency as representing the typical person or unit.
A. True
B. False

11. The median is _____.
A. The middle point
B. The highest number
C. The average
D. Affected by extreme scores

12. Which of the following is NOT a common measure of central tendency?
A. Mode
B. Range

C. Median
D. Mean

13. The variation in two or more variables studies by the index is called:
A. Composite index
B. simple index
C. price index
D. none of these

14. The weights used in a quantity index are:
A. Quantity
B. values
C. Prices
D. none of these

15. Which of the following is an example of time series problem?
1. Estimating number of hotel rooms booking in next 6 months.
2. Estimating the total sales in next 3 years of an insurance company.
3. Estimating the number of calls for the next one week.
A. Only3
B. 1 and 2
C. 2 and 3
D. 1, 2 and 3

Answers for Self Assessment

1.	A	2.	B	3.	A	4.	D	5.	A
6.	B	7.	A	8.	B	9.	C	10.	B
11.	A	12.	B	13.	A	14.	C	15.	D

Further Readings

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

Allan &Blumon, Elementary Statistics: A Step by Step Approach. McGraw-Hill College, June 2003.

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

Web Links

https://www.youtube.com/watch?v=Rjwknl_LuKw
https://www.youtube.com/watch?v=98K7AG32qv8
https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/
https://www.vedantu.com/commerce/index-numbers

Unit 09: 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 way of using experimental data to make statistical decisions. A finding is statistically significant in statistics if it is unlikely to have occurred by chance.

In contrast to exploratory data analysis, hypothesis testing is sometimes referred to as confirmatory data analysis. These decisions are almost always made using null hypothesis tests in frequency probability; that is, tests that answer the query. What is the probability of seeing a value for the test statistic that is at least as extreme as the value that was actually seen, assuming the null hypothesis is true? Hypothesis testing can be used to determine whether experimental data contain enough information to call conventional knowledge into question.

In a two-way ANOVA, the interaction term tells you if one of your independent factors has the same influence on the dependent variable for all values of the other independent variable (and vice versa). Is the effect of educational level (undergraduate/postgraduate) on test anxiety influenced by gender (male/female)? If a statistically significant interaction is discovered, you must also evaluate whether any "simple main effects" exist, and if so, what these effects are.

9.1 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.

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:

- Null hypothesis
- Alternate hypothesis

Assume that the population's mean is m0 and that the sample's mean is x. This is our null hypothesis because we believed the population has a mean of m0. Hom = mo, where Ho is the null hypothesis, is how we write it. Ha = m is an alternative hypothesis. The null hypothesis will be rejected, indicating that the population mean is not m0. This means that the alternative hypothesis is valid.

Significance Level

The validity of the hypothesis at a specific level of significance is the next stage after it has been formulated. The significance level determines the level of confidence with which a null hypothesis is accepted or rejected. A significance level of 5%, for example, suggests that the chance of making a bad judgement is 5%. On 5 out of 100 occurrences, the researcher will be erroneous in accepting a false hypothesis or rejecting a true hypothesis. A significance level of one percent suggests that the researcher has a one-in-a-hundred chance of being mistaken in accepting or rejecting the hypothesis. As a result, a 1% significance level provides more confidence in a judgement than a 5% significance level.

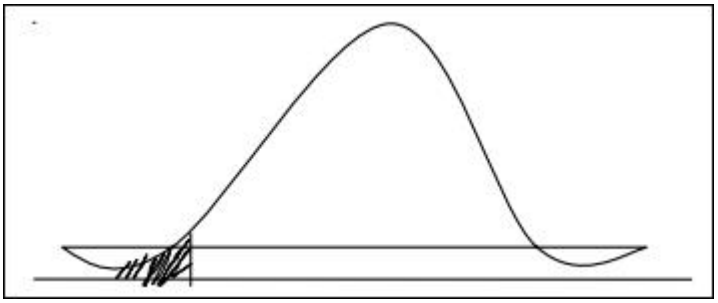
There are two types of tests.

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.

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.



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.

Example 1: 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 2: "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 3:"Are the two markets - Mumbai and Delhi different to test market a

product?"

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".

∴ Ho = m1 = m2

Sign of alternate hypothesis	Type of test
= /	Two-sided
<	One-sided to right
>	One-sided to left

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.

Discuss when you would prefer two tailed test to one tailed test.

9.2 Errors in Hypothesis Testing

There are two types of errors:

a) Hypothesis is rejected when it is true.
b) Hypothesis is not rejected when it is false.

(a) is called Type 1 error (a), (b) 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.

The only way to reduce Type 1 and Type 2 error is by increasing the sample size

Type 1 and Type 2 Errors

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 retailer ship. 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.

9.3 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
   - Z-Test,
   - 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.

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:

One Sample Test

One sample tests can be categorized into 2 categories.

Z Test

When sample size is > 30

1. P_1 = Proportion in sample 1

   P_2 = Proportion in sample 2

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.

Testing the hypothesis about difference between two means: This can be used when two population means are given and null hypothesis is Ho: P1 = P2.

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 Ho × μ = 15%

Alternate Hypothesis is Ha × μ> 15%.

This is a one-tailed (right) test.

$$Z = \frac{\frac{210}{1000} - 0.20}{\sqrt{\frac{0.20(1-0.20)}{1000}}}$$

$$Z = \frac{0.21 - 0.20}{\sqrt{\frac{0.2 \times 0.8}{1000}}}$$

$$= \frac{0.01 - \mu}{\sqrt{\frac{0.16}{1000}}}$$

$$= \frac{0.1}{\frac{0.4}{31.62}}$$

$$= \frac{0.1}{0.012} = 8.33$$

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.

Problem

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**

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{\bar{s}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$n_1=5 n_2=7$$

$$\Sigma x=230, \Sigma y=399$$

$$\Sigma(x-\bar{x})^2=310, \Sigma(y-\bar{y})^2=399$$

$$\bar{x}=\frac{\Sigma x}{n_1}=\frac{230}{5}=46$$

$$\bar{y}=\frac{\Sigma y}{n_2}=\frac{399}{7}=57$$

$$s^2=\frac{1}{n_1+n_2-2}\{\Sigma(x-\bar{x})^2+\Sigma(y-\bar{y})^2\}$$

$$\text{D.F.}=(n_1+n_2-2)=(5+7-2)=10$$

$$s^2=\frac{1}{10}\{310+674\}=98.4$$

$$t=\frac{46-57}{\sqrt{98.4\times\left(\frac{1}{5}+\frac{1}{7}\right)}}$$

$$=\frac{-11}{\sqrt{98.4\times\left(\frac{12}{35}\right)}}$$

$$=\frac{-11}{\sqrt{33.73}}=-\frac{11}{5.8}$$

$$=-1.89$$

t at 10 d.f. at 5% level is 1.81.

Since, calculated t is greater than 1.81, it is significant. Hence HA 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;

$$t=r_{yx}\sqrt{\frac{n-2}{1-r_{yx}}}$$

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.

Problem

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:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$r = 0.52, n = 18$$

$$t = 0.52\sqrt{\frac{18-2}{1-(0.52)^2}}$$

$$= \frac{0.52 \times 4}{0.854} = 2.44$$

$$v = (n-2) = (18-2) = 16$$

$$v = 16, t_{0.05} = 2.12$$

The calculated value of t is greater than the table value. The given value of r is significant.

Two Sample Test

Two sample tests if known as F test

F-Test

Let there be two independent random samples of sizes n_1 and n_2 from two normal populationswith variances σ_1^2 and σ_2^2 respectively. Further, lets$s_1^2 = \frac{1}{n_1-1}\sum(X_u - \bar{X}_1)^2$ 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

$$\left(\frac{v_2}{v_2-2}\right)\sqrt{\frac{2(v_1+v_2-2)}{v_1(v_2-4)}}$$

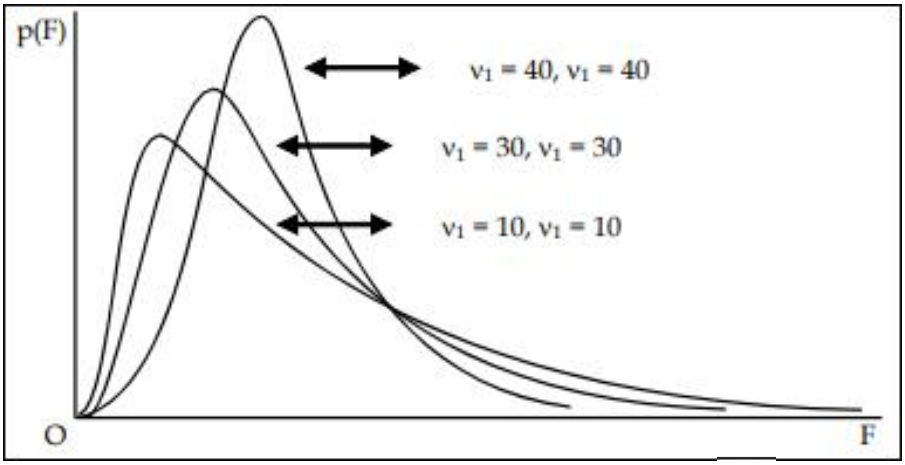
We note that the mean will exist if $v_2 > 2$ and standard error will exist if $v_2 > 4$. Further, the mean > 1.

3. The random variate F can take only positive values from 0 to ∞.The curve is positively skewed.

4. For large values of v_1 and v_2, the distribution approaches normal distribution.

5. If a random variate follows t-distribution with v degrees of freedom, then its square follows F-distribution with 1 and vd.f. i.e.$t_v = F_{1,v}$

6. Fandχ^2 are also related as $F_{v_1 v_2} = \frac{(\chi^2_n)}{v_1}$ as $v_2 \to \infty$.



Chi-square Test

When the null hypothesis is true, a chi-square test (also chi-squared or x^2 test) is any statistical hypothesis test in which the test statistic's sampling distribution is a chi-square distribution, or any in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to estimate a chi-square distribution as closely as desired.

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.

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?

Problem:

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 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.

9.4 Analysis of Variance (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.

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.

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.

Problem:

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., $\overline{XX} = \frac{\bar{X}_1+\bar{X}_2+\bar{X}_3}{K}$ 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 0 and by calculating the squares of such deviation, which are multiplied by the respective number of items or categories in the samples and then by Notes 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 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

$$i = 1.23$$

for our example, total ss will thus be:
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, dfis$(k - 1)$, where k ' represents number of samples (for us it is 3). For within samples dfis$(n - k)$, 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) MS between $= \frac{\text{ss between}}{(k-1)} = \frac{8}{2} = 4$
where $(K - 1)$ is thedf.

(ii) MS within $= \frac{\text{ss within}}{(n-k)} = \frac{24}{9} = 2.67$

1) where$(n - k)$ is thedf.

Step 6: Now we will have to compute F ratio by analysing our samples. The formula for computing 'F' ratio is: ss between 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 (v1* & v2**, i.e., and 9) is 1.5. The table value of F at 5% level with df 2 and 9 (v1 = 2, v2 = 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			

9.5 Two-way ANOVA

The two-way ANOVA compares mean differences between groups divided by two independent variables (called factors). The basic goal of a two-way ANOVA is to figure out if the two independent factors have an effect on the dependent variable. For example, you could use a two-way ANOVA to see if there is an interaction between gender and educational level on test anxiety among university students, where the independent variables are gender (males/females) and education level (undergraduate/postgraduate), and test anxiety is the dependent variable. Alternatively, you may look into if there is an interaction between physical activity and gender and blood cholesterol levels in youngsters, using physical activity (low/moderate/high) and gender (male/female) as independent factors and cholesterol concentration as the dependent variable.

In a two-way ANOVA, the interaction term tells you if one of your independent factors has the same influence on the dependent variable for all values of the other independent variable (and vice versa). Is the effect of educational level (undergraduate/postgraduate) on test anxiety influenced by gender (male/female)? If a statistically significant interaction is discovered, you must also evaluate whether any "simple main effects" exist, and if so, what these effects are.

Problem:

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.$\bar{x}$

3 Calculate variance between columns using the formula $\bar{\sigma}^2 = \frac{\sum n_i(x_i-\bar{x})^2}{k-1}$
where K = $(n_1 + n_2 + n_3 - 3)$

4 Calculate sample variance. It is calculated using formula:
Sample variance $s_i^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$ 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.

