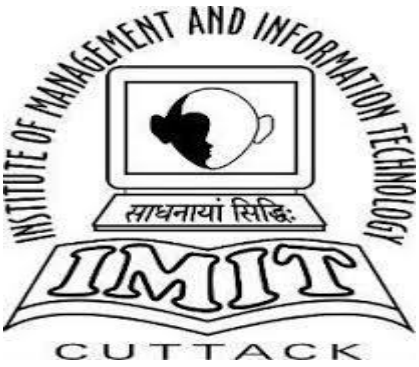


INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY: CUTTACK

(*A Constituent College of Biju Patnaik University of Technology: Odisha)*

BUSINESS RESEARCH (18MBA204)

STUDY MATERIAL

Prepared By

Dr. Chandrakanta Sahoo

(email: chandrakanta2008@gmail.com)

2nd Semester	18MBA204	Business Research	L-T-P 3-0-0	3 Credits	35 hrs

Course Objectives:

1. To equip the students with the basic understanding of the research methodology in changing business scenario.
2. To provide an insight into the application of dynamic analytical techniques to face the challenges, aimed at fulfilling the objective of business decision making.

Module I:
Introduction to RM: Meaning and significance of research. Importance of scientific research in business decision making. Types of research and research process. Identification of research problem and formulation of hypothesis. Research Designs.
Primary data, Secondary data, Design of questionnaire; Sampling fundamentals and sample designs. Measurement and Scaling Techniques, Data Processing.

Module II:
Data Analysis – I: Hypothesis testing; Z-test, t-test, F-test, chi-square test. Analysis of variance (One and Two way). Non-parametric, Test – Sign Test, Run test, Krushall – Wallis test

Module III:
Data Analysis – II: Factor analysis, Multiple Regressions Analysis. Discriminant Analysis (Concept)

Report writing and presentation: Research Report, Types and significance, Structure ofresearch report, Presentation of report.

It may be emphasized on practical aspects such as:

Use of software package to learn the following :-

(I) Draw frequencies, bar charts, histogram.
(ii) Creating and editing graphs and charts.
(iii) Bi-variate correlation.
(iv) The t-test procedure.
(v) Non-parametric Tests : Chi-square Test.
(vi) One way ANOVA Procedure.
(vii) Simple Regression, Multiple Regression, Reliability Analysis, Factor Analysis.

Director, Curriculum Development
Biju Patnaik University of Technology, Odisha
Rourkela

MODULE 1: INTRODUCTION TO RM

1.0.Learning Objectives:

After the end of this unit, students will be able to learn

1.1.Business Research: An Introduction
1.2.Concept
1.3.Importance
1.4.Types of Research Design
1.5.Research Process
1.6.Identifying Research Problem
1.7.Formulating Hypothesis
1.8.Data Collection
1.9. Designing Questionnaire
1.10. Sampling Techniques/ Design
1.11. Measurement and Scaling Techniques
1.12. Data Processing

1.1.Business Research: An Introduction

While research is important in both business and academia, there is no consensus in the literature on how it should be defined. The main reason for this is that different people can interpret research differently. However, from the many definitions there appears to be conformity that:

• research is a process of enquiry and investigation;
• it is systematic and methodical; and
• research increases knowledge.

Research can be defined as a 'step-by-step process that involves the collecting, recording, analyzing and interpreting of information'. As researchers, we are interested in improving our knowledge and understanding of our chosen topic. To do this effectively, researchers must have a clear set of research questions. The importance of research questions cannot be stressed highly enough. The research questions are the main focus of any project, and can probably best be described as 'the glue that holds the project together'

1.2.Business Research: Concept

The purpose of business research is to gather information in order to aid business related decision-making. Business research is defined as 'the systematic and objective process of

collecting, recording, analyzing and interpreting data for aid in solving managerial problems. These managerial problems can be linked to any business function, e.g. human resources, finance, marketing or research and development. Your research project can also be interpreted as business research in the sense that it will be related to business and management. In some cases, this may encompass more than one particular business discipline. For instance, a study might focus on the level of marketing knowledge among finance managers (marketing and finance). Some examples of areas of business and possible research issues are shown in Table 1.1.

TABLE 1.1 Examples of business research

Business aspect	Research issues
Consumer behaviour	Buying habits, brand preference, consumer attitudes
Human resources	Employee attitudes, staff retention, material incentives
Promotion	Media research, public relations studies, product recall through advertising
Product	Test markets, concept studies, performance studies
Finance	Forecasting, budgeting, efficiency of accounting software

1.3.Importance of research in business:

In business, research is important in identifying opportunities and threats. Often, a company's success or failure is dependent on the actions undertaken as a result of conducting research. Although carrying out business research does not guarantee success, it is likely to increase the probability that a new product, service, brand identity or even an event is successful. In some cases, the level of research conducted can be questionable, especially if public opinion is markedly different to that of an organization's viewpoint.

TABLE 1.2 Your research questions answered

Question	Answer
What is research?	Research can be defined as a 'step-by-step process that involves the collecting, recording, analyzing and interpreting of information'.
Why do I need to learn about business research?	An essential part of most business-related study programmes is the research project. Learning about business research helps you to successfully complete your project as well as provide transferable skills that can be used in a wide variety of business and management positions.

The Honeycomb of Research Methodology: In order to understand the key concepts of research and how they fit into your methodology, we now consider the Honeycomb of Research Methodology (see Figure 1.1). In this honeycomb, the three highlighted elements or key concepts of research are joined with three other elements to make up research methodology. Put another way, in the honeycomb, the six main elements – namely: (1) research philosophy;

(2) research approach; (3) research strategy; (4) research design; (5) data collection and (6) data analysis techniques – come together to form research methodology.

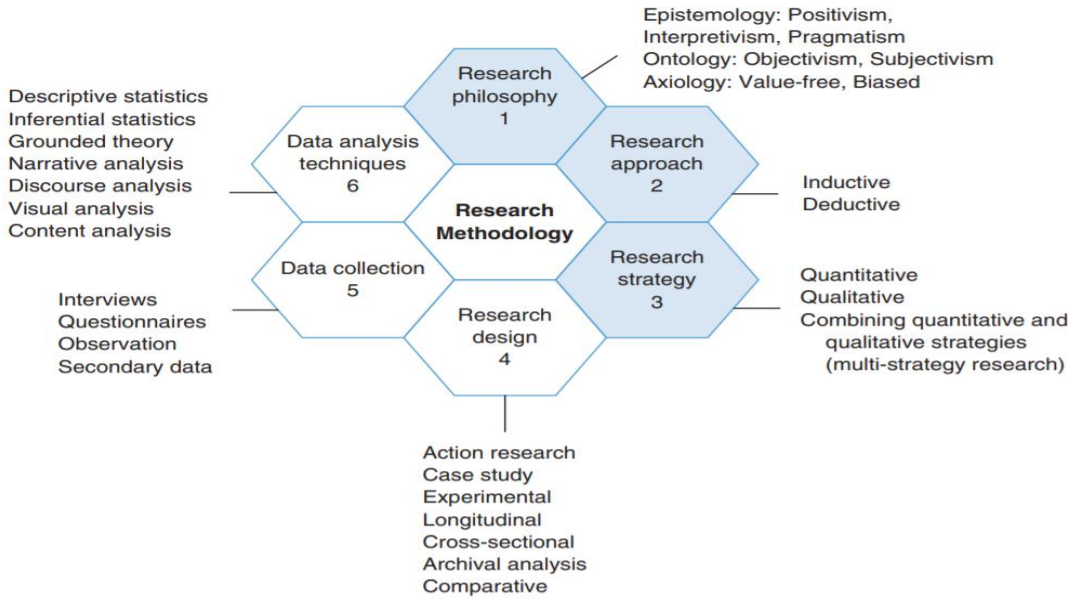


FIGURE 1.1 The Honeycomb of Research Methodology (©2013 Jonathan Wilson)

The Honeycomb RM model is explained here very briefly.

Research philosophy: In general, your research philosophy is linked to your views on the development of knowledge. In other words, what you think constitutes knowledge will impact the way that you go about your research. Subconsciously, this is something that comes naturally. Nonetheless, an understanding of research philosophy is important because it is fundamental to how you approach your research. Mark Easterby-Smith et al. (2002) suggest there are three reasons why an understanding of philosophical issues is very useful.

- It can help to clarify research designs. This entails considering the type of evidence required and how it is to be collected and interpreted.
- Knowledge of philosophy can help the researcher to recognize which designs work best and finally
- Knowledge of philosophy can help the researcher identify and adapt research designs according to the constraints of different subject or knowledge structures. In short, an understanding of research philosophy is important as it gets you thinking about your own role as a researcher.

Epistemology (what is the nature of knowledge?)

Ontology (the way we think the world is)

Axiology (role of values in inquiry)

Research approach: Research methods are often associated with two approaches – inductive and deductive. Let us look at each of these in turn. First, Kenneth F. Hyde (2000: 83) defined inductive as 'a theory-building process, starting with observations of specific instances, and seeking to establish generalisation about the phenomenon under investigation'. In other words, if you decide to follow an inductive approach to your study, you will be seeking to make observations about your research, and then perhaps contribute to a new theory. Conversely, a deductive approach 'begins with and applies a well-known theory'. For example, if your research project was focused on cross-cultural management and based on a deductive approach, then you may decide to apply Geert Hofstede's (1980) cultural theory. In other words, you are applying theory rather than attempting to generate new theory through an inductive approach.

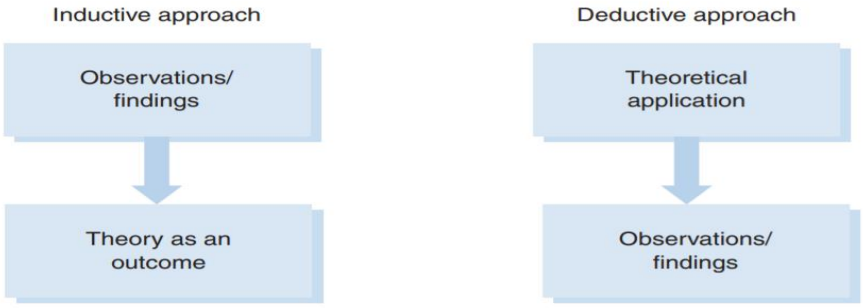


FIGURE 1.2 How theory fits into your research

TABLE 1.3 Major differences between deductive and inductive approaches to research

Deduction emphasizes	Induction emphasizes
• Scientific principles • Moving from theory to data • The need to explain causal relationships between variables • The collection of quantitative data • The application of controls to ensure validity of data • The operationalization of concepts to ensure clarity of definition • A highly structured approach • Researcher independence of what is being researched • The necessity to select samples of sufficient size in order to generalize conclusions	• Gaining an understanding of the meanings humans attach to events • A close understanding of the research context • The collection of qualitative data • A more flexible structure to permit changes of research emphasis as the research progresses • A realization that the research is part of the research process • Less concern with the need to generalize

Source: Saunders et al. (2007)

Research strategy: Two terms often used to describe the main research strategies to business research are qualitative and quantitative. Norman K. Denzin and Yvonna S. Lincoln (2000: 8) described the distinction between qualitative and quantitative as follows:

> "the word ***'qualitative'*** implies an emphasis on the qualities of entities and on processes and meanings that are not experimentally examined or measured (if measured at all) in terms of quantity, amount, intensity or frequency. Qualitative researchers stress the socially constructed nature of reality, the intimate relationship between the research and what is studied, and the situational constraints that shape inquiry. Such researchers emphasize the value-laden nature of inquiry. They seek answers to questions that stress how social experience is created and given meaning. In contrast**, *quantitative*** studies emphasize the measurement and analysis of causal relationships between variables, not processes. Proponents of such studies claim that their work is undertaken from within a value-free framework."