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

$$\frac{15 + 18 + 19 + 22 + 11 + 22 + 27 + 18 + 21 + 17 + 24 + 19 + 16 + 22 + 15 + 18}{16} = \frac{304}{16}$$

$$= 19$$

3-Calculate variance between columns

n	$\bar{x}$	$\bar{\bar{x}}$	$\bar{x} - \bar{\bar{x}}$	$(\bar{x} - \bar{\bar{x}})^2$	$n(\bar{x} - \bar{\bar{x}})^2$
5	17	19	-2	4	5 × 4 = 20
5	21	19	2	4	5 × 4 = 20
6	19	19	0	0	6 × 0 = 0
				$\sum n_i(\bar{x}_i - \bar{\bar{x}})^2$	= 40

$$\bar{\sigma}^2 = \frac{\sum n_i(x_i - \bar{x})^2}{k - 1} = \frac{40}{3 - 1} = 20$$

Variance between columns = 20

4-Calculation sample variance

Training method −1		Training method −2		Training method −3	
Training $x - \bar{x}$	thod − $(x - \bar{x})^2$	Trainir $x - \bar{x}$	netho $(x - \bar{x})^2$	Training $x - \bar{x}$	thod − $(x - \bar{x})^2$
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	$(-4)^2 = 16$
	$\Sigma(\overline{\quad}$ $= 70$		$\Sigma(\overline{\quad}^{\overline{\quad}}$ $= 62$		$\Sigma(\overline{\quad}^{\overline{\quad}}$ $= 60$

Sample variance $= \frac{\sum(x-\bar{x})^2}{n-1} = \frac{70}{5-1}, \frac{\sum(x-\bar{x})^2}{n-1} = \frac{62}{5-1}, \frac{\sum(x-\bar{x})^2}{n-1} = \frac{60}{5-1}$

$$s_1^2 = \frac{70}{4} = 17.5, s_2^2 = \frac{62}{4} = 15.5, s_3^2 = \frac{60}{5} = 12$$

5-Within column variance $\bar{\sigma}^2 = \sum\left(\frac{n_i-1}{n_i-k}\right)s_1^2$

$$= \left(\frac{5-1}{16-3}\right) \times 17.5 + \left(\frac{5-1}{16-3}\right) \times 15.5 + \left(\frac{6-1}{16-3}\right) \times 12$$

Within column variance $= \left(\frac{4}{13}\right) \times 17.5 + \left(\frac{4}{13}\right) \times 15.5 + \frac{5}{13} \times 12$

$$= \frac{192}{13} = 14.76$$

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 is3.81. This is the upper limit of acceptance region. Since calculated value

1.354lies within it we can accept $H_{0'}$ the null hypothesis.

Conclusion: There is no significant difference in the effect of the three training methods.

Non-parametric Test

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 several conditions, especially in marketing research, when parametric tests' assumptions aren't valid. In a parametric test, for example, we assume that the data is distributed normally. Non-parametric tests are employed in these situations. Binomial test, Mann-Whitney U test, Sign test, and other non-parametric tests are examples. When there are only two classes in a population, such as males and females, buyers and non-buyers, success and failure, a binomial test is performed. Every observation about the population must pass one of these two tests. When the sample size is small, the binomial test is employed.

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.

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.
- 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.

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.

Self Assessment

1-When the prediction does not specify a direction, the research have:

A. One-tailed hypothesis
B. Two-tailed hypothesis
C. Null hypothesis
D. None of these

2-The significance level can be denoted by:

A. Alpha
B. Beta
C. Gamma
D. Hyphen

3-A hypothesis is a specific statement of

A. Estimates
B. Assessment
C. Accuracy
D. Prediction

4-The prediction of the opposite direction is called as:

A. One-tailed hypothesis
B. Two-tailed hypothesis
C. Multi-tailed hypothesis
D. None of these

5-Null and alternative hypotheses are statements about:

A. population parameters.
B. sample parameters.
C. sample statistics.
D. it depends - sometimes population parameters and sometimes sample statistics

6-A statement made about a population for testing purpose is called?

A. Statistic
B. Hypothesis
C. Level of Significance
D. Test-Statistic

7-If the Critical region is evenly distributed then the test is referred as?

A. One tailed
B. Two tailed
C. Three tailed
D. Zero tailed

8-Alternative Hypothesis is also called as?

A. Composite Hypothesis
B. Research Hypothesis
C. Simple Hypothesis
D. Null Hypothesis

9-The statement "If there is sufficient evidence to reject a null hypothesis at the 10% significance level, then there is sufficient evidence to reject it at the 5% significance level" is:

A. Always True
B. Never True
C. Sometimes true; the p-value for the statistical test needs to be provided for a conclusion.
D. Not Enough Information: this would depend on the type of statistical test used

10-Any statement whose validity is tested based on a sample is called

A. Null hypothesis
B. Alternative hypothesis
C. Statistical hypothesis
D. Simple hypothesis

11-A hypothesis may be classified as:

A. Simple
B. Composite
C. Null
D. All of the above

12-A null hypothesis is rejected if the value of a test statistic lies in the

A. Rejection region
B. Acceptance region
C. Both
D. None

13-A p value of o.o5 means

A. There is only 5% chance that the results are incorrect
B. There is probability of 5 in 100 that this result would occur if the null hypotheses were true
C. There is only 5% chance of getting this result
D. All of the above are true

14- Which of the following distribution is useful for small sample while testing for population means?

A. Z distribution
B. F distribution
C. Chi-square distribution
D. T distribution

15-The t distribution could be used

A. When sample size is small (n< 30)
B. Sample is drawn from a normal population
C. Population variance is unknown
D. All the above statements are correct.

Answer for Self Assessment

1.	B	2.	A	3.	D	4.	A	5.	A
6.	B	7.	B	8.	B	9.	C	10.	C
11.	D	12.	A	13.	A	14.	D	15.	D

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.

		Red	Blue	Green
Colour	Male	5	14	6
Sex	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?

5. In hypothesis testing, if β is the probability of committing an error of Type II. The power of the test, $1 - \beta$ is then the probability of rejecting H_0 when H_A is true or not? Why?

6. In a statistical test of hypothesis, what would happen to the rejection region if α, the level of significance, is reduced?

7. 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?

8. Can the probability of a Type II error be controlled by the sample size? Why/ why not?

9. 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?

10. 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?

11. 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%.

12. 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.

13. 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)

Further Readings

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.

Web Links

https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/

https://machinelearningmastery.com/statistical-hypothesis-tests/

https://www.uth.tmc.edu/uth_orgs/educ_dev/oser/L2_2.HTM

https://online.stat.psu.edu/statprogram/reviews/statistical-concepts/hypothesis-testing

Unit10: Test of Association

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
- Calculate Chi-square as a non- parametric test

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.

10.1 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. "Correlation is an analysis of covariation between two or more variables."

– A.M. Tuttle

2. "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. 3. "Correlation analysis attempts to determine the 'degree of relationship' between variables".

— YaLun 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 tohave 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.

Scatter Diagram

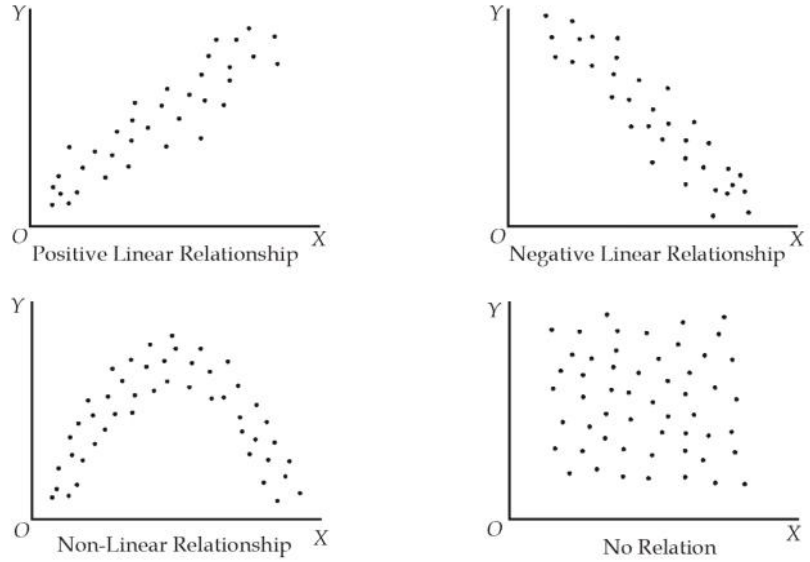
Let the bivariate data be denoted by (X_j, Y_i), where $i = 1,2 \ldots \ldots n$. In order to have some idea about the extent of association between variables X and Y, each pair $(X_p Y_i), i = 1,2 \ldots \ldots n$, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

Each pair of values (X,Y) 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:

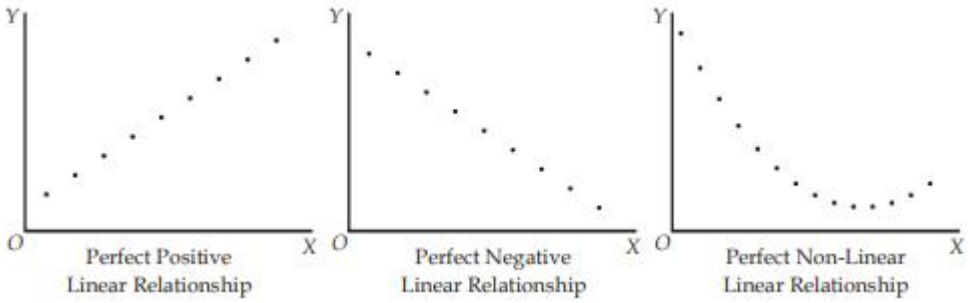
Did you Know?

What the sets of point in generally known?

The sets of points in scatter diagram are known as dots of the diagram



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:



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

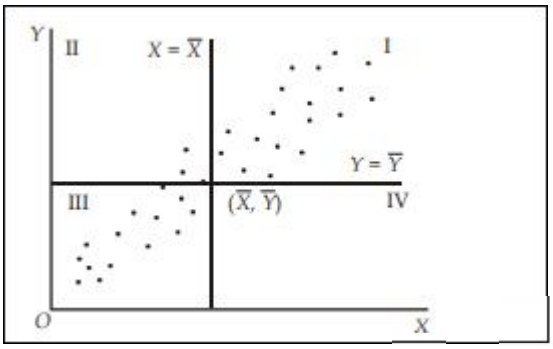
(high) value of Y, the association between them is said to be positive. Contrary to this, when low (high) values of X are associated with high (low) values of Y, the association between them is said to be negative.

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.

10.2 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$(X_i Y)$, $i = 1,2, \ldots \ldots n$. Further, let the scatter diagram of the data be as shown in Figure.

Let $\bar{X}$ and $\bar{Y}$ be the arithmetic means of X and Y respectively. Draw two lines $X = \bar{X}$ and $Y = \bar{Y}$ on the scatter diagram. These two lines, intersect at the point $(\bar{X}, \bar{Y})$ and are mutually perpendicular, divide the whole diagram into four parts, termed as I, II, III and IV quadrants, as shown.



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 $\bar{X}$ are generally associated with values of Y that are greater (less) than $\bar{Y}$, 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., $(X_i - \bar{X})$and $(Y_i - \bar{Y})$, we note that:

1 Both $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ will be positive for all points in quadrant I.

2 $(X_i - \bar{X})$will be negative and $(Y_i - \bar{Y})$ will be positive for all points in quadrant II.

3 Both $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ will be negative for all points in quadrant III.

4 $(X_i - \bar{X})$will be positive and $(Y_i - \bar{Y})$ will be negative for all points in quadrant IV.
It is obvious from the above that the product of deviations, i.e., $(X_i - \bar{X})(Y_i - \bar{Y})$ 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, $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$ will be positive for all the n observations.

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 $\Sigma(x_i - \bar{x})(Y, \bar{Y})$ 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., $\frac{1}{x}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$.This term is called covariance in statistics and is denoted as$\text{Cov}\,(X,Y)$.

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.

$$r_{xy} = \frac{\text{Cos}\,(X,Y)}{\sigma_x \sigma_Y}$$

or

$$r_{xy} = \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum (X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum (Y_1 - \bar{Y})^2}}$$

Cancelling $\frac{1}{n}$ from the numerator and the denominator, we get

$$r_{xy} = \frac{\sum (X_i - \bar{X})(Y_1 - \bar{Y})}{\sqrt{\sum (X_1 - \bar{X})^2}\sqrt{\sum (Y_1 - \bar{Y})^2}}$$

$$\text{Consider} \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X})Y_i - \bar{Y}\sum (X_i - \bar{X})$$
$$= \Sigma XY_i - \bar{X}\Sigma Y_i \text{ (second term is zero)}$$
$$= \sum X_i Y_i - n\bar{X}\bar{Y}\left(\sum Y_i = n\bar{Y}\right)$$

Similarly we can write $\sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$ and

$$\sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$

Substituting these values in equation (3), we have

$$r_{xy} = \frac{\sum X_i Y_1 - n\bar{X}\bar{Y}}{\sqrt{[\sum X_i^2 - H\bar{X}^2]}\sqrt{[Y_i^2 - n\bar{Y}^2]}}$$

$$= \frac{\sum X_1 Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sqrt{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}\sqrt{\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}}}$$

On multiplication of numerator and denominator by n, we can write

$$r_{xy} = \frac{n\sum XY_i - (\sum X_i)(\sum Y_i)}{\sqrt{n\sum X_i^1 - (\sum X_1)^2}\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}}$$

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

$$r_{xy} = \frac{\sum x_1 y_i}{\sqrt{\frac{1}{n}\sum x_i^2}\sqrt{\frac{1}{n}\sum y_i^2}}$$

or

$$r_{XY} = \frac{\sum X_1 Y_1}{\sqrt{\sum X_1^2}\sqrt{\sum Y_i^2}}$$

or

$$r_{xY} = \frac{1}{n}\frac{\sum x_i y_1}{\sigma_x \sigma_y}$$

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

$$\frac{n\sum XY_1 - (\sum X_i)(\sum Y_i)}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}}$$

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 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

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:

$$\frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

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) Σ_{X^2} is the sum of the squared ___res

(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:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{(10)(506) - (55)(103)}{\sqrt{(10)(385) - (55)^2}\sqrt{(10)(1401) - (103)^2}}$$

$$= \frac{(5060 - 5665}{\sqrt{3850 - 3205}\sqrt{14010 - 10609}} = \frac{-605}{\sqrt{825}\sqrt{3401}}$$

$$= \frac{-605}{(28.723)(58.318)} = \frac{-605}{1675.0679} = -0.36$$

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: 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.

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.

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 Xand 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:

$A + hu_i, \quad \therefore \quad \bar{X} = A + h\bar{u}$

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

$$r_{XY} = \frac{\sum h(u_i - \bar{u})k(v_i - \bar{v})}{\sqrt{\sum h^2(u_i - \bar{u})^2}\sqrt{\sum k^2(v_i - \bar{v})^2}} = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum (u_i - \bar{u})^2}\sqrt{\sum (v_i - \bar{v})^2}}$$

$\Rightarrow \quad r_{XY} = r_{uv}$

$\Rightarrow \quad r_{XY} = r_{uv}$

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 ard 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} :

$$r_{XY} = \frac{n\sum u_i v_i - (\sum u_i)(\sum v_i)}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}}$$

2-The coefficient of correlation lies between $-1$and $+1$.

To prove this property, we define

$$x_i' = \frac{X_i - \bar{X}}{\sigma_X} \text{ and } y'; = \frac{Y_i - \bar{Y}}{\sigma_Y}$$

$$\Rightarrow x_i'^2 = \frac{(X_i - \bar{X})^2}{\sigma_X^2} \text{ and } y_i'^2 = \frac{(Y_i - \bar{Y})^2}{\sigma_Y^2}$$

$$\text{or } \sum x_i'^2 = \frac{\sum (X_i - \bar{X})^2}{\sigma_X^2} \text{ and } \sum y_i'^2 = \frac{\sum (Y_i - \bar{Y})^2}{\sigma_Y^2}$$

From these summations we can write $\sum x_i'^2 = \sum y'^2 = n$

$$\text{Also, } r = \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_X \sigma_Y} = \frac{1}{n} \cdot \sum \left(\frac{X_i - \bar{X}}{\sigma_x}\right)\left(\frac{Y_i - Y}{\sigma_y}\right) = \frac{1}{n}\sum x', y_1'$$

Consider the sum $x_1' + y_i'$. 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

$$\frac{1}{n}\sum (x_i' + y_i')^2 \geq 0 \quad \text{or} \quad \frac{1}{n}\sum (x_i'^2 + y_i'^2 + 2x', y_i') \geq 0$$

$$\text{or } \frac{1}{n}\sum x_i'^2 + \frac{1}{n}\sum y^2 + \frac{2}{n}\sum x', y' \geq 0$$

$$\text{or } 1 + 1 + 2r^3 0 \text{ or } 2 + 2r^3 0 \text{ or } r^3 - 1$$

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

$$\frac{1}{n}\sum (x_i' - y_i')^2 \geq 0$$

$$\text{or } \frac{1}{n}\sum\left(x_i'^2 + y_i'^2 - 2x_i'y'\right) \geq 0$$

$$\text{or } \frac{1}{n}\sum x_i'^2 + \frac{1}{n}\sum y_i'^2 - \frac{2}{n}\sum x_i'y' \geq 0$$

or $1 + 1 - 2r^3 0$ or $2 - 2r^3 0$ or $r £ 1$

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 $\Sigma X_j Y_t = 112$

$$\text{Cov}(X,Y) = \frac{1}{n}\left[\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}\right] = \frac{1}{7}\left[112 - \frac{28 \times 28}{7}\right] = 0 \text{ Thus, } r_{XY} = 0$$

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: [illegible] ly a measure of the degree of linear [illegible] on 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.

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.

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.

10.3 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 p(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 $(X_\mu, Y_1), i = 1,2, ...n$. We assume here that there are no ties, i.e., no two or more individuals are tied to a particular rank. Thus, X_t' ' and $Y_i'^s$ are simply integers from 1 to n, appearing in any order. The means of $X$and$Y$, i.e., $\bar{X} = \bar{Y} = \frac{1+2+\cdots..+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$.

Also, $\sigma_x^2 = \sigma_y^2 = \frac{1}{n}[1^2 + 2^2 + \cdots + n^2] - \frac{(n+1)^2}{4} = \frac{1}{n}\left[\frac{n(n+1)(2n+1)}{6}\right] - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}$

Letd_1 be the difference in ranks of the i th individual, i.e.

$$d_i = X_i - Y_i = (X_i - \bar{X}) - (Y_i - \bar{Y})(\because \bar{X} = \bar{Y})$$

Squaring both sides and taking sum over all the observations, we get

$$\Sigma d_i^2 = \sum [(X_i - \bar{X}) - (Y_i - \bar{Y})]^2$$
$$= \sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2 - 2\sum (X_i - \bar{X})(Y_i - \bar{Y})$$

Dividing both sides by n, we get

$$\frac{1}{n}\sum d_1^2 = \frac{1}{n}\sum (X_i - \bar{X})^2 + \frac{1}{n}\sum (Y_i - \bar{Y})^2 - \frac{2}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})$$
$$- \sigma_x^2 + \sigma_r^2 - 2\text{Cov}\,(X,Y) = 2\sigma_x^2 - 2\text{Coo}\,(X,Y)(\because \sigma_x^+ = \sigma_Y^2)$$

From this, we can write $1 - \rho = \frac{1}{n} \times \frac{\sum d_i^2}{2\sigma_x^2}$

or

$$\rho = 1 - \frac{1}{n} \times \frac{\sum d_i^2}{2\sigma_x^2} = 1 - \frac{1}{n} \times \frac{\sum d_i^2}{2} \times \frac{12}{n^2 - 1} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Notes: This formula is not applicable in case of a bivariate frequency distribution

Question: 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.

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 series20,21,21,24,25,25,25,26,27,28. 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.

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 $\frac{n(m^2-1)}{12}$, 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.

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 1 st 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 $(n+1)$, ie., $X_i + Y_i = n+1$ or $Y_1 = (n+1) - X_y i = 1,2, \ldots \ldots n$

Further, $d_i = X_t - Y_i = X_i - (n+1) + X_i = 2X_i - (n+1)$

Squaring both sides, we have

$$d_i^2 = [2X_i - (n+1)]^2 = 4X_i^2 + (n+1)^2 - 4(n+1)X_i$$

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

$$= 1 - \frac{6n(n^2-1)}{3} \times \frac{1}{n(n^2-1)} = -1$$

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:

Calculation Table

		Rank of Marks in			
Marks in Statistics	Marks in Commerce	Statistics X	Commerce Y	$d_i = X_i - Y_i$	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 $\Sigma d_i^2 = 16$

 Rank Correlation $r = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6\times16}{10\times99} = 0.903$

10.4 Chi-square Test

When the null hypothesis is true, a chi-square test (also chi-squared or x^2 test) is any statistical hypothesis test in which the test statistic's sampling distribution is a chi-square distribution, or any in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to estimate a chi-square distribution as closely as desired.

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.

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?

Problem:

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 $200 \times 0.7 = 140$.

Those who do not drink that brand are $200 \times 0.3 = 60$

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 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.

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.
- 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.
- Calculation of rank correlation coefficient in different situations

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.

Spearman's Rank Correlation: The Spearman's rank coefficient of correlation is a nonparametric measure of rank correlation. It measures the strength and direction of the association between two ranked variables.

Observed Frequency: The frequency actually obtained from the performance of an experiment.

Contingency table: A table having rows and columns where in each row corresponds to a level of one variable and each column to a level of another variable.

Self Assessment

1. The correlation coefficient is used to determine:

A. A specific value of the y-variable given a specific value of the x-variable

B. A specific value of the x-variable given a specific value of the y-variable

C. The strength of the relationship between the x and y variables

D. None of these

2. The coefficient of correlation

A. is the square of the coefficient of determination
B. is the square root of the coefficient of determination
C. is the same as r-square
D. can never be negative

3. When the values of two variables move in the opposite directions, correlation is said to be .....

A. Linear
B. Non-linear
C. Positive
D. Negative

4. When the values of two variables move in the opposite directions, correlation is said to be .....

A. Linear
B. Non-linear
C. Positive
D. Negative

5. Rank correlation coefficient was discovered by.....

A. Fisher
B. Spearman
C. Karl Pearson
D. Bowley

6. Spearman's Rank Correlation Coefficient is usually denoted by.....

A. K
B. A
C. S
D. R

7. Study of correlation among three or more variables simultaneously is called.....

A. Partial correlation
B. Multiple correlation
C. Nonsense correlation
D. Simple correlation

8. Which of these distributions is used for a testing hypothesis?

A. Normal Distribution
B. Chi-Squared Distribution
C. Gamma Distribution

D. Poisson Distribution

9. The Variance of Chi Squared distribution is given as k.

A. True
B. False

10. On account of simple calculation involved, $\chi 2$ test is very frequently used by the statistician.

A. True
B. False

11-Zero correlation coefficient between two variables could mean
A. The variables are non-linearly related to each other
B. There is a cause and effect relationship between variables
C. That there is error of measurement in variables
D. None of the above is true

12-If all the scatter of points on two variables lie on a negatively stopped straight line, the correlation coefficient between the variables would be
A. +1
B. -1
C. Zero
D. None of the above

13-A positive and a negative relationship may have the same strength.

A. True
B. False

14-A non-directional hypothesis predicts a negative correlation between two variables.
A. True
B. False

15-Spearman's rank order correlation is a parametric statistic.

A. True
B. False

Answers for Self Assessment

1.	C	2.	B	3.	D	4.	D	5.	C
6.	D	7.	B	8.	B	9.	B	10.	A
11.	A	12.	B	13.	A	14.	B	15.	B

Review Questions

1. Show that the coefficient of correlation, r, is independent of change of origin and scale.

2.Prove that the coefficient of correlation lies between – 1 and + 1.

3. What is Spearman's rank correlation? What are the advantages of the coefficient of rank correlation over Karl Pearson's coefficient of correlation?

4.What can you conclude on the basis of the fact that the correlation between body weight and annual income were high and positive?

5. From the data given below,

X	52	60	58	39	41	53	47	34
Y	40	46	43	54	49	55	48	57

find out Karl Pearson's coefficient of correlation.

1. Suppose we have ranks of 8 students of B.Sc. in Statistics and Mathematics. On the basis of rank we would like to know that to what extent the knowledge of the student in Statistics and Mathematics is related.

Rank in Statistics	52	60	58	39	41	53	47	34
Rank in Mathematics	40	46	43	54	49	55	48	57

2. Enumerate the steps in chi-square calculation.

3. In a study equal number of boys and girls were asked to express their preference for lecture method and discussion method. The data are given below:

Students	Preferred Lecture Method	Preferred Discussion Method
Boys	31	19
Girls	24	16

Further Readings

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

Web Links

https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

https://conjointly.com/kb/correlation-statistic/

https://www.youtube.com/watch?v=4EXNedimDMs

https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Unit11:Analysis of Variance (ANOVA) and Prediction Techniques

CONTENTS

Objectives

After studying this unit, you will be able to:

- Explain the Concept of Analysis of Variance (ANOVA)
- Discuss reliability and validity
- Define the bivariate regression
- Carry outmultiple regression analysis

Introduction

ANOVA stands for "analysis of variance," and it's a statistical technique for testing a hypothesis and determining how various groups react to one another by connecting independent and dependent variables. ANOVA is a statistical test that compares the means of two groups to see if there is a difference between them.It is an advanced technique for the experimental treatment of testing differences among all of the means.