A comparison of qualitative and quantitative research

One way of describing qualitative and quantitative research is to compare the differences between the two. These differences include:

- the rejection of quantitative, positivist methods by qualitative researchers;
- qualitative researchers believe they can get closer to the actors' perspective through detailed interviewing and observation;
- qualitative researchers are more likely to confront the constraints of everyday life, while quantitative researchers tend to abstract themselves from this world and consequently, they seldom study it directly; and
- qualitative researchers tend to believe that rich descriptions are valuable while quantitative researchers are less concerned with such detail (Näslund, 2002: 328).

When comparing a qualitative and quantitative study, in a qualitative study, the research question often starts with a how or what so that initial forays into the topic describe what is going on. This is in contrast to quantitative questions that ask why and look for a comparison of groups (e.g. is Group 1 better at something than Group 2?) or a relationship between variables, with the intent of establishing an association, relationship, or cause and effect, e.g. did variable X explain what happened in variable Y? (Creswell, 1998).

Finally, your research strategy is likely to be a matter of choice. Once again, it is not simply a question of one or the other. In many respects your strategy does not need to follow a qualitative/quantitative divide. Increasingly, students are recognizing that using mixed methods for their data collection can add value to their study. For example, you may wish to administer

a questionnaire survey that explores customer satisfaction in your workplace, while you are also interested in conducting follow-up interviews with those individuals who appear to be particularly dissatisfied. In short, do not be 'pigeon-holed' into one strategy or the other, but consider the merits of adopting an eclectic approach.

Combining qualitative and quantitative research (multi-strategy research)

According to Tashakkori and Teddlie (1998: 17–18), mixed method studies are those that 'combine the qualitative and quantitative approaches into the research methodology of a single study or multi-phased study.'

TABLE 1.4 Positivism, interpretivism and pragmatism epistemologies

	Research approach	Ontology	Axiology	Research strategy
Positivism	Deductive	Objective	Value-free	Quantitative
Interpretivism	Inductive	Subjective	Biased	Qualitative
Pragmatism	Deductive/inductive	Objective and subjective	Value-free/biased	Qualitative and/or quantitative

1.4.Types of Research Design

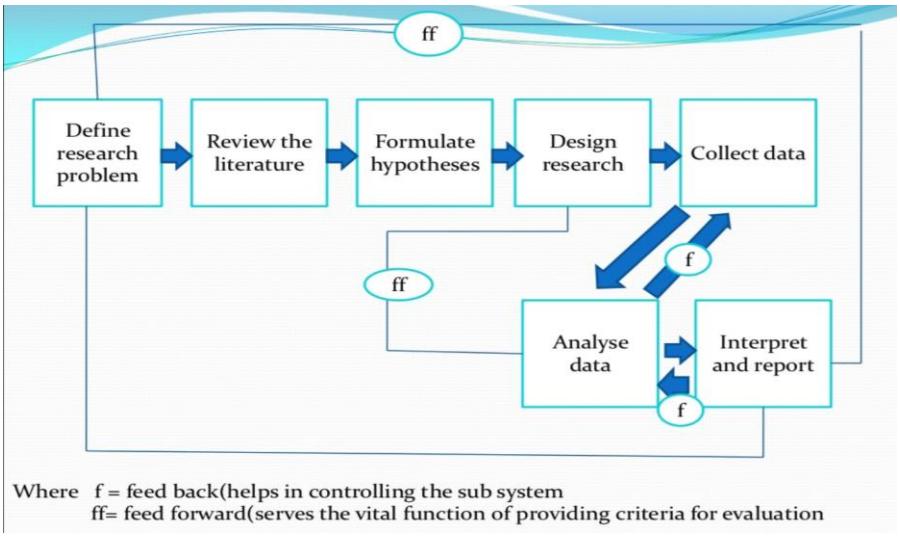
DESCRIPTIVE	ANALYTICAL
•Descriptive research includes surveys and fact-finding enquiries of different kinds. •The major purpose of descriptive research is description of the state of affairs as it exists at present. •The main characteristic of this method is that the researcher has no control over the variables; he can only report what has happened or what is happening. •Example 1: Examining the fluctuations of U. S. international trade balance during 1974-1995. •2.Starting from late 1986, the value of U.S. dollar value has steadily increased against the Japanese yen and German Mark. Examining the magnitude of this trend in the value of U.S. dollar is another example of descriptive research;	•In analytical research, on the other hand, the researcher has to use facts or information already available, and analyze these to make a critical evaluation of the material. •Analytical research attempts to explain why and how. It usually concerns itself with cause–effect relationships among variables. •Example1:Explaining why and how U.S. trade balance move in a particular way over time. •2. While explaining *how* and *why* this surge in the value of U.S. dollar is going to affect the U.S. Is analytical research.

APPLIED	FUNDAMENTAL
•Applied research aims at finding a solution for an immediate problem facing a society or an industrial/business organisation.	•Fundamental research is mainly concerned with. generalisations and with the formulation of a theory
•The central aim of applied research is to discover a solution for some pressing practical problem.	•Basic research is directed towards finding information that has a broad base of applications and thus, adds to the already existing organized body of scientific knowledge.
•Examples: Research aimed at certain conclusions (say, a solution) facing a concrete social or business. Research to identify social, economic or political trends that may affect a particular institution or marketing research or evaluation research are examples of applied research.	•Examples: Research concerning some natural phenomenon, human behaviour carried on with a view to make generalisations about human behaviour

QUALITATIVE RESEARCH	QUANTITATIVE RESEARCH
•Qualitative research is concerned with qualitative phenomenon, i.e., phenomena relating to or involving quality or kind.	•Quantitative research is based on the measurement of quantity or amount.
•Qualitative research is specially important in the behavioural sciences where the aim is to discover the underlying motives of human behaviour.	•It is applicable to phenomena that can be expressed in terms of quantity.
•This type of research aims at discovering the underlying motives and desires, using in depth interviews for the purpose.	•It usually involves collecting and converting data into numerical form so that statistical calculations can be made and conclusions drawn.
•For instance, when we are interested in Investigating the reasons for human behaviour, we quite often talk of 'Motivation Research'	•Example- Total sales of soap industry in terms of rupees cores and or quantity in terms of lakhs tones for particular year, say 2008,could be researched, compared with past 5 years and then projection for 2009 could be

CONCEPTUAL RESEARCH	EMPIRICAL RESEARCH
1. Research related to some abstract idea or theory generally used by philosophers and thinkers to develop new concepts or to reinterpret existing ones.	1. Research done on experience or observation alone, without due regard for system and theory. It is also called Experimental research as the conclusions can be verified by observation or experiment.
2. The researcher breaks down a theorem or concept into its constituent parts to gain a better & deeper understanding of the issue concerning the theorem. Conceptual research is a useful method but should be used in conjunction with other methods to produce better & understandable results.	2. The researcher provides himself with a working hypothesis to get the probable results. Facts are found to prove or disprove the hypothesis after which experimental designs are made to bring forth the desired information.

1.5.Research Process



1.6.How to Identify 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. Alternatively, **research problems** can be identified by reviewing recent literature, reports, or databases in your field. Often the section of "recommendations for the future studies" provided at the end of journal articles or doctoral dissertations suggest potential **research problems**

1.7.How to formulate Hypothesis?

A hypothesis is an assumption/ statement that introduces a research question and proposes an expected result. It is an integral part of the scientific method that forms the basis of scientific experiments. Therefore, you need to be careful and thorough when building your hypothesis. A minor flaw in the construction of your hypothesis could have an adverse effect on your experiment.

To devise and perform an experiment using the scientific method, you need to make sure that your hypothesis is testable. To be considered testable, some essential criteria must be met:

1. There must be a possibility to prove that the hypothesis is true. It just confirms the theory. No new discovery found. (Null Hypothesis)
2. There must be a possibility to prove that the hypothesis is false. It does not confirm the theory. New Alternative solution developed. (Alternative Hypothesis)
3. The results of the hypothesis must be reproducible.

Without these criteria, the hypothesis and the results will be vague. As a result, the experiment will not prove or disprove anything significant.

1.8.Data Collection:

Data collection is a systematic method of collecting and measuring data gathered from different sources of information in order to provide answers to relevant questions. An accurate evaluation of collected data can help researchers predict future phenomenon and trends.

Data collection can be classified into two, namely: primary and secondary data. Primary data are raw data i.e. fresh and are collected for the first time. Secondary data, on the other hand, are data that were previously collected and tested.

Methods of data collection

The system of data collection is based on the type of study being conducted. Depending on the researcher's research plan and design, there are several ways data can be collected.

The most commonly used methods are: published literature sources, surveys (email and mail), interviews (telephone, face-to-face or focus group), observations, documents and records, and experiments.

1. Literature sources

This involves the collection of data from already published text available in the public domain. Literature sources can include: textbooks, government or private companies' reports, newspapers, magazines, online published papers and articles.

This method of data collection is referred to as secondary data collection. In comparison to primary data collection, tt is inexpensive and not time consuming.

2. Surveys

Survey is another method of gathering information for research purposes. Information are gathered through questionnaire, mostly based on individual or group experiences regarding a particular phenomenon.

There are several ways by which this information can be collected. Most notable ways are: web-based questionnaire and paper-based questionnaire (printed form). The results of this method of data collection are generally easy to analyse.

3. Interviews

Interview is a qualitative method of data collection whose results are based on intensive engagement with respondents about a particular study. Usually, interviews are used in order to collect in-depth responses from the professionals being interviewed.

Interview can be structured (formal), semi-structured or unstructured (informal). In essence, an interview method of data collection can be conducted through face-to-face meeting with the interviewee(s) or through telephone.

4. Observations

Observation method of information gathering is used by monitoring participants in a specific situation or environment at a given time and day. Basically, researchers observe the behaviour of the surrounding environments or people that are being studied. This type of study can be controlled, natural or participant.

Controlled observation is when the researcher uses a standardised procedure of observing participants or the environment. Natural observation is when participants are being observed in their natural conditions. Participant observation is where the researcher becomes part of the group being studied.

5. Documents and records

This is the process of examining existing documents and records of an organisation for tracking changes over a period of time. Records can be tracked by examining call logs, email logs, databases, minutes of meetings, staff reports, information logs, etc.

For instance, an organisation may want to understand why there are lots of negative reviews and complains from customer about its products or services. In this case, the organisation will look into records of their products or services and recorded interaction of employees with customers.