The ANOVA technique allows us to do this simultaneous test and is thus regarded as a valuable analytical tool in the hands of a researcher. Using this method, one can estimate if the samples weretaken from populations with the same mean.

Regression analysis is a proven method for determining which variables have an impact on a certain subject. Regression analysis allows you to confidently establish which elements are most important, which factors may be ignored, and how these factors interact. Data is at the heart of regression analysis. It aids businesses in comprehending the data they have and using it – specifically, the correlations between data points – to make better decisions, ranging from sales forecasting to inventory levels and supply and demand analysis. Regression analysis is frequently referred to as one of the most important business analysis approaches.

11.1 Analysis of variance(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.

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. The difference in sample means is not significant when the calculated value of F is less than the table value of F, and the null hypothesis is accepted. When the estimated value of F is greater than the critical value of F, on the other hand, the difference in sample means is regarded 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. For a new toffee box introduced by Nutrine Company, three prices are explored. The price of three different types of toffee boxes is 39, 44, and 49 dollars. The goal is to figure out how price levels affect sales. These toffee boxes will be shown in five supermarkets. The sales are as follows:

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., x1, x2 and x3, 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., $\bar{\bar{X}}\bar{X} = \frac{\bar{X}_1+\bar{X}_2+\bar{X}_3}{K}$

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$= \Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2i} - \bar{x}_2)^2 + \Sigma(x_{3i} - \bar{x}_3)^2$

The sum of squares (ss) for variance between samples is calculated by subtracting the sample mean deviations from the mean of sample means () and computing the squares of such deviations, which are then multiplied by the number of items or categories in the samples to get their total. The sum of squares (ss) for variance within samples is calculated by subtracting all sample item values from their respective sample averages, squaring the deviations, and then adding them together. For our illustration then

ss between $= 4(5-4)^2 + 4(4-4)^2 + 4(3-4)^2$

$= 4 + 0 + 4 = 8$

ss within $= \frac{\{(5-5)^2 + (6-5)^2 + (2-5)^2 + (7-5)^2\}}{\Sigma(x_{1i} - \bar{x}_1)^2} + \frac{\{(4-4)^2 + (4-4)^2 + (2-4)^2\}}{\Sigma(x_{2i} - \bar{x}_2)^2}$

$+ \frac{\{(4-3)^2 + (3-3)^2 + (2-3)^2 + (3-3)^2\}}{\Sigma(x_{3d} - \bar{x}_3)^2}$

$= (0 + 1 + 9 + 4) + (0 + 0 + 4 + 4) + (1 + 0 + 1 + 0)$

$= 14 + 8 + 2$

$= 24$

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:

$$\Sigma\left(x_{ij} - \bar{\bar{x}}\right)^2$$

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

2. Two-way ANOVA

The approach for calculating variance is identical to that used for one-way classification. The following is an example of ANOVA two-way classification: Assume a company has four different types of machines: A, B, C, and D. It has placed four of its employees on each machine for a given amount of time, such as one week. The average production of each worker on each type of machine was calculated at the end of one week. These data are given below:

Average Production by the MachineType

	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

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.

Following steps are followed.

1 Calculate Sample mean i.e. $\bar{x}$

2 Calculate General mean i.e. $\bar{x}$

3 Calculate variance between columns using the formula $\bar{\sigma}^2 = \frac{\sum n_i(x_i-\bar{x})^2}{k-1}$ where $K = (n_1 + n_2 + n_3 - 3)$

4 Calculate sample variance. It is calculated using formula:
Sample variance $s_i^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$ 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 = $S(n_1 - k)$

9 Refer to F table $f8$ find value.

10 Draw conclusions.

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

$$= \frac{15 + 18 + 19 + 22 + 11 + 22 + 27 + 18 + 21 + 17 + 24 + 19 + 16 + 22 + 15 + 18}{16}$$

$$= \frac{304}{16} = 19$$

3 Calculate variance between columns:

n	$\bar{x}$	$\bar{\bar{x}}$	$\bar{x} - \bar{\bar{x}}$	$(\bar{x} - \bar{\bar{x}})^2$	$n(\bar{x} - \bar{\bar{x}})^2$
5	17	19	-2	4	5 × 4 = 20
5	21	19	2	4	5 × 4 = 20
6	19	19	0	0	6 × 0 = 0
				$\sum n_i(\bar{x}_i - \bar{\bar{x}})^2$	= 40

$$\bar{\sigma}^2 = \frac{\sum n_i(x_i - x)^2}{k - 1} = \frac{40}{3 - 1} = 20$$

4. Calculation sample variance:

Training method -1		Training method -2		Training method -3	
$x - \bar{\bar{x}}$	$(x - \bar{\bar{x}})^2$	$x - \bar{\bar{x}}$	$(x - \bar{\bar{x}})^2$	$x - \bar{\bar{x}}$	$(x - \bar{\bar{x}})^2$
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	$(-4)^2 = 16$
	$\sum(x - \bar{\bar{x}})^2 = 70$		$\sum(x - \bar{\bar{x}})^2 = 62$		$\sum(x - \bar{\bar{x}})^2 = 60$

Sample variance $= \frac{\sum(x-\bar{x})^2}{n-1} = \frac{70}{5-1}, \frac{\sum(x-\bar{x})^2}{n-1} = \frac{62}{5-1}, \frac{\sum(x-\bar{x})^2}{n-1} = \frac{60}{5-1}$

$$s_1^2 = \frac{70}{4} = 17.5, s_2^2 = \frac{62}{4} = 15.5, s_3^2 = \frac{60}{5} = 12$$

5. Within column variance $\bar{\sigma}^2 = \sum\left(\frac{n_i - 1}{n_i - k}\right) s_1^2$

$$= \left(\frac{5-1}{16-3}\right) \times 17.5 + \left(\frac{5-1}{16-3}\right) \times 15.5 + \left(\frac{6-1}{16-3}\right) \times 12$$

$$= \left(\frac{4}{13}\right) \times 17.5 + \left(\frac{4}{13}\right) \times 15.5 + \frac{5}{13} \times 12$$

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= $(3 - 1) = 2$.

11.2 Reliability and Validity

There are two criteria to decide whether the scale selected is good or not. They are:

1. Reliability

2. Validity

Reliability Analysis

The degree to which the measurement method is error-free is referred to as reliability. Accuracy and consistency are two aspects of reliability. If the scale produces the same findings when repeated measurements are taken under the same conditions, it is said to be reliable.

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.

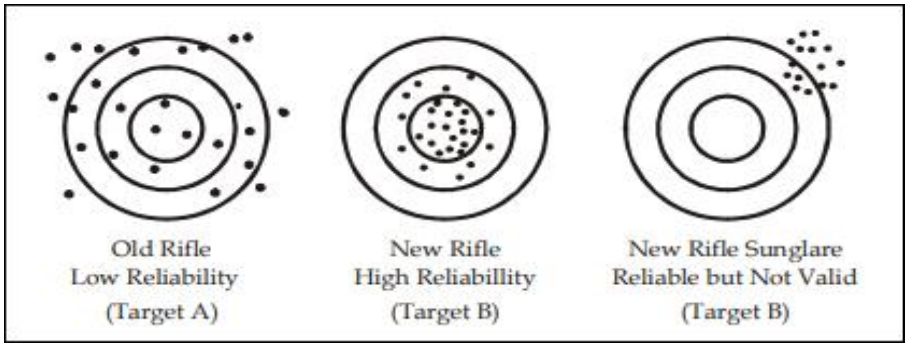
Validity Analysis

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 feels that there is a direct link between job satisfaction and the degree to which a person is an extrovert, as well as the sales force's performance. As a result, those who have high job satisfaction and outgoing personalities should perform well. If they don't, the measure's construct validity is called into question.
2. **Content Validity:**The problem should be clearly defined by the researcher. Determine the object to be measured. Create a scale that is appropriate for this purpose. Regardless of these factors, the scale may be criticised for its lack of content validity. Face validity is another term for content validity. The advent of new packaged foods is one example. When a new packaged food is introduced, it represents a significant change in flavour. Hundreds of thousands of people may be urged to try the new packaged meals. People may report that they liked the new flavour overwhelmingly. Even with such a positive response, the product may nevertheless fail when it is launched on a commercial basis. So, what's the issue? Perhaps a vital question was overlooked.

3. **Predictive Validity**: This pertains to "How best a researcher can guess the future performance from the knowledge of attitude score"?
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.



11.3 Regression Analysis

Regression is often put into two- bivariate and multiple regression analysis

Bivariate Regression

Bivariate Regression, often known as simple regression analysis, is a technique for determining the strength of a relationship between two variables. The two variables are commonly referred to as X and Y, with one acting as an independent (or explanatory) variable and the other as a dependent variable (or outcome variable).

Bivariate Regression Analysis employs a linear regression line (since the relationship between the variables is considered to be linear) to help measure how the two variables change together in order to establish the relation.

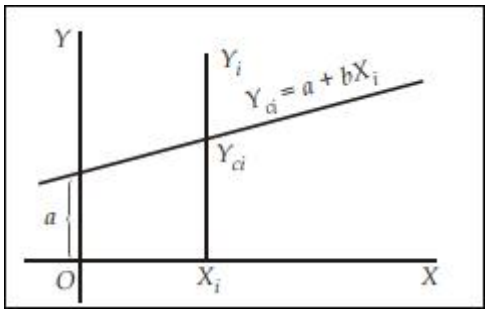
For a bivariate data (Xi, Yi), i = 1, 2, ... n, we can have either X or Y as independent variable. If X is independent variable then we can estimate the average values of Y for a given value of X. The relation used for such estimation is called regression of Y on X. If on the other hand Y is used for estimating the average values of X, the relation will be called regression of X on Y. For a 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.

The general form of the line of regression of Y on X is YCi = a + bXi, where YCi denotes the average or predicted or calculated value of Y for a given value of X = Xi. 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 (Xi, Yi), i = 1, 2, ... n.

Notes: It is important to distinguish between YCi and Yi. Whereas Yi is the observed value, YCi is a value calculated from the regression equation.

Using the regression YCi = a + bXi, we can obtain YC1 , YC2 , ... YCn corresponding to the X values X1 , X2 , ... Xn respectively. The difference between the observed and calculated value for a particular value of X say Xi is called error in estimation of the ith 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 = Yi – YCi (i = 1, 2,.....n), the error in estimation of the ith 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= $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - Y_{Ci})^2$



The regression line can, alternatively, be written as a deviation of Yi from YCi i.e. Yi – YCi = e i or Yi = YCi + e i or Yi = a + bXi + e i. The component a + bXi is known as the deterministic component and ei 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 S = S(Yi – a – bXi)2 (Θ YCi = a + bXi).

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 $$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n}(Y_i - a - bX_i) = 0$$

or $$\sum_{i=1}^{n}(Y_i - a - bX_i) = \sum_{i=1}^{n} Y_i - na - b\sum_{i=1}^{n} X_i = 0$$

or $$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i \quad \text{.... (1)}$$

Also, $$\frac{\partial S}{\partial b} = 2\sum_{i=1}^{n}(Y_i - a - bX_i)(-X_i) = 0$$

or $$-2\sum_{i=1}^{n}(X_iY_i - aX_i - bX_i^2) = \sum_{i=1}^{n}(X_iY_i - aX_i - bX_i^2) = 0$$

or $$\sum_{i=1}^{n} X_iY_i - a\sum_{i=1}^{n} X_i - b\sum_{i=1}^{n} X_i^2 = 0$$

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 YCi = a + bXi, 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 $(\bar{X}, \bar{Y})$.

From equation (3), we have $a = \bar{Y} - b\bar{X}$ .... (4)

Substituting this value of a in equation (2), we have

$$S X_i Y_i = \left(\bar{Y} - b\bar{X}\right)\Sigma X_i + b\Sigma X_i^2$$

$$= \bar{Y}\sum X_i - b\bar{X}\sum X_i + b\sum X_i^2 = n\bar{X}\bar{Y} - bn\bar{X}^2 + b\sum X_i^2$$

or $$\sum X_i Y_i - n\bar{X}\bar{Y} = b\left(\sum X_i^2 - n\bar{X}^2\right)$$

or $$b = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} \quad \text{.... (5)}$$

Also, $$\sum X_i Y_i - n\bar{X}\bar{Y} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

and $$\sum X_i^2 - n\bar{X}^2 = \sum (X_i - \bar{X})^2$$

$\therefore$ $$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad \text{.... (6)}$$

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

$$b = \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum (X_i - \bar{X})^2} = \frac{Cov(X,Y)}{\sigma_X^2} \quad \text{.... (8)}$$

The expression for b, which is convenient for use in computational work, can be written from equation (5) is given below:

$$b = \frac{\sum X_i Y_i - n\frac{\sum X_i}{n}\cdot\frac{\sum Y_i}{n}}{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2} = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

Multiplying numerator and denominator by n, we have

$$b = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n\sum X_i^2 - (\sum X_i)^2} \quad \text{.... (9)}$$

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 $$(X_i - \bar{X}) = h(u_i - \bar{u})$$

and $$Y_i - \bar{Y} = k(v_i - \bar{v})$$

Substituting these values in equation (6), we have

$$b = \frac{hk\sum(u_i-\bar{u})(v_i-\bar{v})}{h^2\sum(u_i-\bar{u})^2} = \frac{k\sum(u_i-\bar{u})(v_i-\bar{v})}{h\sum(u_i-\bar{u})^2}$$

$$= \frac{k}{h}\left[\frac{n\sum u_iv_i-(\sum u_i)(\sum v_i)}{n\sum u_i^2-(\sum u_i)^2}\right] \quad \text{.... (10)}$$

(**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 $$Y_{Ci} = \bar{Y} - b\bar{X} + bX_i \quad \text{or} \quad Y_{Ci} - \bar{Y} = b(X_i - \bar{X}) \quad \text{.... (11)}$$

or $$(Y_{Ci} - \bar{Y}) = r\cdot\frac{\sigma_Y}{\sigma_X}(X_i - \bar{X}) \quad \text{.... (12)}$$

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.

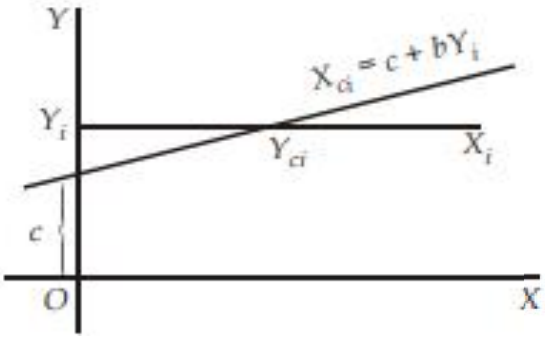
In this case, we have to calculate the value of c and d so that

$S' = (X_i - X_{Ci})^2$ 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 $$\Sigma X_iY_i = c\Sigma Y_i + d\Sigma Y_i^2 \quad \text{.... (14)}$$



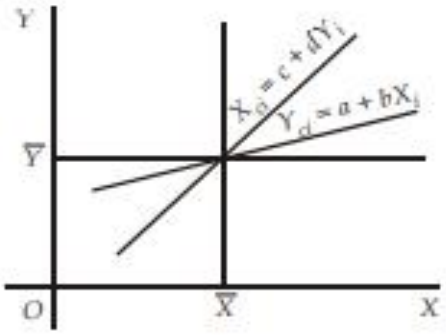
Dividing both sides of equation (13) by n, we have $\bar{X} = c + d\bar{Y}$

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.

$$d = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum Y_i^2 - n\bar{Y}^2} \quad \text{.... (16)}$$



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 $v_i = \frac{Y_i - B}{k}$

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

$$X_{\odot} = \bar{X} - d\bar{Y} + dY_i \text{ or } \left(X_{\odot} - \bar{X}\right) = d\left(Y_i - \bar{Y}\right) \quad \text{.... (23)}$$

or

$$\left(X_{\odot} - \bar{X}\right) = r \cdot \frac{\sigma_X}{\sigma_Y}\left(Y_i - \bar{Y}\right) \quad \text{.... (24)}$$

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 minimisingS(Yi – YCi) 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(Xi – XCi) 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 = ± 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 = ± 1 in equation (12) or (24) it can be shown that the lines of regression in both the cases become.

$$\left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right) \approx \pm\left(\frac{X_i - \bar{X}}{\sigma_X}\right)$$

Further when $r = 0$, equation (12) becomes $Y_{\odot} = \bar{Y}$ and equation (24) becomes $X_{\mathrm{C}} = \bar{X}$. 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.

Multiple Regression Analysis

Multiple Regression Analysis is an extension of two variable regression analysis. In this analysis, two of more independent variables are used to estimate the values of a dependent variable, instead of one independent variable.
The objective of multiple regression analysis are:

1. To derive an equation which provides estimates of the dependent variable from values of the two or more variables independent variables.
2. To obtain the measure of the error involved in using the regression equation as a basis of estimation.
3. To obtain a measure of the proportion of variance in the dependent variable accounted for or explained by the independent variables.

Multiple regression equation explains the average relationship between the given variables and the relationship is used to estimate the dependent variable. Regression equation refers the equation for estimating a dependent variable.

Example: Estimating dependent variable X_1 from the independent variablesX_2, X_3 ... ...

it is known as regression equation of X_1on $X_2 x_2$ ... ...Regression equation, when three variables are involved, is given below:

$$x_{123} = a_{122} + b_{123}x_2 + b_{13.2}x_3$$

Where $X_{1.23}$ =estimated value of the dependent variable X_2X_3 =independent variables. $a_{1\Omega}$ = (Constant) the intercept made by the regression plan it gives the value of the dependent variable, when all the independent variables assume a value equal to zero.

b_{122} and b_{132} = Partial regression coefficients or net regression coefficients. $b_{1.zs}$ = measures the amount by which a unit change in X_2 is expected to aflect

X_1 when X_3 is held constant.

Deviations Taken From Actual Means

$$x_{1.23} = b_{1.23}x_2 + b_{132}x_3$$
$$x_1 = (x_1 - \bar{x}_1)$$
$$x_2 = (x_2 - \bar{x}_2)$$
$$x_3 = (x_3 - \bar{x}_3)$$

b_{123} and b_{132}can be obtained by solving the following equations.

$$\Sigma X_1X_2 = b_{123}x_2^{\ 2} + b_{132}x_2x_3$$
$$\Sigma X_1X_2 = b_{123}\Sigma x_2x_3 + b_{132}\Sigma x_3$$
$$b_{12.3} = \frac{\sigma_{1.23}}{\sigma_{3.12}}$$
$$(x_1 - \bar{x}_1) = \left[\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^{\ 2}}\right]\left[\frac{s_1}{s_2}\right](x_2 - \bar{x}_2) + \left[\frac{r_{12} - r_{13}r_{28}}{1 - r_{23}^{\ 2}}\right]\left[\frac{s_1}{S_3}\right](x_3 - \bar{x}_3)$$

Regression equation of X_3 and X_2 and X_1 is:

$$(x_3 - x_3) = \left[\frac{r_{23} - r_{13}r_{12}}{1 - r_{23}^2}\right]\left[\frac{s_3}{s_2}\right](x_2 - \bar{x}_2) + \left[\frac{r_{13} - r_{23}r_{12}}{1 - r_{23}^2}\right]\left[\frac{s_3}{s_1}\right](x_1 - \bar{x}_1)$$

Summary

- ANOVA is a technique of statistics and it is applied to test the equality of three or more sample means.
- It is an advanced technique for the experimental treatment of testing differences among all of the means.
- The ANOVA allows to do this simultaneous test and is thus considered as a valuable analytical tool in the hands of a researcher.
- Regression analysis allows you to confidently establish which elements are most important, which factors may be ignored.
- Regression is a term used for predicting the value of one variable from the other.
- Least square method is used to fit the line.

Keywords

ANOVA: It is a statistical technique used to test the equality of three or more sample means.

Bivariate Regression: a technique for determining the strength of a relationship between two variables.

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.

Reliability Analysis:the extent to which the measurement process is free from errors.

Internal Consistency in Reliability: The consistency with which each item represents the construct of interest.

Validity Analysis: means "Are we measuring, what we think, we are measuring?"

Self Assessment

1-Analysis of variance is a statistical method of comparing the _____ of several populations.

A. standard deviations
B. variances
C. means
D. proportions

2- The one-way ANOVA is used to test statistical hypotheses concerning:

A. Variances
B. Group Means
C. Standard Deviations
D. None of these

3- ANOVAs cannot be used when testing data collected in educational research as it cannot be applied to social science.

A. True
B. False

4- What type of data are best analysed in ANOVA?

A. Correlational
B. Random
C. Experimental
D. Simple

5-What is the definition of 'mean square'?

A. A sum of squares divided by its degrees of freedom
B. The square root of the mean
C. The square of the mean
D. A table of means with four cells

6-In regression, the equation that describes how the response variable (y) is related to the explanatory variable (x) is:

A. the correlation model
B. the regression model
C. used to compute the correlation coefficient
D. None of these alternatives is correct.

7- In regression analysis, the variable that is being predicted is the

A. response, or dependent, variable
B. independent variable
C. intervening variable
D. is usually x

8- Regression equation is also named as

A. predication equation
B. estimating equation
C. line of average relationship
D. all the above

9- Regression coefficient is independent of

A. origin
B. scale
C. both origin and scale
D. neither origin nor scale.

10- The regression analysis measures _____ between X and Y.
A. Dependence
B. Independence
C. Both a & b
D. None

11- The _____ sum of squares measures the variability of the sample treatment means around the overall mean.
A. treatment
B. error
C. interaction
D. total

12- Which of the following is an assumption of one-way ANOVA comparing samples from three or more experimental treatments?
A. All the response variables within the k populations follow a normal distributions.
B. The samples associated with each population are randomly selected and are independent from all other samples.
C. The response variables within each of the k populations have equal variances.
D. All of the above.

13- As variability due to chance decreases, the value of F will

A. increase
B. stay the same
C. decrease
D. can't tell from the given information

14-What do ANOVA calculate?

A. F ratios
B. Z Scores
C. T Scores
D. None

15- Which of the following assumptions must be met to use an ANOVA?

A. There is homogeneity of variance
B. The dependent variable must be interval or ratio
C. The data must be normally distributed
D. All of these

Answers for Self Assessment

1.	C	2.	B	3.	B	4.	C	5.	A
6.	C	7.	A	8.	D	9.	A	10.	A
11.	A	12.	D	13.	A	14.	A	15.	D

Review Questions

1. What do you think as the reason behind the two lines of regression being different?
2. From the data given below:-

X	52	60	58	39	41	53	47	34
Y	40	46	43	54	49	55	48	57

and find out the following:

i) The two regression equations.
ii)The most likely value of X when Y = 41.
iii)The most likely value of Y when X = - 45.

3-Obtain the equations of the two lines of regression for the data given below:

X	45	42	44	43	41	45	43	40
Y	40	38	36	35	38	39	37	41

4-In the estimation of regression equation of two variables X and Y the following results were obtained. X = 90, Y = 70, n = 10, Ìx 2 =6360; Ìy 2 = 2860, Ìxy = 3900 Obtain the two regression equations.

5- A test was given to five students taken at random from the fifth class of three schools of a town. The individual scores are

School I	9	7	6	5	8
School II	7	4	5	4	5
School III	6	5	6	7	6

Carry out the analysis of variance

6- Three varieties of coal were analysed by four chemists and the ash-content in the varieties was found to be as under.

	Chemists			
Varieties	1	2	3	4
A	8	5	5	7
B	7	6	4	4
C	3	6	5	4

Carry out the analysis of variance.

7-What is analysis of variance?

8. Distinguish between t-test for difference between means and ANOVA.
9. What is multiple regression? How does it differ from bivariate regression?

FurtherReadings

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.

A Parasuraman, Dhruv Grewal, Marketing Research, Biztantra

Paneerselvam, R, Research Methods, PHI.

Web Links

https://www.youtube.com/watch?v=TKom54uOzXY

https://murraylax.org/rtutorials/regression_intro.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/

https://sciencing.com/difference-between-bivariate-multivariate-analyses-8667797.html

Unit12: 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.

12.1 Multivariate Analysis

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.

12.2 Classification

Multiple-variate analysis: This can be classified under the following heads:

A. Factor Analysis
B. Cluster Analysis
C. Discriminant Analysis
D. Multidimensional Scaling
E. Conjoint Analysis

12.3 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.

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:

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 you 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-dimensionalspace, and one could compute the (Euclidean) distances between the 5 row points in the 4dimensional 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).

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 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 maximizing 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.

12.4 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, lifestyles 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.