6. Experiments

Experimental research is a research method where the causal relationship between two variables are being examined. One of the variables can be manipulated, and the other is measured. These two variables are classified as dependent and independent variables.

In experimental research, data are mostly collected based on the cause and effect of the two variables being studied. This type of research are common among medical researchers, and it uses quantitative research approach.

7. Questionnaire

Questionnaire can be an instrument of data collection for obtaining qualitative and quantitative data. Qualitative data (e.g. job satisfaction, customers' satisfaction etc.) can be collected using ordinal scale or by interview/ focused group discussion (FGD) method. and quantitative data (e.g. economic status of individual etc) can be collected by using ratio scale.

1.9.Designing of Questionnaire

The design of a questionnaire will depend on whether the researcher wishes to collect exploratory information (i.e. qualitative information for the purposes of better understanding or the generation of hypotheses on a subject) or quantitative information (to test specific hypotheses that have previously been generated).

Exploratory questionnaires: If the data to be collected is qualitative or is not to be statistically evaluated, it may be that no formal questionnaire is needed. For example, in interviewing the female head of the household to find out how decisions are made within the family when purchasing breakfast foodstuffs, a formal questionnaire may restrict the discussion and prevent a full exploration of the woman's views and processes. Instead one might prepare a brief guide, listing perhaps ten major open-ended questions, with appropriate probes/prompts listed under each.

Formal standardised questionnaires: If the researcher is looking to test and quantify hypotheses and the data is to be analysed statistically, a formal standardised questionnaire is designed. Such questionnaires are generally characterised by:

- ✓ prescribed wording and order of questions, to ensure that each respondent receives the same stimuli
- ✓ prescribed definitions or explanations for each question, to ensure interviewers handle questions consistently and can answer respondents' requests for clarification if they occur
- ✓ prescribed response format, to enable rapid completion of the questionnaire during the interviewing process.

Given the same task and the same hypotheses, six different people will probably come up with six different questionnaires that differ widely in their choice of questions, line of questioning, use of open-ended questions and length. There are no hard-and-fast rules about how to design a questionnaire, but there are a number of points that can be borne in mind:

- ✓ A well-designed questionnaire should meet the research objectives. This may seem obvious, but many research surveys omit important aspects due to inadequate preparatory work, and do not adequately probe particular issues due to poor understanding. To a certain degree some of this is inevitable. Every survey is bound to leave some questions unanswered and provide a need for further research but the objective of good questionnaire design is to 'minimise' these problems.

- ✓ It should obtain the most complete and accurate information possible. The questionnaire designer needs to ensure that respondents fully understand the questions and are not likely to refuse to answer, lie to the interviewer or try to conceal their attitudes. A good questionnaire is organised and worded to encourage respondents to provide accurate, unbiased and complete information.
- ✓ A well-designed questionnaire should make it easy for respondents to give the necessary information and for the interviewer to record the answer, and it should be arranged so that sound analysis and interpretation are possible.
- ✓ It would keep the interview brief and to the point and be so arranged that the respondent(s) remain interested throughout the interview.

1.10. Sampling Techniques/ Design

It would normally be impractical to study a whole population, for example when doing a questionnaire survey. Sampling is a method that allows researchers to infer information about a population based on results from a subset of the population, without having to investigate every individual. In a general bird's eye view, it can be said as the representative of the population. For example, a doctor wants blood test report of a patient. From any part of the body, a drop of blood can be collected. The test report will reveal the same result. If total blood of the body is the population, a drop of blood taken from the body is called a sample.

There are several different sampling techniques available, and they can be subdivided into two groups:

- Probability sampling and
- Non-probability sampling.

In probability (random) sampling, you start with a complete sampling frame of all eligible individuals from which you select your sample. In this way, all eligible individuals have a chance of being chosen for the sample, and you will be more able to generalise the results from your study. Probability sampling methods tend to be more time-consuming and expensive than non-probability sampling.

In non-probability (non-random) sampling, you do not start with a complete sampling frame, so some individuals have no chance of being selected. Consequently, you cannot estimate the effect of sampling error and there is a significant risk of ending up with a non-representative sample which produces non-generalisable results. However, non-probability sampling methods tend to be cheaper and more convenient, and they are useful for exploratory research and hypothesis generation.

Probability Sampling Methods

1. Simple random sampling

In this case, each individual is chosen entirely by chance and each member of the population has an equal chance, or probability, of being selected.

As with all probability sampling methods, simple random sampling allows the sampling error to be calculated and reduces selection bias. A specific advantage is that it is the most straightforward method of probability sampling. A disadvantage of simple random sampling is that you may not select enough individuals with your characteristic of interest, especially if that characteristic is uncommon. It may also be difficult to define a complete sampling frame and inconvenient to contact them, especially if different forms of contact are required (email, phone, post) and your sample units are scattered over a wide geographical area.

2. Systematic sampling

Individuals are selected at regular intervals from the sampling frame. The intervals are chosen to ensure an adequate sample size. If you need a sample size *n* from a population of size *x*, you should select every x/n^{th} individual for the sample. For example, if you wanted a sample size of 100 from a population of 1000, select every $1000/100 = 10^{th}$ member of the sampling frame. Systematic sampling is often more convenient than simple random sampling, and it is easy to administer.

However, it may also lead to bias, for example if there are underlying patterns in the order of the individuals in the sampling frame, such that the sampling technique coincides with the periodicity of the underlying pattern. As a hypothetical example, if a group of students were being sampled to gain their opinions on college facilities, but the Student Record Department's central list of all students was arranged such that the sex of students alternated between male and female, choosing an even interval (e.g. every 20^{th} student) would result in a sample of all males or all females. Whilst in this example the bias is obvious and should be easily corrected, this may not always be the case.

3. Stratified sampling

In this method, the population is first divided into subgroups (or strata) who all share a similar characteristic. It is used when we might reasonably expect the measurement of interest to vary between the different subgroups, and we want to ensure representation from all the subgroups. For example, in a study of stroke outcomes, we may stratify the population by sex, to ensure

equal representation of men and women. The study sample is then obtained by taking equal sample sizes from each stratum. In stratified sampling, it may also be appropriate to choose non-equal sample sizes from each stratum. For example, in a study of the health outcomes of nursing staff in a country, if there are three hospitals each with different numbers of nursing staff (hospital A has 500 nurses, hospital B has 1000 and hospital C has 2000), then it would be appropriate to choose the sample numbers from each hospital *proportionally* (e.g. 10 from hospital A, 20 from hospital B and 40 from hospital C). This ensures a more realistic and accurate estimation of the health outcomes of nurses across the county, whereas simple random sampling would over-represent nurses from hospitals A and B. The fact that the sample was stratified should be taken into account at the analysis stage.

Stratified sampling improves the accuracy and representativeness of the results by reducing sampling bias. However, it requires knowledge of the appropriate characteristics of the sampling frame (the details of which are not always available), and it can be difficult to decide which characteristic(s) to stratify by.

4. Clustered sampling

In a clustered sample, subgroups of the population are used as the sampling unit, rather than individuals. The population is divided into subgroups, known as clusters, which are randomly selected to be included in the study. Clusters are usually already defined, for example individual Gram Panchayat (GP) practices or towns could be identified as clusters. In single-stage cluster sampling, all members of the chosen clusters are then included in the study. In two-stage cluster sampling, a selection of individuals from each cluster is then randomly selected for inclusion. Clustering should be taken into account in the analysis. The Household survey, which is undertaken annually in England, is a good example of a (one-stage) cluster sample. All members of the selected households (clusters) are included in the survey.[1]

Cluster sampling can be more efficient that simple random sampling, especially where a study takes place over a wide geographical region. For instance, it is easier to contact lots of individuals in a few GP practices than a few individuals in many different GP practices. Disadvantages include an increased risk of bias, if the chosen clusters are not representative of the population, resulting in an increased sampling error.

Non-Probability Sampling Methods

1. Convenience sampling

Convenience sampling is perhaps the easiest method of sampling, because participants are selected based on availability and willingness to take part. Useful results can be obtained, but the results are prone to significant bias, because those who volunteer to take part may be different from those who choose not to (volunteer bias), and the sample may not be representative of other characteristics, such as age or sex. Note: volunteer bias is a risk of all non-probability sampling methods.

2. Quota sampling

This method of sampling is often used by market researchers. Interviewers are given a quota of subjects of a specified type to attempt to recruit. For example, an interviewer might be told to go out and select 20 adult men, 20 adult women, 10 teenage girls and 10 teenage boys so that they could interview them about their television viewing. Ideally the quotas chosen would proportionally represent the characteristics of the underlying population.

Whilst this has the advantage of being relatively straightforward and potentially representative, the chosen sample may not be representative of other characteristics that weren't considered (a consequence of the non-random nature of sampling).

3. Judgement (or Purposive) Sampling

Also known as selective, or subjective, sampling, this technique relies on the judgement of the researcher when choosing who to ask to participate. Researchers may implicitly thus choose a "representative" sample to suit their needs, or specifically approach individuals with certain characteristics. This approach is often used by the media when canvassing the public for opinions and in qualitative research.

Judgement sampling has the advantage of being time-and cost-effective to perform whilst resulting in a range of responses (particularly useful in qualitative research). However, in addition to volunteer bias, it is also prone to errors of judgement by the researcher and the findings, whilst being potentially broad, will not necessarily be representative.

4. Snowball sampling

This method is commonly used in social sciences when investigating hard-to-reach groups. Existing subjects are asked to nominate further subjects known to them, so the sample increases in size like a rolling snowball. For example, when carrying out a survey of risk behaviours amongst intravenous drug users, participants may be asked to nominate other users to be interviewed.

Snowball sampling can be effective when a sampling frame is difficult to identify. However, by selecting friends and acquaintances of subjects already investigated, there is a significant risk of selection bias (choosing a large number of people with similar characteristics or views to the initial individual identified).

1.11. Measurement and Scaling Techniques

Scaling technique is a method of placing respondents in continuation of gradual change in the pre-assigned values, symbols or numbers based on the features of a particular object as per the defined rules. All the scaling techniques are based on four pillars, i.e., order, description, distance and origin.

The marketing research is highly dependable upon the scaling techniques, without which no market analysis can be performed.

Types of Scaling Techniques

The researchers have identified many scaling techniques; today, we will discuss some of the most common scales used by business organizations, researchers, economists, experts, etc.

These techniques can be classified as primary scaling techniques and other scaling techniques.

Let us now study each of these methods in-depth below:

Primary Scaling Techniques

The major four scales used in statistics for market research consist of the following:

Nominal Scale

Nominal scales are adopted for non-quantitative (containing no numerical implication) labelling variables which are unique and different from one another.

Types of Nominal Scales:

1. **Dichotomous**: A nominal scale that has only two labels is called 'dichotomous'; *for example*, Yes/No.
2. **Nominal with Order**: The labels on a nominal scale arranged in an ascending or descending order is termed as 'nominal with order'; *for example*, Excellent, Good, Average, Poor, Worst.
3. **Nominal without Order**: Such nominal scale which has no sequence, is called 'nominal without order'; *for example*, Black, White.

Ordinal Scale

The ordinal scale functions on the concept of the relative position of the objects or labels based on the individual's choice or preference.