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:

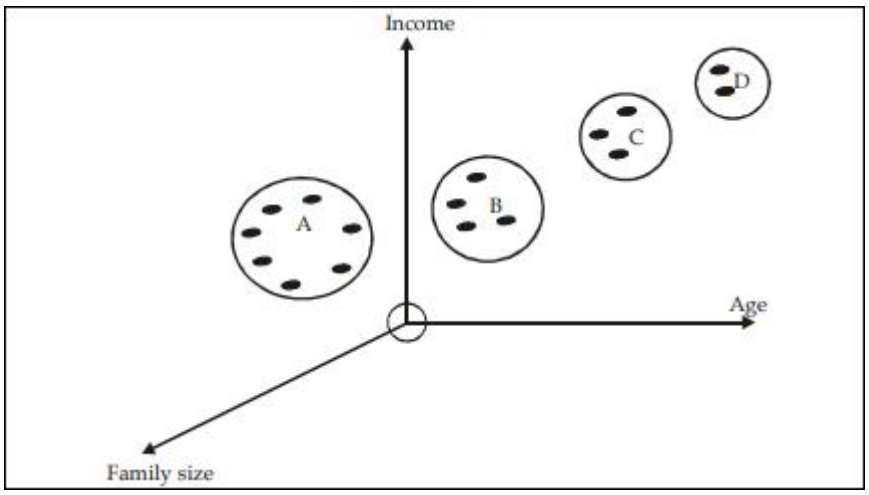
1. Selection of the sample to be clustered (buyers, products, employees).
2. Definition on which the measurement to be made (E.g.: product attributes, buyer characteristics, employees' qualification).
3. Computing the similarities among the entities.
4. Arrange the cluster in a hierarchy.
5. Cluster comparison and validation.

Did you know?

Names can also be given to clusters to describe each one. For example, there can be a cluster called "neo-rich". Segments are prioritized 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, and 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.



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.

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.
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.

12.5 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 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.

$$Z = b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots\ldots\ldots$$

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

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.

12.6 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 nonmetric model has 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:

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.

12.7 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 traveler: (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.

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)

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, 2 = 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 leastpreferred.

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.

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.
- 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.

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.

Self Assessment

1-In discriminant analysis the averages for the independent variables for a group define the

A. centroid
B. median
C. mode
D. central tendency

2____ is a method for deriving the utility values that consumers attach to varying levels of a product's attributes

A. Regression
B. Conjoint analysis
C. Correlation
D. T test

3-The conjoint analysis procedure is based on trade-offs respondents make when evaluating alternatives.

A. True
B. False

4-What is the idea behind conjoint analysis?

A. Understanding Consumer Preferences
B. Manufacturing process determination
C. Building Social Media
D. Ignoring Price Increase

5-What is the first step in setting up a conjoint analysis?

A. Using data to improve your products
B. Choosing features, functions or attributes
C. Asking consumers to choose their top features
D. Collecting responses from consumers

6-In Conjoint Analysis, responses are collected from……

A. Researchers
B. Industries
C. Marketers
D. Consumers

7-Conjoint Analysis is a …… technique which is used to determine customers' preferences:

A. Descriptive
B. Predictive
C. Inferential
D. None of These

8-Factor analysis requires that variables:

A. Are measured at nominal level
B. Are abstract concepts
C. Are not related to each other
D. Are related to each other

9-Factor analysis is a(n) _____ in that the entire set of interdependent relationships is examined.

A. KMO measure of sampling adequacy
B. orthogonal procedure
C. interdependence technique
D. varimax procedure

10-_____ are simple correlations between the variables and the factors.

A. Factor scores
B. Factor loadings
C. Correlation loadings
D. Both a and b are correct

11-A factor can be considered to be an underlying latent variable:

A. on which people differ
B. that is explained by unknown variables
C. that cannot be defined
D. that is influenced by observed variables

12-The decision about how many factors to retain is based on:

A. personal choice
B. Kaiser's rule
C. Scree test
D. Both Kaiser's rule and Scree test

13-The goal of clustering is to-

A. Divide the data points into groups
B. Classify the data point into different classes
C. Predict the output values of input data points
D. All of the above

14-Which of the following is a bad characteristic of a dataset for clustering analysis-

A. Data points with outliers
B. Data points with different densities
C. Data points with non-convex shapes
D. All of the above

15-On which data type, we cannot perform cluster analysis

A. Time series data
B. Text data
C. Multimedia data
D. None

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 summarize 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. 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.

12. 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.

13. Six observations on two variables are available, as shown in the following table:

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 for Self Assessment

1.	A	2.	B	3.	A	4.	A	5.	B
6.	D	7.	B	8.	C	9.	C	10.	B
11.	A	12.	D	13.	A	14.	D	15.	D

Further Readings

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

Web Links

https://www.qualtrics.com/au/experience-management/research/factor-analysis/

https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis

https://www.qualtrics.com/au/experience-management/research/cluster-analysis/

https://www.statisticshowto.com/multidimensional-scaling/

https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Multidimensional_Scaling.pdf

Unit 13: Reporting a Quantitative Study

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

A report is a formal document written for a number of objectives in the sciences, social sciences, engineering, and business fields. The findings of a specified or specific task are usually written up in a report. It's worth noting that reports are seen as legal papers in the workplace, therefore they must be exact, accurate, and difficult to misunderstand.

At its most fundamental level, report writing is defined by three characteristics: a set framework, independent parts, and achieving 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.

13.1 Significance of Report Writing

The most significant component of the research process is the preparation and presentation of a research report. No matter how brilliant the idea or well-designed the research study is, it is useless unless it is properly communicated to others in the form of a research report. Furthermore, time and effort spent gathering and analyzing data will be wasted if the report is ambiguous or badly worded. As a result, it's critical to summarize and explain the findings to management in the form of a logical and clear study report.

The research assignment is not complete until the report has been presented and/or written, hence it is considered a major component of the research project. Even the most brilliant idea, well-designed and executed research study, and startling generalizations and conclusions are of limited value unless they are properly communicated to others. The purpose of study isn't served adequately until the results are disseminated. In most cases, research findings must be added to the overall body of knowledge. All of this explains why producing a research report is so important. Some people do not consider report writing to be an important aspect of the research process. However, the common consensus is that the presenting of research findings or the preparation of a report should be considered part of the research endeavor. The writing of a report is the final phase of a research project, and it necessitates a set of skills that are distinct from those required during the other stages of research. This task should be carried out with utmost caution by the researcher; he may seek aid and instruction from professionals for this reason.

13.2 Techniques and Precautions of Interpretation

The term "interpretation" refers to the process of extracting the meaning from data. Interpretation can also be defined as the process of converting facts into information. The essence of any research project is to interpret the findings. This necessitates a high level of expertise. In order to reach a conclusion, you must use one of two methods: (i) induction or (ii) deduction.

In the induction approach, one begins with observed facts and then applies generalization to explain the relationship between the observed items.

Deductive reasoning, on the other hand, begins with a general law and is then applied to a specific case, i.e., deduction begins with the general and ends with the specific.

Example of Induction: All products manufactured by Sony are excellent. DVD player model 2602 MX is made by Sony. Therefore, it must be excellent.

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.

We argue from observation during the inductive phase. We reason towards the observation during the deductive phase. The quality of the data analysis determines the success of the interpretation. The interpretation of data that has not been adequately analyzed may be incorrect. If an analysis needs to be corrected, then good data collecting is required. Similarly, if the data is correct but the analysis is incorrect, the interpretation or conclusion will be incorrect as well. Even with good data and analysis, the data might sometimes lead to incorrect interpretation. The researcher's experience and the methods he uses for interpretation play a role in the interpretation.

Did you 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.

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".

Basic Analysis of "Qualitative" Information

(Respondents' verbal answers in interviews, focus groups, or written commentary on questionnaires):

- Go over all of the information.
- Sort comments into categories based on their content, such as worries, suggestions, strengths, shortcomings, comparable experiences, programme inputs, recommendations, outputs, result indicators, and so on.
- Label the categories or themes with words like "concerns," "suggestions," and so on.
- Look for patterns, associations, and causal relationships in the themes, such as whether all people who attended evening programmes had similar concerns, whether most people were from the same geographic area, whether most people were in the same salary range, what processes or events respondents experienced during the programme, and so on.
- Save all comments for several years after they've been completed in case they're needed in the future

Interpreting Information

- Attempt to put the data into context, e.g., compare results to what you expected, promised results; management or programme staff; any common standards for your products or services; original goals (especially if you're conducting a programme evaluation); indications or measures of achieving outcomes or results (especially if you're conducting an outcomes or performance evaluation); desirability (especially if you're conducting an outcomes or performance evaluation); desirability (especially if you're conducting an outcomes or performance

- Take into account suggestions to assist employees in improving the programme, product, or service; judgments about programme operations or reaching targets, and so on.
- Write out your conclusions and recommendations in a report, together with the interpretations that support them.

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 generalizing 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".

13.3 Layout,Style and Precautions of the Report writing

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 organized 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. Particular attention should be paid to make sure there is:
ii. Particular attention should be paid to make sure there is:
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.

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.
- Verbs 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).

Notes: Comparison of data, highlighting of unexpected outcomes, your own or others' comments on problems uncovered, and balancing of pros and cons of proposed solutions are all required for the unit discussion. However, all too often, the discussion consists of a dry review of the findings.

Precautions of the Report writing

Another problem is endless description without interpretation. Conclusions are needed in tables, not a thorough presentation of all figures or percentages in the cells for readers to observe.

Qualitative data is frequently overlooked. Still, including quotes from informants to illustrate your findings and conclusions adds life to your report. They are also scientifically valuable because they allow the reader to draw his or her own conclusions from the information you provide. (Assuming your presentation is not biased!)

Sometimes qualitative data (e.g., open opinion questions) are simply categorized and counted like quantitative data, without any interpretation, despite the fact that they might provide fascinating insights into the causes behind informants' behavior or beliefs. This is a significant case of data abuse that needs to be addressed.

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.

13.4 Types of Report

There are two types of reports (1) Oral report (2) Written report.

Oral Report

When the researchers are asked to give an oral presentation, this form of reporting is essential. When compared to a written report, giving an oral presentation is more challenging. Because the reporter must communicate directly with the audience, this is the case. Any stuttering during an oral presentation can give the listeners a poor impression. The presenter's self-confidence may be lowered as a result of this. Communication is crucial in an oral presentation. To decide 'What to say,' 'How to say,' and 'How much to say,' a lot of planning and thinking is required. In addition, the presenter may be bombarded with questions from the crowd. An oral presentation necessitates much preparation; the following is a general classification.

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 script, memorized, or spoken extemporaneously? Memorization is not advised because a slip could occur during the presentation. Second, it results in a speaker-centric strategy. It is not suggested to read from the manuscript because it gets repetitive, uninteresting, and lifeless. Making main points notes so that they can be expanded is the best technique to speak in ex-tempo. Sequences should be followed in a logical manner.

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.

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.

Did you Know?

Report on children's education is a kind of government and social welfare report.

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 nontechnical persons.

Summary

- A report is a formal document written for a number of objectives in the sciences, social sciences, engineering, and business fields.
- The most important thing to remember when writing a research report is to communicate with the audience.
- The report should be able to pique the readers' curiosity. As a result, the report should be written with the reader in mind.
- Accuracy and clarity are two other factors to consider while writing a report.
- The following points should be kept in mind when giving an oral presentation: the language used, time management, graph use, report purpose, and so on. The audience must be able to understand the visuals used.
- The presenter must ensure that the presentation is finished within the allowed time. It's a good idea to set aside some time for questions and answers.
- Depending on whether the report is brief or extensive, it might be characterised as a written report. It can also be divided into two categories: technical and non-technical reports.
- The report's style should be straightforward and to the point.
- In report writing, there should not be an excessive amount of detail, and qualitative data should not be overlooked.

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.

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.

Self Assessment

1-Through interpretation, a researcher can …… relations and processes that underlie his findings.

A. Expose
B. Hide
C. Transfer
D. None of These

2-Data Analysis attempts to obtain……to research questions:

A. Answers
B. Extension
C. Problem
D. None of these

3-Which of the following does not represent the sequence from data to Knowledge?

A. From Data to Information
B. From Information to Facts
C. From Facts to Knowledge
D. From Analysis to Interpretation

4-Usefulness and utility of research findings lies in proper…….

A. Analysis
B. Interpretation
C. Findings
D. Results

5-Which of the following is NOT responsible for errors in data interpretation?

A. false generalization
B. wrong interpretation of statistical measures
C. data with consistent homogeneity
D. identification of correlation with causation

6-A Research Report is always written in a ……. language

A. Casual
B. Formal
C. Informal
D. Technical

7-A quantitative research report is……. in nature

A. Numerical
B. Technical
C. Specialized
D. Descriptive

8-Managers and marketers rely on ……as an authentic source of information.

A. Periodic Reports
B. Formal Reports

C. Long Reports

D. Business Reports

9-A research report helps to identify ……..for further inquiry

A. Structured Data
B. New Variables
C. Knowledge Gaps
D. None of These

10-In a report there must be …….in margins and spacing.

A. Consistency
B. Regularity
C. Clarity
D. None

11- Aim must be logical and ……in the report presentation

A. Organised
B. Systematic
C. Structured
D. None

12-A business report should do all of the following except:

A. Be written for a specific reader or readers

B. Take into account the reader's technical sophistication and interest in the project

C. Use technical jargon

D. Have an action /applied orientation.

13-Effective oral presentation techniques include all of the following *except* _____.

A. the use of visual aids displayed with a variety of media

B. allowing sufficient opportunity for questions, both during and after the presentation
C. not spending much time on the reason for the research and getting to the results quickly
D. constant eye contact and interaction with the audience

14-Effective oral presentation techniques include all of the following *except* _____.

A. the speaker should vary the volume, pitch, voice quality, articulation, and rate while speaking

B. terminate the presentation with a strong closing
C. the presentation should be sponsored by a top-level manager in the client's organization
D. All of the above are correct.

15-Report are often used to display the result of:

A. Experiment

B. Investigation

C. Inquiry

D. All of these

Answers for Self Assessment

1.	A	2.	A	3.	D	4.	B	5.	C
6.	B	7.	A	8.	D	9.	C	10.	A
11.	B	12.	C	13.	C	14.	D	15.	D

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. Oral presentation requires the researcher to be good public speaker explain.
9. Explain the style and layout of report.

Further Readings

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

Web Links

https://www.ets.org/Media/Research/pdf/RM-12-05.pdf

https://eduvoice.in/types-research-report-writing/

https://www.yourarticlelibrary.com/marketing/research-report-introduction-definition-and-report-format/48713

https://www.questionpro.com/blog/research-reports/

Unit 14: Writing Research Proposal

CONTENTS

Objectives

After studying this unit, you will be able to:

- Appraise the purpose of a research proposal.
- Summarise the criteria for evaluation of a research proposal.
- Comprehend contents of a research proposal
- Design a realistic, itemized budget linked to research specific objectives activities
- Underline in detail important aspects related to proposal defence.

Introduction

In order to prepare for your dissertation, you may be expected to produce a brief proposal or plan explaining your research idea and how you propose to carry it out. This is a good method to get ready for your research and it will get you thinking about a lot of the topics addressed in the next section. The proposal will ask you to demonstrate some understanding of the literature in your chosen field, in addition to stating your intended research methodology and techniques, the topic area in which your study will be conducted, and the research questions that you aim to answer.

A research proposal is a comprehensive plan, scheme, structure, and strategy for obtaining answers to your research project's study questions or challenges. A research proposal should detail the steps you propose to take to achieve your research goals, test hypotheses (if any), and answer your research questions. It should also explain why you're conducting the research. In general, the major purpose of a research proposal is to explain the operational plan for acquiring answers to your study questions. As a result, the reader is assured of the methodology's validity in finding accurate and objective responses to your research questions.

In order to achieve this function, a research proposal must tell you, your research supervisor and reviewers the following information about your study:

What you are proposing to do;

How you plan to find answers to what you are proposing;

Why you selected the proposed strategies of investigation.

14.1 Contents of a Research Proposal

A research proposal should contain the following information about your study:

An introduction, including a brief literature review;

Theoretical framework that underpins your study;

Conceptual framework which constitutes the basis of your study;

Objectives or research questions of your study;

Hypotheses to be tested, if applicable;

Study design that you are proposing to adopt;

Setting for your study;

Research instrument(s) you are planning to use;

Sampling design and sample size;

Ethical issues involved and how you propose to deal with them;

Data processing procedures;

Proposed chapters of the report;

Problems and limitations of the study;

Proposed timeframe for the project.

Proposal Presentation

Proposal Defence

Appendix

A research proposal should clearly and specifically express the aforementioned elements in such a way that anyone reading it will be able to do all tasks in the same manner as you. It should also:

-allow you to refer back to the proposal for guidance in making decisions at various stages of the research process;

-convince your research supervisor or reviewer that your proposed methodology is meritorious, valid, appropriate, and workable in terms of answering your research questions or objectives.

Preamble/introduction

The proposal should begin with an introduction, which should include some of the following details. Keep in mind that some of the content mentioned in this area may not be applicable to all studies, so choose just what is relevant to your research. The literature review is crucial when drafting this section since it serves two purposes:

1. It acquaints you with the available literature in the area of your study, thereby broadening your knowledge base.

2. It provides you with information on the methods and procedures other people have used in similar situations and tells you what works and what does not.

The kind, scope, and quality of a literature review are largely determined by the academic level for which the proposal is being written. The contents of this part may also vary significantly depending on the study topic. Begin by taking a broad view of the main subject area before eventually focusing on the central problem under consideration. Cover the following aspects of your research field when doing so.:

An overview of the main area under study;

A historical perspective (development, growth, etc.) pertinent to the study area;

Philosophical or ideological issues relating to the topic;

Trends in terms of prevalence, if appropriate;

Major theories, if any;

The main issues, problems and advances in the subject area under study;

Important theoretical and practical issues relating to the central problem under study;

The main findings relating to the core issue(s).

The Problem

After giving a comprehensive overview of the topic, concentrate on topics related to the major theme, noting some of the gaps in the existing body of knowledge. Determine some of the most important unanswered questions. Some of the primary research questions that you'd like to answer with your study should be stated here, along with a reason and relevance for each.

Knowledge gained from other studies and the literature about the issues you are proposing to investigate should be an integral part of this section. Specifically, this section should:

-Identify the issues that are the basis of your study;

-Specify the various aspects of/perspectives on these issues;

-Identify the main gaps in the existing body of knowledge;

-Raise some of the main research questions that you want to answer through your study;

-Identify what knowledge is available concerning your questions, specifying the differences of 0pinion in the literature regarding these questions if differences exist;

-Develop a rationale for your study with particular reference to how your study will fill the identified gaps.

14.2 Objectives of the Study

Include a statement of your study's major and sub objectives in this section. The main aim identifies the overall focus of your research, whilst the sub objectives specify the specific issues you want to investigate.

The study's objectives should be defined clearly and specifically. Each sub-goal should focus on a single topic. In establishing sub goals, use action verbs like 'to decide,' 'to find out,' and 'to ascertain,' which should be numerically stated. If the goal is to test a hypothesis, the precise objectives must be written using the hypothesis formulation standard.

The statement of objectives in qualitative investigations is less specific than in quantitative studies. In qualitative research, you should simply state the study's overarching goal, as your goal is to learn as much as possible as you proceed. As you may be aware, qualitative research's strength lies in its adaptability and capacity to absorb new ideas while gathering data. In qualitative research, structured statements that limit you to a preconceived framework of exploration are not a recommended convention. Exploring 'what does it mean to have a child with ADHD in the family?', 'how does it feel to be a victim of domestic violence?', 'how do people cope with racial discrimination?', 'the relationship between resilience and yoga?', or ‘reconstructing life after bushfire' are sufficient statements to communicate your qualitative research's intent of objectives. If necessary, more specific goals can be set.

14.3 Study Design

Describe how you propose to answer your research questions using the study design you've chosen. (Explain whether the design is a case study, descriptive, cross-sectional, before-and-after, experimental, or non-experimental.) Determine the advantages and disadvantages of your research plan.

Include information on the numerous logistical techniques you plan to use to carry out the study design. One of the characteristics of a good research design is that it clearly explains the details so that anyone else who wishes to follow the recommended process can do as exactly as you did.Your study design should include information about the following:

Who makes up the study population?

Can each element of the study population be identified? If yes, how?

Will a sample or the total population be studied?

How will you get in touch with the selected sample?

How will the sample's consent to participate in the study be sought?

How will the data be collected (e.g. by interview, questionnaire or observation)?

In the case of a mailed questionnaire, to what address should the questionnaire be returned?

Are you planning to send a reminder regarding the return of questionnaires?

How will confidentiality be preserved?

How and where can respondents contact you if they have queries?

The Setting

Describe the organisation, agency, or community where you will conduct your research in a few words. If the study is on a group of individuals, emphasize some of the group's most important qualities (such as its history, size, makeup, and structure) and bring attention to any relevant material that is accessible.

Include the following information in your description if your study is about an agency, office, or organisation: the agency, office, or organization's core services; its administrative structure; the types of clients served; information about the topics that are key to your research. If you're researching a community, quickly describe some of its most important aspects, such as the community's size, a social profile of the community (i.e. the make-up of the various groups within it), and difficulties related to the study's major theme.

Measurement Procedures

This section should cover your instrument as well as the specifics of how you want to operationalize your primary variables. To begin, defend your research tool selection by noting its advantages and disadvantages. Then, make a list of the important components of your research instrument and how they relate to the study's main goals. If you're using a standard instrument, talk about the evidence for its reliability and validity briefly. Describe and explain any modifications you make if you adapt or modify it in any manner.

You should also talk about how you'll put the primary themes into practice. If you're going to measure effectiveness, for example, be specific about how you'll do it. Mention the key indications of self-esteem and the techniques for measuring it if you plan to measure the self-esteem of a group of people (e.g. the Likert or Thurstone scale, or any other procedure).

Ideally, for quantitative studies you should attach a copy of the research instrument to your proposal.

Ethical Issues

Any ethical difficulties that research may have are taken seriously by all academic institutions. To cope with them, every institution has an ethics policy in place. You must be familiar with the policies of your institution. It is critical that you highlight any ethical difficulties in your proposal and explain how you plan to address them. You must consider ethical considerations from the perspective of your responders, and you must specify the process in place to deal with any potential 'damage,' whether psychological or otherwise.

Sampling

Under this section of the proposal include the following:

the size of the sample population (if known), as well as where and how this data will be gathered; the size of the sample you intend to choose, as well as your justifications for doing so; an explanation of the sampling design you want to apply in the sample selection (simple random sampling, stratified random sampling, quota sampling, etc.).

Analysis of data

Describe the data analysis strategy you propose to utilize in broad terms. Indicate whether the data will be manually or automatically analyzed. Determine the application you'll use to analyse the data and, if necessary, the statistical processes you'll use. Determine the essential variables for cross-tabulation in quantitative investigations.

Describe how you plan to analyse your interviews or observation notes in qualitative research to derive meaning from what your respondents have stated regarding the issues covered or observation notes taken. Identifying main themes by analysing the contents of the information acquired in the field is one of the most prevalent ways. You must first select whether you want to analyse this data manually or with the help of a computer programme.

Structure of the Report

As clearly as possible, state how you intend to organise the final report. In organising your material for the report, the specific objectives of your study are of immense help.

Plan to develop your chapters around the main themes of your study. The title of each chapter should clearly communicate the main thrust of its contents.

The first chapter, maybe headed 'Introduction,' should provide an overview of your research, covering the majority of your project proposal and highlighting any variations from the original plan. The second chapter should include some background information about the research population, such as socioeconomic and demographic characteristics. The primary goal of this chapter is to provide readers with some background on the demographic from which you gathered data. 'Socioeconomic–demographic characteristics of the study population', 'The study population,' or any other title that conveys this idea to readers could be used for the second chapter. The titles of the other chapters will vary from research to study, but each chapter should be organised around a central theme, as previously stated. Although the language of chapter titles is a personal preference, they must all communicate the chapter's main idea. The precise objectives of these topics were developed in the development of these themes.If your research is qualitative, the significant issues discovered throughout the data collecting and analysis stages should be used to create chapters. Following the development of key issues, the following stage is to organise the main themes within each issue and create a structure for communicating your results to your readers.

14.4 Problems and Limitations

This section should list any problems you think you might encounter concerning, for example, the availability of data, securing permission from the agency/organisation to carry out the study, obtaining the sample, or any other aspect of the study.

You will not have unlimited resources and as this may be primarily an academic exercise, you might have to do less than an ideal job. However, it is important to be aware of – and communicate – any limitations that could affect the validity of your conclusions and generalisations.

Here, *problems* refer to difficulties relating to logistical details, whereas *limitations* designate structural problems relating to methodological aspects of the study. In your opinion the study design you chose may not be the best but you might have had to adopt it for a number of reasons. This is classified as a limitation of the study. This is also true for sampling or measurement procedures. Such limitations should be communicated to readers.

Work Schedule/Timelines

Because you must do the research within a specified time range, you must give yourself dates. List the many operational processes you'll need to perform, along with the deadlines for each. Remember to set aside some time near the end as a "cushion" in case the research process does not go as planned.

- A timeline is a very important part of a research/project proposal.
- It shows the chronological order of events that a researcher plan to do in his/her project.

- It is supposed to give the reader a broad overview of the project at a glance. It does not have to be very detailed.
- A timeline presents a clear indication of the time frame for the project, Identify tasks, the times when each activity of the project will be implemented and the responsible member of the team.
- A timeline is displayed most effectively in a graphic, table or spreadsheet will help demonstrate the feasibility of the project in a very visible way.