For example, At Amazon.in, every product has a customer review section where the buyers rate the listed product according to their buying experience, product features, quality, usage, etc.

The ratings so provided are as follows:

- 5 Star – Excellent
- 4 Star – Good
- 3 Star – Average
- 2 Star – Poor
- 1 Star – Worst

Interval Scale

An interval scale is also called a cardinal scale which is the numerical labelling with the same difference among the consecutive measurement units. With the help of this scaling technique, researchers can obtain a better comparison between the objects.

For example; A survey conducted by an automobile company to know the number of vehicles owned by the people living in a particular area who can be its prospective customers in future. It adopted the interval scaling technique for the purpose and provided the units as 1, 2, 3, 4, 5, 6 to select from.

In the scale mentioned above, every unit has the same difference, i.e., 1, whether it is between 2 and 3 or between 4 and 5.

Ratio Scale

One of the most superior measurement technique is the ratio scale. Similar to an interval scale, a ratio scale is an abstract number system. It allows measurement at proper intervals, order, categorization and distance, with an added property of originating from a fixed zero point. Here, the comparison can be made in terms of the acquired ratio.

For example, A health product manufacturing company surveyed to identify the level of obesity in a particular locality. It released the following survey questionnaire: Select a category to which your weight belongs to:

Less than 40 kilograms

- 40-59 Kilograms
- 60-79 Kilograms
- 80-99 Kilograms
- 100-119 Kilograms
- 120 Kilograms and more

The following table will better clarify the difference between all the four primary scaling techniques:

PARTICULAR	NOMINAL SCALE	ORDINAL SCALE	INTERVAL SCALE	RATIO SCALE
Characteristics	Description	Order	Distance	Description, Order, Distance and Origin
Sequential Arrangement	Not Applicable	Applicable	Applicable	Applicable
Fixed Zero Point	Not Applicable	Not Applicable	Not Applicable	Applicable
Multiplication and Division	Not Applicable	Not Applicable	Not Applicable	Applicable
Addition and Subtraction	Not Applicable	Not Applicable	Applicable	Applicable
Difference between Variables	Non-Measurable	Non-Measurable	Measurable	Measurable
Mean	Not Applicable	Not Applicable	Applicable	Applicable
Median	Not Applicable	Applicable	Applicable	Applicable
Mode	Applicable	Applicable	Applicable	Applicable

1.12. Data Processing

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output.

Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

Six stages of data processing

1. Data collection

Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.

2. Data preparation

Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as "pre-processing" is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create high-quality data for the best business intelligence.

3. Data input

The clean data is then entered into its destination (perhaps a CRM like Salesforce or a data warehouse like Redshift), and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.

4. Processing

During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation. Processing is done using machine learning algorithms, though the process itself may vary slightly depending on the source of data being processed (data lakes, social networks, connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).

5. Data output/interpretation

The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.). Members of the company or institution can now begin to self-serve the data for their own data analytics projects.

6. Data storage

The final stage of data processing is storage. After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. Plus, properly stored data is a necessity for compliance with data protection

legislation like GDPR. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

Summery

Research can be defined as a 'step-by-step process that involves the collecting, recording, analyzing and interpreting of information.' The purpose of business research is to gather information in order to aid business related decision-making. Business research is defined as 'the systematic and objective process of collecting, recording, analyzing and interpreting data for aid in solving managerial problems. In business, research is important in identifying opportunities and threats. Often, a company's success or failure is dependent on the actions undertaken as a result of conducting research. Although carrying out business research does not guarantee success, it is likely to increase the probability that a new product, service, brand identity or even an event is successful. There are various types of research: (a) Descriptive and Analytical Research; (b) Applied and Fundamental Research; (c) Quantitative and Qualitative Research; (d) Conceptual and Empirical Research. Research as a process involved certain steps. The steps are (a) defining research problems; (b) Review of Literature; (c) formulate hypothesis; (d) research design; (e) Data Collection; (f) Data Analysis and; (g) Interpretation and reporting. Data collection can be classified into two, namely: primary and secondary data. Primary data are raw data i.e. fresh and are collected for the first time. Secondary data, on the other hand, are data that were previously collected and tested. There are several different sampling techniques available, and they can be subdivided into two groups:

- Probability sampling and
- Non-probability sampling.

In probability (random) sampling, you start with a complete sampling frame of all eligible individuals from which you select your sample. In this way, all eligible individuals have a chance of being chosen for the sample, and you will be more able to generalise the results from your study. In non-probability (non-random) sampling, you do not start with a complete sampling frame, so some individuals have no chance of being selected. Consequently, you cannot estimate the effect of sampling error and there is a significant risk of ending up with a non-representative sample which produces non-generalisable results.

Scaling technique is a method of placing respondents in continuation of gradual change in the pre-assigned values, symbols or numbers based on the features of a particular object as per the defined rules. All the scaling techniques are based on four pillars, i.e., order, description, distance and origin.

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output.

Review Questions

Long Type Questions

1. Describe the concept and significance of research in Business.
2. Describe various types of research design.
3. Describe research process in details with a business example.
4. Describe various kinds of measurement and scaling techniques.
5. Describe contextual uses of various types of sampling techniques with suitable examples.
6. Describe various stages in data processing.
7. Compare and contrast probability and non probability sampling methods.

Short Types questions:

1. Explain the significance of business research.
2. Explain the functional research domains of business with suitable examples.
3. Distinguish between deductive and inductive approaches of business research.
4. Distinguish between qualitative and quantitative research in business.
5. Explain mixed approach to business with suitable examples.
6. Distinguish between descriptive and analytical research design.
7. Distinguish between conceptual and empirical research design.
8. Explain research process.
9. Distinguish between null and alternative hypothesis.
10. Explain the various sources of data obtained for business research.
11. Explain various types of questionnaires.
12. Explain various types of scaling techniques with examples.
13. Explain the stages involved in data processing.

Very Short Types Questions

1. Why is business research significant?
2. What is business research?
3. Write two of the functional domains of business research.
4. Write one example of qualitative research.
5. Write one example of quantitative research.
6. Why is hypothesis important in business research?
7. What are the sources of data?
8. How is primary study conducted?
9. Write two methods of data collection.
10. What is snowball sampling.
11. What are the uses of random sampling?
12. What are the problems associated with judgemental sampling?
13. Write an example of ordinal scale.
14. Write an example of nominal scale.
15. Write the stages of data processing.

MODULE 2: DATA ANALYSIS

Statistical Tests — When to use Which Test?

2.0. Learning Objectives

After the end of this unit, the students will be able to learn

2.0. Learning Objectives

After the end of this unit, the students will be able to learn

2.1. Introduction

After collection of data using appropriate research instruments, selection of appropriate method of analysis plays a vital role in addressing the research problem. Inappropriate selection of methods of data analysis does not serve the very purpose of conducting a research. This unit discusses some of the relevant methods such as descriptive analysis, parametric and non-para metric test useful for data analysis.

2.2. Descriptive Statistics

Descriptive statistics describe a sample, i.e., straightforward. You simply take a group that you're interested in, record data about the group members, and then use summary statistics and graphs to present the group properties. With descriptive statistics, there is no uncertainty because you are describing only the people or items that you actually measure. You're not trying to infer properties about a larger population.

The process involves taking a potentially large number of data points in the sample and reducing them down to a few meaningful summary values and graphs. This procedure allows us to gain more insights and visualize the data than simply pouring through row upon row of raw numbers!

Common tools of descriptive statistics

Descriptive statistics frequently use the following statistical measures to describe groups:

Central tendency: Use the mean or the median to locate the center of the dataset. This measure tells you where most values fall. The **mean**, or average, is calculated by finding the sum of the study data and dividing it by the total number of data. The **mode** is the number that appears most frequently in the set of data. The **median** is the middle value in a set of data

Box: 1

Knowing what we know, let's calculate the mean, median, and mode using the example from before. Again, the anxiety ratings of your classmates are 8, 4, 9, 3, 5, 8, 6, 6, 7, 8, and 10.

Mean: (8+ 4 + 9 + 3 + 5 + 8 + 6 + 6 + 7 + 8 + 10) / 11 = 74 / 11 = The mean is 6.73.

Median : In a data set of 11, the median is the number in the sixth place. 3, 4, 5, 6, 6, **7**, 8, 8, 8, 9, 10. The median is 7.

Mode: The number 8 appears more than any other number. The mode is 8.

Dispersion*:* How far out from the center do the data extend? You can use the range or standard deviation to measure the dispersion. The simplest measure of dispersion is the **range**. This tells us how spread out our data is. In order to calculate the range, you subtract the smallest number from the largest number. Just like the mean, the range is very sensitive to outliers.

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

Its symbol is **σ** (the Greek letter sigma)

Box 2

To find the standard deviation of a set of values:

a. Find the mean of the data

b. Find the difference (deviation) between each of the scores and the mean c. Square each deviation d. Sum the squares

e. Dividing by one less than the number of values, find the "mean" of this sum (the variance*)

f. Find the square root of the variance (the standard deviation) *Note: In some books, the variance is found by dividing by n. In statistics it is more useful to divide by n -1.

The formula is easy: it is the **square root** of the **Variance.** So now you ask, "What is the Variance?"

The **variance** is a measure of the average distance that a set of data lies from its mean. The variance is not a stand-alone statistic. It is typically used in order to calculate other statistics, such as the standard deviation. The higher the variance, the more spread out your data

In other words, Variance is defined as:

The average of the **squared** differences from the Mean.

To calculate the variance follow these steps:

- Work out the Mean (the simple average of the numbers)
- Then for each number: subtract the Mean and square the result (the *squared difference*).
- Then work out the average of those squared differences.

Illustration

EXAMPLE Find the variance and standard deviation of the following scores on an exam: 92, 95, 85, 80, 75, 50 SOLUTION First we find the mean of the data:

Mean = 92+95+85+80+75+50 6 = 477 6 = 79.5

Then we find the difference between each score and the mean (deviation).

The sum of the squares is 1317.50. Next, we find the "mean" of this sum (the variance).

1317.50 5 = 263.5 Finally, we find the square root of this variance. $\sqrt{263.5} \approx 16.2$ So, the standard deviation of the scores is 16.2; the variance is 263.5.

A low dispersion indicates that the values cluster more tightly around the center. Higher dispersion signifies that data points fall further away from the center. We can also graph the frequency distribution.

Score	Score - Mean	Difference from mean
92	92 – 79.5	+12.5
95	95 – 79.5	+15.5
85	85 – 79.5	+5.5
80	80 – 79.5	+0.5
75	75 – 79.5	-4.5
50	50 – 79.5	-29.5

Next we square each of these differences and then sum them.

Difference	Difference Squared
+12.5	156.25
+15.5	240.25
+5.5	30.25
+0.5	0.25
-4.5	20.25
-29.5	870.25
Sum of the squares →	1317.50

Skewness: The measure tells you whether the distribution of values is symmetric or skewed.

2.3. Inferential statistics

Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn. Because the goal of inferential statistics is to draw conclusions from a sample and generalize them to a population, we need to have confidence that our sample accurately reflects the population. This requirement affects our process. At a broad level, we must do the following:

1. Define the population we are studying.
2. Draw a representative sample from that population.
3. Use analyses that incorporate the sampling error.