Rationale for Research Timelines

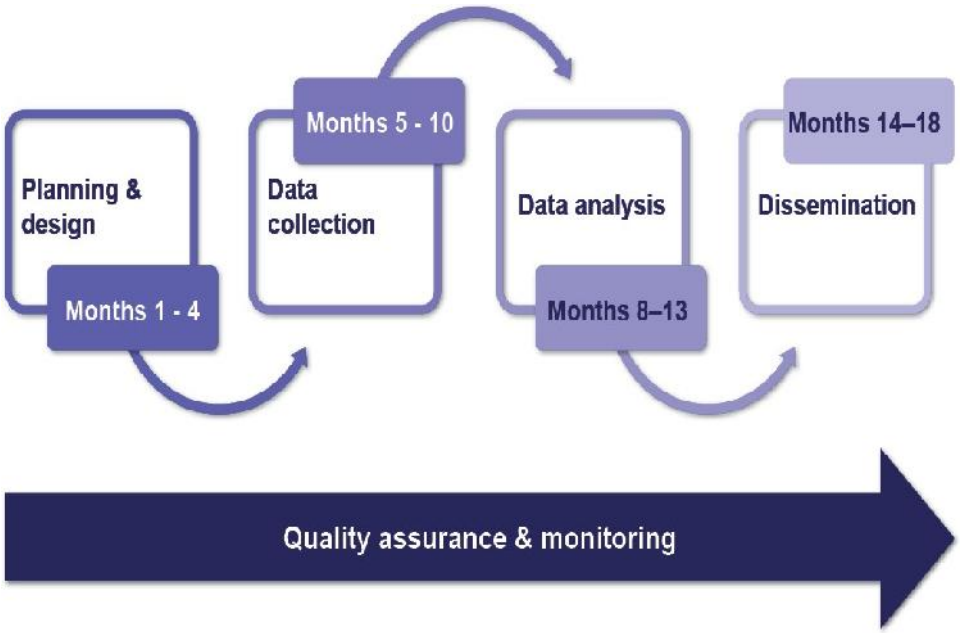
- Facilitates the development of a research focus
- Ensures consensus and ownership of the strategy and plan.
- Clarifies responsibilities and roles and how each action impacts the research as a whole
- Facilitates research monitoring and evaluation as well as identification of issues and reporting.
- Provides management/donors with key information for research review.

Timeline-Phases

- Planning phase
- Implementation phase
- Follow-through phase

Timelines-Considerations

- Timelines need to be realistic and represent the entire duration of the project
- Show project timelines using most appropriate style, for example: Bar chart (Gantt chart).



Budgets

- Outlines the funds needed effectively conduct the research proposed
- Outlines exactly what researcher realistically need from the funding agency to conduct the research
- Budget should be realistic
- Aligns with agency suggested/required budget categories
- The budget should align with the activities proposed in research design

Budget categories

- Personnel (salary and benefits)
- Researcher (time, salary and benefits)
- Training
- Consultants and/or resource person (salary)
- Instruction
- Equipment
- Supplies (paper, tapes, film, batteries, printing costs, publication cost etc.)
- Communication (telephone/postage/Internet/ media)
- Materials preparation (software, medical supplies, copying and printing)
- Travel/subsistence
- Community liaison
- Rental of facilities
- Evaluation
- Indirect costs (costs that your organization requires you to include)
- Other expenses (lunches for Meetings, interviews etc.)

Budget Justification

- Justify each budget item
- Demonstrate how the budget items align with the activities to be undertaken in your research design
- Provide details on additional sources of funding available to the organization or Principal investigator
- If the funds will go to different institutions, indicate allocation of funds by site

Proposal Presentation

- A proposal presentation has a distinct audience and purpose, a researcher should assume his/her audience to be:

  - ❖ Experts in the field of proposed study
  - ❖ Generalists with exposure to the field of study

  A researcher need to persuade/convince evaluators:

  - ❖ That the project is worth doing
  - ❖ The researcher is capable of carrying it out

  A research should help evaluators understand the motivation of his/her idea:

  - ❖ General: What is the problem? What is its (social, scientific) significance? •
  - ❖ Specific: How will he/she approach the research question?
  - ❖ Originality/Creativity/Innovation: Is the project novel? How is it related/compared to prior works

  A Researcher Should:

  - Provide a clear overview of your research plan
  - Propose pertinent experiments with good controls
  - Explain the methods succinctly
  - Demonstrate the kind of data he/she might see
  - Show how data will illuminate the central question
  - Offer alternative solutions/backup plans

Presentation Outline

- Brief research overview
- Sufficient background information for everyone to understand the proposal •
- Statement of the research problem and goals
- Project details and methods
- Predicted outcomes if everything goes according to plan and if nothing does
- Needed resources to complete the work (budget proposal)
- Societal impact if all goes well
- Timetable of activities (Gantt Chart)

Proposal Defence

- The Proposal should be prepared in accordance with formatting style and guidelines.

- The Proposal should include chapters (Introduction, Literature Review, and Methodology) and their traditional elements, the References, and appropriate Appendices (surveys, assessments, measurement scales).
- The defense should begin with a description of the context or background for the research question(s) in the study.
- It should also define key terms and variables and identifies hypotheses.
- The Proposal defense serves as an opportunity for the researcher to share the proposed study that is a comprehensive and well-defined plan for the research.
- The format of the Proposal defense is a brief and succinct presentation followed by questions from the review committee.

Points to defend

- Significance of the proposed research
- A summary of key points extracted from the literature on the topic
- A description of the conceptual framework and how the problem will be measured or assessed
- A proposal for analysis and interpretation of data or evidence
- Following the presentation, each member from the evaluation committee will be given the opportunity to present questions to the candidate;
- It is aimed to probe the candidate's understanding of the Proposal and to clarify, to both the candidate and Committee members, information which has been presented.
- Committee members may also suggest changes in any aspect of the Proposal.
- Opinions may differ; should differences arise, the chair provides guidance.
- The Proposal defense requires demonstration of two main elements:
    - The candidate, Chair, and Committee have thought deeply and carefully about the Proposal; the "big picture" is defensible.
    - The candidate is able to weigh the suggestions of the Committee and accept those that will strengthen the study.

Appendix As an appendix, in the case of quantitative studies, attach your research instrument. Also, attach a list of references in the appendix of the proposal.

Summary

A research proposal details the operational plan for obtaining answers to research questions.

- A research proposal must tell your supervisor and others what you propose to do, how you plan to proceed and why the chosen strategy has been selected.
- A research proposal thus assures readers of the validity of the methodology used to obtain answers accurately and objectively.
- Any given research proposal provides only a framework within which a research proposal for both quantitative and qualitative studies should be written and assume that you are reasonably well acquainted with research methodology and an academic style of writing.

Keywords

Title: The summary of the study's actual notion; it should use the fewest exact words that effectively convey the research's overall meaning.

Introduction:A brief background of the selected topic, including the objective, significance, relevancy, and applicability of the outcomes. It should clearly state the main points of the study. It also comprises objectives, which define the researcher's goals for the investigation.

Review of Literature: The overview of the chosen issue should be based on essential writings from other sources, such as authentic web sites, government records, and academic journal articles. Every source is described, summarized, and evaluated in a literature review.

Research Gap: The missing item from the present field of research's literature. This is an area where the style should be unique in order to fill a gap in a certain field of research.

Theoretical and Conceptual Framework: The relationship that will be revealed inside the research is theoretical framework, which will also support the study's theory, and conceptual framework, which reflects the overall architecture of the study as well as a visual representation of the relationship between variables.

Hypothesis:A test-based forecast of what is likely to happen in the research. Uncertainty in a statement that emphasizes the link between the study's factors.

Methodology: In general, it makes useful to consider what type of research is conducted and how it is conducted in order for readers to assess the research's validity and dependability. This chapter comprises several sections, including sample design, data collection process, sample size, statistical techniques to be utilised, and so on.

Conclusions: Briefly, the entire research procedure that has been included in the synopsis and will be included in the major research is completed.

Timeline: Defining the total time line that will be required for each phase of the research, which must correspond to the time limit set by the relevant authorities.

References: It gives credit to the authors who contributed ideas and words to the research work. For reference, use the appropriate format, such as APA, Harvard, or MLA.

Self Assessment

1. A good research proposal will always

A. Provide with respondents name and address.

B. Focus on addressing the research objectives.

C. Consider all possible research that had previously been done on the topic.

D. Discuss all unnecessary data.

2-One step that is not included in planning a research study is:

A. Identifying a researchable problem.

B. A review of current research.

C. Statement of the research question.

D. Developing a research plan.

3. The statement of purpose in a research study should:

A. Identify the design of the study.

B. Identify the intent or objective of the study.

C. Specify the type of people to be used in the study.

D. Describe the study.

4. A review of the literature prior to formulating research questions allows the researcher to do which of the following?

A. To become familiar with prior research on the phenomenon of interest.

B. To identify potential methodological problems in the research area.

C. To develop a list of pertinent problems relative to the phenomenon of interest.

D. All of the above.

5. The feasibility of a research study should be considered in light of:

A. Cost and time required to conduct the study.

B. Skills required of the researcher.

C. Potential ethical concerns.

D. All of the above.

6. A formal statement of the research question or "purpose of research study" generally

A. Is made prior to the literature review.

B. Is made after the literature review.

C. Will not help guide the research process.

D. All of the above.

7. The Introduction section of the research proposal

A. Gives an overview of prior relevant studies.

B. Contains a statement of the purpose of the study.

C. Concludes with a statement of the research questions and, for quantitative research, it includes the research hypothesis.

D. All the above

8. Research hypotheses are

A. Formulated prior to a review of the literature.

B. Statements of predicted relationships between variables.

C. Stated such that they cannot be confirmed or refuted.

D. Statements of no relationships between variables.

9. The research participants are described in detail in which section of the research proposal?

A. Introduction.

B. Research Methodology.

C. Data Analysis.

D. Conclusion.

10. According to the text, which of the following orders is the recommended in the flowchart of the development of a research idea?

A. Research topic, research problem, research purpose, research question, and hypothesis.

B. Research topic, research purpose, research problem, research question, and hypothesis.

C. Research topic, research problem, research purpose, research question, and hypothesis.

D. Research topic, hypothesis, research problem, research question, research purpose.

11. The timing section of a project will NOT include:

A. Progress report dates.

B. Guidelines on ethics.

C. Deadline for ending data collection.

D. Deadline for submitting the final report.

12. The research proposal's literature review is important because

A. The advisor insists upon it.

B. It looks authoritative.

C. It shows that you are knowledgeable about the literature that relates to your research topic.

D. It is expected by the university.

13. Which section of the research proposal describes the purpose with a full statement of the research question?

A. Introduction.

B. Research Methodology.

C. Literature review.

D. References.

14. Which of the following phrase should be avoided in a research proposal?

A. I hope to

B. The intention is to complete the study by

C. This research draws on the work of

D. The research seeks to

15. What helps to agree timings, agree resource allocation and also draws boundaries?

A. The questionnaire.

B. The Proposal.

C. The final report.

D. The interview schedule.

Answers for Self Assessment

1.	B	2.	D	3.	B	4.	D	5.	D
6.	B	7.	D	8.	B	9.	B	10.	A
11.	B	12.	B	13.	A	14.	D	15.	B

Review Questions

1. Enumerate the contents of any research proposal?
2. How far is it important to formulate objectives in the proposal?
3. Draft a sample research proposal on any given topic of your choice?
4. Why is it important to write about limitation in any given research proposal?
5. What are the considerations in presenting research proposal?
6. Throw light on the research proposal defence?
7. State as to why it is important to have time lines in any research proposal?

FurtherReadings

Business Research Methods By Naval Bajpai, Pearson

Marketing Research By Naresh K Malhotra, Pearson

Marketing Research: Text And Cases By Nargundkar, R., Mcgraw Hill Education

A Parasuraman, Dhruv Grewal, Marketing Research, Biztantra

Paneerselvam, R, Research Methods, Phi.

Web Links

https://www.monash.edu/rlo/graduate-research-writing/write-the-thesis/writing-a-research-proposal

https://www.youtube.com/watch?v=d8brslGIi10

https://www.youtube.com/watch?v=eALzUfkQJRU

https://www.youtube.com/watch?v=aj4H2nVuqNE

https://www.youtube.com/watch?v=NIUTwCoLIVo

https://www.birmingham.ac.uk/schools/law/courses/research/research-proposal.aspx

LOVELY PROFESSIONAL UNIVERSITY
Jalandhar-Delhi G.T. Road (NH-1)
Phagwara, Punjab (India)-144411
For Enquiry: +91-1824-521360
Fax.: +91-1824-506111
Email: odl@lpu.co.in