We don't get to pick a convenient group. Instead, random sampling allows us to have confidence that the sample represents the population. This process is a primary method for obtaining samples that mirrors the population on average. Random sampling produces statistics, such as the mean, that do not tend to be too high or too low. Using a random sample, we can generalize from the sample to the broader population. Unfortunately, gathering a truly random sample can be a complicated process.

Pros and cons of working with samples

You gain tremendous benefits by working with a random sample drawn from a population. In most cases, it is simply impossible to measure the entire population to understand its properties. The alternative is to gather a random sample and then use the methodologies of inferential statistics to analyze the sample data.

While samples are much more practical and less expensive to work with, there are trade-offs. Typically, we learn about the population by drawing a relatively small sample from it. We are a very long way off from measuring all people or objects in that population. Consequently, when you estimate the properties of a population from a sample, the sample statistics are unlikely to equal the actual population value exactly.

For instance, your sample mean is unlikely to equal the population mean exactly. The difference between the sample statistic and the population value is the sampling error. Inferential statistics incorporate estimates of this error into the statistical results.

Sampling Error Formula refers to the formula that is used in order to calculate statistical error that occurs in the situation where person conducting the test doesn't select sample that represents the whole population under consideration and as per the formula Sampling Error is calculated by dividing the standard deviation of the population by the square root of the size of sample and then multiplying the resultant with the Z score value which is based on confidence interval.

Box 3: Calculation of Sampling Error

- **Step1**: Gathered all set of data called the population. Compute the population means and population standard deviation.
- **Step2**: Now, one needs to determine the size of the sample and further the sample size has to be less than the population and it should not be greater.
- **Step3**: Determine the confidence level and accordingly one can determine the value of Z score from its table.
- **Step4**: Now multiply Z score by the population standard deviation and divide the same by the square root of the sample size in order to arrive at a margin of error or sample size error.[i]

Sampling Error = Z x (σ /$\sqrt{n}$)

Where,

- Z is the Z score value based on the confidence interval
- σ is the population standard deviation
- n is the size of the sample

The most common methodologies in inferential statistics are hypothesis tests, confidence intervals, and regression analysis. Interestingly, these inferential methods can produce similar summary values as descriptive statistics, such as the mean and standard deviation. However, as I'll show you, we use them very differently when making inferences.

Box 4

Differences between Descriptive and Inferential Statistics

As you can see, the difference between descriptive and inferential statistics lies in the process as much as it does the statistics that you report.

For descriptive statistics, we choose a group that we want to describe and then measure all subjects in that group. The statistical summary describes this group with complete certainty (outside of measurement error).

For inferential statistics, we need to define the population and then devise a sampling plan that produces a representative sample. The statistical results incorporate the uncertainty that is inherent in using a sample to understand an entire population.

A study using descriptive statistics is simpler to perform. However, if you need evidence that an effect or relationship between variables exists in an entire population rather than only your sample, you need to use inferential statistics.

In contrast, summary values in descriptive statistics are straightforward. The average score in a specific class is a known value because we measured all individuals in that class. There is no uncertainty.

2.4. Hypothesis Testing

Before we venture on the difference between different tests, we need to formulate a clear understanding of what a null hypothesis is. A null hypothesis, proposes that no significant difference exists in a set of given observations. For the purpose of these tests in general

Null: Given two sample means are equal

Alternate: Given two sample means are not equal

For rejecting a null hypothesis, a test statistic is calculated. This test-statistic is then compared with a critical value and if it is found to be greater than the critical value the hypothesis is rejected. "*In the theoretical underpinnings, hypothesis tests are based on the notion of critical regions: the null hypothesis is rejected if the test statistic falls in the critical region*. The critical values are the boundaries of the critical region. If the test is one-sided (like a $\chi 2$ test or a one-sided t-test- one tailed) then there will be just one critical value, but in other cases (like a two-sided t-test- two tailed) there will be two".

Critical Value

A critical value is a point (or points) on the scale of the test statistic beyond which we reject the null hypothesis, and, is derived from the level of significance α of the test. ***Critical value can tell us, what is the probability of two sample means belonging to the same distribution. Higher, the critical value means lower the probability of two samples belonging to same distribution****. The general critical value for a two-tailed test is* ***1.96,*** *which is based on the fact that* ***95%*** *(confidence interval) of the area of a normal distribution is within 1.96 standard deviations of the mean.*

Box 5

Critical values for a test of hypothesis depend upon a test statistic, which is specific to the type of test, and the significance level, α, which defines the sensitivity of the test. A value of α = 0.05 implies that the null hypothesis is rejected 5 % of the time when it is in fact true. The choice of α is somewhat arbitrary, although in practice values of 0.1, 0.05, and 0.01 are common. Critical values are essentially cut-off values that define regions where the test statistic is unlikely to lie; for example, a region where the critical value is exceeded with probability α if the null hypothesis is true. The null hypothesis is rejected if the test statistic lies within this region which is often referred to as the rejection region(s).

Critical values can be used to do hypothesis testing in following ways:

1. Calculate test statistic
2. Calculate critical values based on significance level alpha.
3. Compare test statistic with critical values.

Box 6

P- Value: Another quantitative measure for reporting the result of a test of hypothesis is the p-value. The p-value is the probability of the test statistic being at least as extreme as the

one observed given that the null hypothesis is true. A small p-value is an indication that the null hypothesis is false.

It is good practice to decide in advance of the test how small a p-value is required to reject the test. This is exactly analagous to choosing a significance level, α, for test. For example, we decide either to reject the null hypothesis if the test statistic exceeds the critical value (for $\alpha = 0.05$) or analagously to reject the null hypothesis if the p-value is smaller than 0.05. It is important to understand the relationship between the two concepts because some statistical software packages report p-values rather than critical values.

Two-Tailed and One tailed Test of hypothesis

Two tailed test of hypothesis is a test of hypothesis where the area of rejection is on the both side. In case of one- tailed, rejection of null hypothesis is either at left side of the distribution (for negative values) or right side of the distribution (for positive value). Two tailed test of hypothesis is non- directional and one tailed test of hypothesis is directional in nature. Some hypotheses predict only that one value will be different from another, without additionally predicting which will be higher. The test of such a hypothesis is **nondirectional** or **two-tailed** because an extreme test statistic in either tail of the distribution (positive or negative) will lead to the rejection of the null hypothesis of no difference.

Illustration:

Suppose that you suspect that a particular class's performance on a proficiency test is not representative of those people who have taken the test. The national mean score on the test is 74.

The research hypothesis is:

The mean score of the class on the test is not 74.

Or in notation: $H_a : \mu \neq 74$

The null hypothesis is:

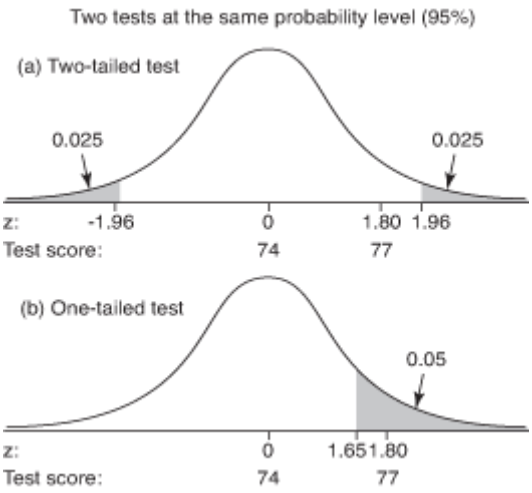
The mean score of the class on the test is 74.

In notation: $H_0: \mu = 74$

As in the last example, you decide to use a 5 percent probability level for the test. Both tests have a region of rejection, then, of 5 percent, or 0.05. In this example, however, the rejection region must be split between both tails of the distribution—0.025 in the upper tail and 0.025 in

the lower tail—because your hypothesis specifies only a difference, not a direction, as shown in Figure 1(a). You will reject the null hypotheses of no difference if the class sample mean is either much higher or much lower than the population mean of 74. In the previous example, only a sample mean much lower than the population mean would have led to the rejection of the null hypothesis.

Figure 1.Comparison of (a) a two-tailed test and (b) a one-tailed test, at the same probability level (95 percent).



The decision of whether to use a one- or a two-tailed test is important because a test statistic that falls in the region of rejection in a one-tailed test may not do so in a two-tailed test, even though both tests use the same probability level. Suppose the class sample mean in your example was 77, and its corresponding *z*-score was computed to be 1.80. Table 2 in "Statistics Tables" shows the critical *z*-scores for a probability of 0.025 in either tail to be –1.96 and 1.96. In order to reject the null hypothesis, the test statistic must be either smaller than –1.96 or greater than 1.96. It is not, so you cannot reject the null hypothesis. Refer to Figure 1(a).

Suppose, however, you had a reason to expect that the class would perform better on the proficiency test than the population, and you did a one-tailed test instead. For this test, the rejection region of 0.05 would be entirely within the upper tail. The critical *z*-value for a probability of 0.05 in the upper tail is 1.65. (Remember that Table 2 in "Statistics Tables" gives areas of the curve below *z*; so you look up the *z*-value for a probability of 0.95.) Your computed

test statistic of $z = 1.80$ exceeds the critical value and falls in the region of rejection, so you reject the null hypothesis and say that your suspicion that the class was better than the population was supported. See Figure 1(b).

In practice, you should use a one-tailed test only when you have good reason to expect that the difference will be in a particular direction. A two-tailed test is more conservative than a one-tailed test because a two-tailed test takes a more extreme test statistic to reject the null hypothesis.

2.5. Parametric and Non- Parametric Test

Reasons to Use Parametric Tests

Reason 1: Parametric tests can perform well with skewed and nonnormal distributions.
parametric tests can perform well with continuous data that are nonnormal if you satisfy the sample size guidelines in the table below. These guidelines are based on simulation studies conducted by statisticians here at Minitab. To learn more about these studies, read our Technical Papers.

Parametric tests (means)	**Nonparametric tests (medians)**
1-sample t test	1-sample Sign, 1-sample Wilcoxon
2-sample t test	Mann-Whitney test
One-Way ANOVA	Kruskal-Wallis, Mood's median test
Factorial DOE with one factor and one blocking variable	Friedman test

Reason 2: Parametric tests can perform well when the spread of each group is different

While nonparametric tests don't assume that your data follow a normal distribution, they do have other assumptions that can be hard to meet. For nonparametric tests that compare groups, a common assumption is that the data for all groups must have the same spread (dispersion). If your groups have a different spread, the nonparametric tests might not provide valid results.

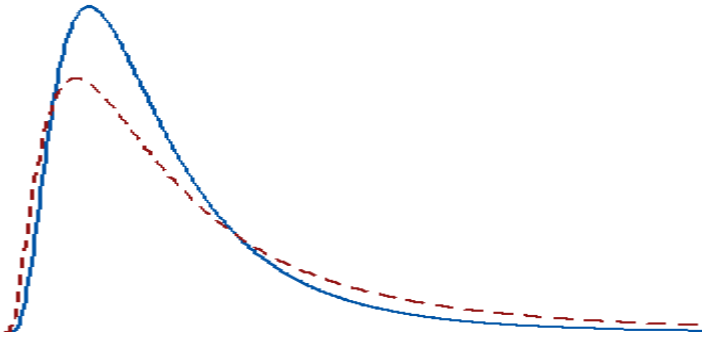
On the other hand, if you use the 2-sample t test or One-Way ANOVA, you can simply go to the **Options** sub dialog and uncheck *Assume equal variances*. Voilà, you're good to go even when the groups have different spreads!

Reason 3: Statistical power

Parametric tests usually have more statistical power than nonparametric tests. Thus, you are more likely to detect a significant effect when one truly exists.

Reasons to Use Nonparametric Tests

Reason 1: Your area of study is better represented by the median



This is one of the reasons to use a nonparametric test. The fact that you can perform a parametric test with non-normal data doesn't imply that the mean is the statistic that you want to test.

For example, the center of a skewed distribution, like income, can be better measured by the median where 50% are above the median and 50% are below. If you add a few billionaires to a sample, the mathematical mean increases greatly even though the income for the typical person doesn't change.

When your distribution is skewed enough, the mean is strongly affected by changes far out in the distribution's tail whereas the median continues to more closely reflect the center of the distribution. For these two distributions, a random sample of 100 from each distribution produces means that are significantly different, but medians that are not significantly different.

Reason 2: You have a very small sample size

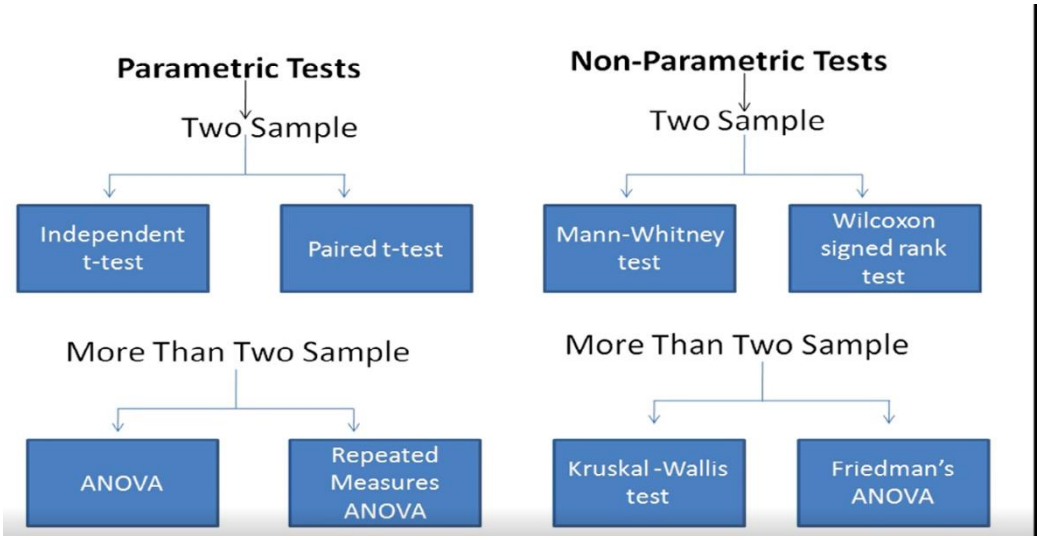
If you don't meet the sample size guidelines for the parametric tests and you are not confident that you have normally distributed data, you should use a nonparametric test. When you have a really small sample, you might not even be able to ascertain the distribution of your data because the distribution tests will lack sufficient power to provide meaningful results.

Reason 3: You have ordinal data, ranked data, or outliers that you can't remove

Typical parametric tests can only assess continuous data and the results can be significantly affected by outliers. Conversely, some nonparametric tests can handle ordinal data, ranked data, and not be seriously affected by outliers. Be sure to check the assumptions for the nonparametric test because each one has its own data requirements.

If you have Likert data and want to compare two groups, read my post Best Way to Analyze Likert Item Data: Two Sample T-Test versus Mann-Whitney.

Figure: 1: Parametric and Non-Parametric Testing Methods



Note: May here be noted that the details of all the methods are not possible to describe here as the business research syllabus limits us to go beyond it. As per the syllabus, the lecture materials are prepared.

2.6. Z-test

In a z-test, the sample is assumed to be normally distributed. A z-score is calculated with population parameters such as **"population mean"** and **"population standard deviation"** and is used to validate a hypothesis that the sample drawn belongs to the same population.

Null: Sample mean is same as the population mean

Alternate: Sample mean is not same as the population mean

The statistics used for this hypothesis testing is called z-statistic, the score for which is calculated as

z = (x — μ) / (σ / √n), where

x= sample mean

μ = population mean

σ / √n = population standard deviation

If the test statistic is lower than the critical value, accept the hypothesis or else reject the hypothesis

Illustration:

Question:

A researcher claims that, in a certain city, the average yearly consumption of milk per person is 28.1 gallons. We wish to test the validity of this claim, so we perform a study. We test 75 people and we find that their average yearly consumption of milk is 25.8 gallons. The standard deviation of the population is 9 gallons.

We wish to determine the validity of the researcher's claim. Calculate the p-value.

Z-test for a single mean

This procedure is done for a single treatment group. The sample size for this test is at least 30. It is assumed that there is no arbitrary difference in the mean of the treatment group and population group.

Answer and Explanation:

It is to be tested whether the average yearly consumption of milk per person is 28.1 gallons.

The number of people tested (n)=75(n)=75

The sample average consumption (¯x)=25.8(x¯)=25.8

The population standard deviation (σ)=9(σ)=9

The null and the alternative hypothesis can be stated as:

Ho:μ=28.1Ha:μ≠28.1Ho:μ=28.1Ha:μ≠28.1

The appropriate test statistic to be used is,

$z=\bar{x}-\mu\sigma\sqrt{n}=25.8-28.19\sqrt{75}=-2.213z=x^{-}-\mu\sigma n=25.8-28.1975=-2.213$

Using the z-tables, the two- tailed p-value for test statistic z = -2.213 is 0.0269.

Hence, **the p-value is 0.0269.**

Since the p-value is greater than the significance level 0.05, hence the test result is significant.

Thus, the null hypothesis is rejected. Therefore, at 5% level of significance, we have sufficient evidence to conclude that the average yearly consumption of milk per person is not equal to 28.1 gallons.

2.7. T-test

A t-test is used to compare the mean of two given samples. Like a z-test, a t-test also assumes a normal distribution of the sample. A t-test is used when the population parameters (mean and standard deviation) are not known.

There are three versions of t-test

1. Independent samples t-test which compares mean for two groups

2. Paired sample t-test which compares means from the same group at different times

3. One sample t-test which tests the mean of a single group against a known mean.

The statistic for this hypothesis testing is called t-statistic, the score for which is calculated as **t = (x1 — x2) / (σ / √n1 + σ / √n2),** where

x1 = mean of sample 1

x2 = mean of sample 2

n1 = size of sample 1

n2 = size of sample 2

There are multiple variations of t-test which are explained in detail here

Illustration:

Student's T-tests can be used in real life to compare means. For example, a drug company may want to test a new cancer drug to find out if it improves life expectancy. In an experiment, there's always a control group (a group who are given a placebo, or "sugar pill"). The control group may show an average life expectancy of +5 years, while the group taking the new drug might have a life expectancy of +6 years. It would seem that the drug might work. But it could be due to a fluke. To test this, researchers would use a Student's t-test to find out if the results are repeatable for an entire population.

The T Score.

The t score is a ratio between the **difference between two groups and the difference within the groups**. The larger the t score, the more difference there is between groups. The smaller the t score, the more similarity there is between groups. A t score of 3 means that the groups are three times as different *from* each other as they are within each other. When you run a t test, the bigger the t-value, the more likely it is that the results are repeatable.

- A large t-score tells you that the groups are different.
- A small t-score tells you that the groups are similar.

T-Values and P-values

How big is "big enough"? Every t-value has a p-value to go with it. A p-value is the probability that the results from your sample data occurred by chance. P-values are from 0% to 100%. They are usually written as a decimal. For example, a p value of 5% is 0.05. **Low p-values are good**; They indicate your data did not occur by chance. For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance. In most cases, a p-value of 0.05 (5%) is accepted to mean the data is valid.

Calculating the Statistic / Test Types

There are **three main types of t-test:**

- An Independent Samples t-test compares the means for two groups.
- A Paired sample t-test compares means from the same group at different times (say, one year apart).
- A One sample t-test tests the mean of a single group against a known mean.

Calculating an Independent Samples T Test By hand

Sample question: Calculate an independent samples t test for the following data sets:

Data set A: 1,2,2,3,3,4,4,5,5,6

Data set B: 1,2,4,5,5,5,6,6,7,9

Step 1: Sum the two groups:

A: 1 + 2 + 2 + 3 + 3 + 4 + 4 + 5 + 5 + 6 = 35

B: 1 + 2 + 4 + 5 + 5 + 5 + 6 + 6 + 7 + 9 = 50

Step 2: Square the sums from Step 1:

$35^2 = 1225$

$49^2 = 2500$

Set these numbers aside for a moment.

Step 3: Calculate the means for the two groups:

A: (1 + 2 + 2 + 3 + 3 + 4 + 4 + 5 + 5 + 6)/10 = 35/10 = 3.5

B: (1 + 2 + 4 + 5 + 5 + 5 + 6 + 6 + 7 + 9) = 50/10 = 5

Set these numbers aside for a moment.

Step 4: Square the individual scores and then add them up:

A: $1^1 + 2^2 + 2^2 + 3^3 + 3^3 + 4^4 + 4^4 + 5^5 + 5^5 + 6^6 = 145$

B: $1^2 + 2^2 + 4^4 + 5^5 + 5^5 + 5^5 + 6^6 + 6^6 + 7^7 + 9^9 = 298$

Set these numbers aside for a moment.

Step 5: Insert your numbers into the following formula and solve:

$$t = \frac{\mu_A - \mu_B}{\sqrt{\left[\frac{\left(\Sigma A^2 - \frac{(\Sigma A)^2}{n_A}\right) + \left(\Sigma B^2 - \frac{(\Sigma B)^2}{n_B}\right)}{n_A + n_B - 2}\right] \cdot \left[\frac{1}{n_A} + \frac{1}{n_B}\right]}}$$

$(\Sigma A)^2$: Sum of data set A, squared (Step 2).

$(\Sigma B)^2$: Sum of data set B, squared (Step 2).

μ_A: Mean of data set A (Step 3)

μ_B: Mean of data set B (Step 3)

ΣA^2: Sum of the squares of data set A (Step 4)

ΣB^2: Sum of the squares of data set B (Step 4)

n^A: Number of items in data set A

n^B: Number of items in data set B

$$t = \frac{3.5 - 5}{\sqrt{\left[\frac{\left(145 - \frac{1225}{10}\right) + \left(298 - \frac{2500}{10}\right)}{10 + 10 \quad -2}\right] \cdot \left[\frac{1}{10} + \frac{1}{10}\right]}}$$

$$t = \frac{-1.5}{\sqrt{\left[\frac{\left(145 - 122.5\right) + \left(298 - 250\right)}{18}\right] \cdot \left[\frac{2}{10}\right]}}$$

$$t = \frac{-1.5}{\sqrt{3.917 \cdot \frac{2}{10}}} = \frac{-1.5}{\sqrt{0.783}} = -1.69$$

Step 6: Find the Degrees of freedom (n_A-1 + n_B-1) = 18

Step 7: Look up your degrees of freedom (Step 6) in the t-table. If you don't know what your alpha level is, use 5% (0.05).

18 degrees of freedom at an alpha level of 0.05 = 2.10.

Step 8: Compare your calculated value (Step 5) to your table value (Step 7). The calculated value of -1.79 is less than the cutoff of 2.10 from the table. Therefore p > .05. As the p-value is greater than the alpha level, we cannot conclude that there is a difference between means.

Box 7: Difference between T and Z Tests

In hypothesis testing, Z tests and t test follow similar assumptions. For instance, they both imply that the underlying data distribution of a certain population follows the normal distribution. Even the graphical representations of the t and Z distribution are symmetrical and bell shaped. So, when do we use Z and when do we use t? Well, the answer is rather simple. Z tests are known as large sample tests and t tests are known as small sample tests. So what constitutes a large or small sample? When n<30n<30, we have a small sample and use t statistics. Similarly, when n>30n>30 we have a large sample and use Z Statistics.

2.8. ANOVA

ANOVA, also known as analysis of variance, is used to compare multiple (three or more) samples with a single test**.** There are 2 major flavors of ANOVA.

1. One-way ANOVA: It is used to compare the difference between the three or more samples/groups of a single independent variable.

2. MANOVA: MANOVA allows us to test the effect of one or more independent variable on two or more dependent variables. In addition, MANOVA can also detect the difference in co-relation between dependent variables given the groups of independent variables.

The hypothesis being tested in ANOVA is

Null: All pairs of samples are same i.e. all sample means are equal

Alternate: At least one pair of samples is significantly different

The statistics used to measure the significance, in this case, is called F-statistics. The F value is calculated using the formula

F= ((SSE1 — SSE2)/m)/ SSE2/n-k, where

SSE = residual sum of squares

m = number of restrictions

k = number of independent variables

There are multiple tools available such as SPSS, R packages, Excel etc. to carry out ANOVA on a given sample.

One Way ANOVA

A one way ANOVA is used to compare two means from two independent (unrelated) groups using the F-distribution. The null hypothesis for the test is that the two means are equal. Therefore, a significant result means that the two means are unequal.

Examples of when to use a one way ANOVA

Situation 1: You have a group of individuals randomly split into smaller groups and completing different tasks. For example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea.

Situation 2: Similar to situation 1, but in this case the individuals are split into groups based on an attribute they possess. For example, you might be studying leg strength of people according to weight. You could split participants into weight categories (obese, overweight and normal) and measure their leg strength on a weight machine.

Box 8: Assumptions of ANOVA

(i) All populations involved follow a normal distribution.

(ii) All populations have the same variance (or standard deviation).

(iii) The samples are randomly selected and independent of one another.

Illustration

Consider this example:

Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use α = 5%.

Table ANOVA.1

	Compact cars	Midsize cars	Full-size cars
	643	469	484
	655	427	456
	702	525	402
$\bar{X}$	666.67	473.67	447.33
S	31.18	49.17	41.68

(1.) State the null and alternative hypotheses The null hypothesis for an ANOVA always assumes the population means are equal. Hence, we may write the null hypothesis as:

H0: $\mu 1 = \mu 2 = \mu 3$ - The mean head pressure is statistically equal across the three types of cars. Since the null hypothesis assumes all the means are equal, we could reject the null hypothesis if only mean is not equal. Thus, the alternative hypothesis is:

Ha: At least one mean pressure is not statistically equal.

(2.) Calculate the appropriate test statistic.

The test statistic in ANOVA is the ratio of the between and within variation in the data. It follows an F distribution.

Total Sum of Squares – the total variation in the data. It is the sum of the between and within variation.

Total Sum of Squares (SST) = ∑∑= = − r i c j Xij X 1 1 2 ( ) , where r is the number of rows in the table, c is the number of columns, X is the grand mean, and Xij is the i th observation in the j th column.

Using the data in Table ANOVA.1 we may find the grand mean:

$$\bar{\bar{X}} = \frac{\sum X_{ij}}{N} = \frac{(643+655+702+469+427+525+484+456+402)}{9} = 529.22$$

$$SST = (643-529.22)^2 + (655-529.22)^2 + (702-529.22)^2 + (469-529.22)^2 + \ldots + (402-529.22)^2 = 96303.55$$

Between Sum of Squares (or Treatment Sum of Squares) – variation in the data between the different samples (or treatments).

Treatment Sum of Squares (SSTR) = ∑ − 2 r (X X) j j , where j r is the number of rows in the j th treatment and X j is the mean of the j th treatment.

Using the data in Table ANOVA.1,

$$SSTR = [3*(666.67-529.22)^2] + [3*(473.67-529.22)^2] + [3*(447.33-529.22)^2] = 86049.55$$

Within variation (or Error Sum of Squares) – variation in the data from each individual treatment.

$$SSE = [(643-666.67)^2 + (655-666.67)^2 + (702-666.67)^2] + [(469-473.67)^2 + (427-473.67)^2 + (525-473.67)^2] + [(484-447.33)^2 + (456-447.33)^2 + (402-447.33)^2] = 10254.$$

Note that SST = SSTR + SSE (96303.55 = 86049.55 + 10254).

Hence, you only need to compute any two of three sources of variation to conduct an ANOVA. Especially for the first few problems you work out, you should calculate all three for practice. The next step in an ANOVA is to compute the "average" sources of variation in the data using SST, SSTR, and SSE.

Total Mean Squares (MST) $= \frac{SST}{N-1}$ → "average total variation in the data" (N is the total number of observations)

$$\text{MST} = \frac{96303.55}{(9-1)} = 12037.94$$

Mean Square Treatment (MSTR) $= \frac{SSTR}{c-1}$ → "average between variation" (c is the number of columns in the data table)

$$\text{MSTR} = \frac{86049.55}{(3-1)} = 43024.78$$

Mean Square Error (MSE) $= \frac{SSE}{N-c}$ → "average within variation"

$$\text{MSE} = \frac{10254}{(9-3)} = 1709$$

Note: MST ≠ MSTR + MSE

The test statistic may now be calculated. For a one-way ANOVA the test statistic is equal to the ratio of MSTR and MSE. This is the ratio of the "average between variation" to the "average within variation." In addition, this ratio is known to follow an F distribution. Hence,

$$\text{F} = \frac{MSTR}{MSE} = \frac{43024.78}{1709} = 25.17$$

The intuition here is relatively straightforward. If the average between variation rises relative to the average within variation, the F statistic will rise and so will our chance of rejecting the null hypothesis.

(3.) Obtain the Critical Value
To find the critical value from an F distribution you must know the numerator (MSTR) and denominator (MSE) degrees of freedom, along with the significance level.

F^{CV} has df1 and df2 degrees of freedom, where df1 is the numerator degrees of freedom equal to c-1 and df2 is the denominator degrees of freedom equal to N-c.

In our example, df1 = 3 - 1 = 2 and df2 = 9 - 3 = 6. Hence we need to find $F_{2,6}^{CV}$ corresponding to $\alpha = 5\%$. Using the F tables in your text we determine that $F_{2,6}^{CV} = 5.14$.

(4.) Decision Rule
You reject the null hypothesis if: F (observed value) > F^{CV} (critical value). In our example 25.17 > 5.14, so we reject the null hypothesis.

(5.) Interpretation
Since we rejected the null hypothesis, we are 95% confident ($1-\alpha$) that the mean head pressure is not statistically equal for compact, midsize, and full size cars. However, since only one mean must be different to reject the null, we do not yet know which mean(s) is/are different. In short, an ANOVA test will test us that *at least one mean is different*, but an additional test must be conducted to determine which mean(s) is/are different.

Determining Which Mean(s) Is/Are Different

If you fail to reject the null hypothesis in an ANOVA then you are done. You know, with some level of confidence, that the treatment means are statistically equal. However, if you reject the null then you must conduct a separate test to determine which mean(s) is/are different. There are several techniques for testing the differences between means, but the most common test is the Least Significant Difference Test.

Least Significant Difference (LSD) for a *balanced* sample: $\sqrt{\frac{2 * MSE * F_{1,N-c}}{r}}$, where MSE is the mean square error and r is the number of rows in each treatment.

In the example above, LSD = $\sqrt{\frac{(2)(1709)(5.99)}{3}} = 82.61$

Thus, if the absolute value of the difference between any two treatment means is greater than 82.61, we may conclude that they are not statistically equal.

Compact cars vs. Midsize cars:
$|666.67 - 473.67| = 193$. Since $193 > 82.61$ → mean head pressure is statistically different between compact and midsize cars.

Midsize cars vs. Full-size cars:
$|473.67 - 447.33| = 26.34$. Since $26.34 < 82.61$ → mean head pressure is statistically equal between midsize and full-size cars.

Compact vs. Full-size:
Work this on your own.

One-way ANOVA in Excel

You may conduct a one-way ANOVA using Excel.

(Preliminary step) First, make sure that the "Analysis ToolPak" is installed.
Under "Tools" is the option "Data Analysis" present?
If yes – ToolPak is installed.
If no – select "Add-ins."
Check the boxes entitled "Analysis ToolPak" and "Analysis ToolPak – VBA" and click "OK". This will install the "Data Analysis ToolPak."

(1.) Under "Tools" select "Data Analysis"
In the window that appears select "ANOVA: One factor" and click "OK."
(2.) Using your mouse highlight the cells containing the data.
(3.) Select "Columns" if each treatment is its own column or "Row" if each treatment is its own row.
(4.) Set your level of significance. (The default is 5% or 0.05.)
(5.) Click "OK" and the ANOVA output will appear on a new worksheet.

Two Way ANOVA

A Two- Way ANOVA is an extension of the One- Way ANOVA. With a One Way, you have one independent variable affecting a dependent variable. With a Two-Way ANOVA, there are two independents. Use a two-way ANOVA when you have one measurement variable (i.e. a quantitative variable) and two nominal variables. In other words, if your experiment has a quantitative outcome and you have two categorical explanatory variables, a two way ANOVA is appropriate.

Assumptions for Two Way ANOVA

- The population must be close to a normal distribution.
- Samples must be independent.
- Population variances must be equal.
- Groups must have equal sample sizes.

Illustration

Suppose you want to determine whether the brand of laundry detergent used and the temperature affects the amount of dirt removed from your laundry. To this end, you buy two detergents with the different brand ("Super" and "Best") and choose three different temperature levels ("cold", "warm" and "hot").

Then you divide your laundry randomly into "6*r" pile of equal size and assign each 'r' piles into the combination of ("super" and "Best") and ("cold", "warm" and "hot"). In this example, we are interested in testing the Null Hypothesis.

H(oD) = The amount of dirt removed does not depend on the type of detergent.
H(oT) = The amount of dirt removed does not depend on the temperature.

The example has two factors(factor detergent, factor temperature) at a=2(Super and Best) and b=3(cold, warm and hot) levels. Thus, there are a*b = 3*2=6 different combination of detergent and temperature with each combination. There are r=4 loads. (r is called the number of replicates). This sums up to "n=a*b*r"=24=2*3*4 loads in total.

The amounts of Y(ijk) of dirt removed when washing sub pile k(k=1,2,3,4) with detergent i(i=1,2) at temperaturej(j=1,2,3) are recorded in table below:-

	cold	warm	hot
Super	4	7	10
	5	9	12
	6	8	11
	5	12	9
Best	6	13	12
	6	15	13
	4	12	10
	4	12	13

Solution:

	cold	warm	hot	M(d) [Y(i)]
Super	4	7	10	
	5	9	12	
	6	8	11	
	5	12	9	
	mean(Yij)=5	**mean(Yij)=9**	**mean(Yij)=10.5 ~10**	**8**
Best	6	13	12	
	6	15	13	
	4	12	10	
	4	12	13	
	mean(Yij)=5	**mean(Yij)=13**	**mean(Yij)=12**	**10**
M(t)[Y(j)]	5	11	11	9

We have calculated all the means like detergent mean(Md), temperature means (Mt) and mean of every group combination.

Now what we only have to do is calculate the sum of squares(ss) and degree of freedom(df) for temperature, detergent and interaction between factor and levels.

First calculate the SS(within)/df(within) we have already know how to calculate SS(within)/df(within) in one way ANOVA we calculated this but in two way anova the formula is different 🙂

STEP 1: Formula for calculation of SS(within) is:

Yijk is the elements in the groups.

$\bar{Y}_{(ij)}$ is mean of combinations

When we put the values and do calculations with this formula we will get SS(within) is

$$SS_{within} = \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{4}\left(Y_{ijk} - \bar{Y_{ij\cdot}}\right)^2$$

$= (4-5)^2 + (5-5)^2 + (6-5)^2 + (5-5)^2$

$+(7-9)^2 + (9-9)^2 + (8-9)^2 + (12-9)^2$

. . . . . .

$+(12-12)^2 + (13-12)^2 + (10-12)^2 + (13-12$

$= 38$

Calculate the df(within):

$df(within) = (r-1)*a*b = 3*2*3 = 18$

Calculate MS(within):

$MS(within) = SS(within)/df(within) = 38/18 = 2.1111$

STEP 2: Calculate SS(detergent) and df(detergent) and MS(detergent)

$\bar{Y}$(i) is the mean of detergent

$\bar{Y}$ is the total mean detergent and temperature

$$SS_{detergent} = r \cdot b \cdot \sum_{i=1}^{2}\left(\bar{Y_{i\cdot\cdot}} - \bar{Y_{\cdots}}\right)^2$$

$= 4*3[(8-9)^2+(10-9)^2]$

$= 24$

Calculate df(detergent):

$df(detergent) = a-1 = 2-1 = 1$

Calculate MS(detergent):

$MS(detergent) = SS(detergent)/df(detergent)$

$= 24/1 = 24$

STEP 3: Calculate the SS(temperature), df(temperature) and MS(temperature)

$\bar{Y}$(i) is the mean of detergent

$\bar{Y}$ is the total mean detergent and temperature

$$SS_{temperature} = r \cdot a \cdot \sum_{j=1}^{3}\left(\bar{Y_{\cdot j\cdot}} - \bar{Y_{\cdots}}\right)^2$$

$= 4*2*[(5-9)^2 + (11-9)^2 + (11-9)^2]$

$= 192$

= 192

Calculate df(temperature):

df(temperature) = b-1 = 3-1 = 2

Calculate MS(temperature):

MS(temperature) = SS(temperature)/df(temperature)

= 192/2 = 81

STEP 4: Calculate SS(interaction), df(interaction) and MS(interaction)

$\bar{Y}$(ij) is mean of combinations

$\bar{Y}$(i) is the mean of detergent

$\bar{Y}$(j) is the mean of temperature

$\bar{Y}$ is the total mean detergent and temperature

Calculate SS(interaction):

$$SS_{interaction} = r \times \sum_{i=1}^{2}\sum_{j=1}^{3}\left(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots}\right)^2$$

$= 4 \times (5 - 8 - 5 + 9)^2 + (9 - 8 - 11 + 9)^2 + (110 - 8 - 11 + 9)^2 + \cdots + (12 - 11 - 10 + 9)^2$

= 12

Calculate df(interaction):

df(interaction) = (a-1)*(b-1) = (2-1)*(3-1) = 2

Calculate MS(interaction):

MS(interaction) = SS(interaction)/df(interaction)

= 12/2

= 6

Its time to calculate the F-test: Calculate critical F-value

MS(detergent)/MS(within) ~ F(df(detergent), df(within))

MS(temperature)/MS(within) ~ F(df(temperature), df(within))

MS(interaction)/MS(within) ~ F(df(interaction), df(within))

2.9. Nonparametric Test

The term "nonparametric statistics" has been imprecisely defined in the following two ways, among others.

1. The first meaning of *nonparametric* covers techniques that do not rely on data belonging to any particular parametric family of probability distributions.

   These include, among others:

   - *distribution free* methods, which do not rely on assumptions that the data are drawn from a given parametric family of probability distributions. As such it is the opposite of parametric statistics.
   - *nonparametric statistics* (a statistic is defined to be a function on a sample; no dependency on a parameter).

   Order statistics, which are based on the ranks of observations, is one example of such statistics.

2.10. Sign Test

The **Sign test** is a non-parametric test that is used to test whether or not two groups are equally sized. The sign test is used when dependent samples are ordered in pairs, where the bivariate random variables are mutually independent. It is based on the direction of the plus and minus sign of the observation, and not on their numerical magnitude. It is also called the binominal sign test, with $p = .5$.. The sign test is considered a weaker test, because it tests the pair value below or above the median and it does not measure the pair difference. The sign test is available in SPSS: click “menu,” select “analysis,” then click on “nonparametric,” and choose “two related sample” and “sign test.

Assumptions:

- **Data distribution:** The Sign test is a non–parametric (distribution free) test, so we do not assume that the data is normally distributed.
- **Two sample:** Data should be from two samples. The population may differ for the two samples.
- **Dependent sample:** Dependent samples should be a paired sample or matched. Also known as ‘before–after’ sample.

Types of sign test:

1. **One sample:** We set up the hypothesis so that + and – signs are the values of random variables having equal size.
2. **Paired sample:** This test is also called an alternative to the paired sample t-test. This test uses the + and – signs in paired sample tests or in before-after study. In this test, null hypothesis is set up so that the sign of + and – are of equal size, or the population means are equal to the sample mean.

Procedure:

1. Calculate the + and – sign for the given distribution. Put a + sign for a value greater than the mean value, and put a – sign for a value less than the mean value. Put 0 as the value is equal to the mean value; pairs with 0 as the mean value are considered ties.
2. Denote the total number of signs by 'n' (ignore the zero sign) and the number of less frequent signs by 'S.'
3. Obtain the critical value (K) at .05 of the significance level by using the following formula in case of small samples:

$$K = \frac{n-1}{2} - 0.98\sqrt{n}$$

Sign test in case of large sample:

$$Z = \frac{S - np}{\sqrt{np(1-p)}}$$

Binominal distribution formula = ${}^{n}C_{x} q^{n-x}, p^{x}$ with p =1/2

Compare the value of 'S' with the critical value (K). If the value of S is greater than the value of K, then the null hypothesis is accepted. If the value of the S is less than the critical value of K, then the null hypothesis is accepted. In the case of large samples, S is compared with the Z value.

2.11. Run Test of Randomness

Running a Test of Randomness is a non-parametric method that is used in cases when the parametric test is not in use. In this test, two different random samples from different populations with different continuous cumulative distribution functions are obtained. Running a test for randomness is carried out in a random model in which the observations vary around a constant mean. The observation in the random model in which the run test is carried out has a constant variance, and the observations are also probabilistically independent. The run in a run test is defined as the consecutive sequence of ones and twos. This test checks whether or

not the number of runs are the appropriate number of runs for a randomly generated series. The observations from the two independent samples are ranked in increasing order, and each value is coded as a 1 or 2, and the total number of *runs* is summed up and used as the test statistics. Small values do not support suggest different populations and large values suggest identical populations (the arrangements of the values should be random). Wald Wolfowitz run test is commonly used.

Questions Answered:

Does the X group differ from the Y group in regards to the diet treatment implemented on both groups?

Assumptions:

Data is collected from two independent groups.

If the run test is being tested for randomness, then it is assumed that the data should enter in the dataset as an ordered sample, increasing in magnitude. This means that for carrying-out the run test for randomness, there should not be any groupings or other pre-processing.

If the run test is carried out in SPSS, then it is assumed that the variables that are being tested in the run test should be of numeric type. This means that if the test variables are of the string type, then the variables must be coded as numbers in order to make those variables of the numeric type.

Generally, in non-parametric tests, no underlying distribution is assumed. This holds for the run test as well, but if the number of observations is more than twenty, then it is assumed (in the run test) that the underlying distribution would be normal and would have the mean and variance that is given by the formulas as discussed above.

Null Hypothesis: The order of the ones and twos is random.

Alternative Hypothesis: The order of ones and twos is not random.

This checking is done in the following manner:

Let us consider that 'H' denotes the number of observations. The 'H_a' is considered to be the number that falls above the mean, and 'H_b' is considered to be the number that falls below the mean. The 'R' is considered to be the observed number of runs. After considering these symbols, then the probability of the observed number of runs is derived.

Formula of the mean and the variance of the observed number of the runs:

$$E(R) = H + 2 H_a H_b / H$$

$$V(R) = 2 H_a H_b (2 H_a H_b - H) / H^2 (H - 1)$$

The researcher should note that in the run test for the random type of model, if the value of the observations is larger than twenty, then the distribution of the observed number of runs would approximately follow normal distribution. The value of the standard normal variate of the observed number of runs in the run test is given by the following:

$Z = R – E (R) / Stdev (R)$.

This follows the normal distribution that has the mean as zero and the variance as 1. This is also called the standard normal distribution that the Z variate must follow.

2.12. Kruskal Wallis Test

The Kruskal-Wallis test is a nonparametric (distribution free) test, and is used when the assumptions of one-way ANOVA are not met. Both the Kruskal-Wallis test and one-way ANOVA assess for significant differences on a continuous dependent variable by a categorical independent variable (with two or more groups). In the ANOVA, we assume that the dependent variable is normally distributed and there is approximately equal variance on the scores across groups. However, when using the Kruskal-Wallis Test, we do not have to make any of these assumptions. Therefore, the Kruskal-Wallis test can be used for both continuous and ordinal-level dependent variables. However, like most non-parametric tests, the Kruskal-Wallis Test is not as powerful as the ANOVA.

Null hypothesis: Null hypothesis assumes that the samples (groups) are from identical populations.

Alternative hypothesis: Alternative hypothesis assumes that at least one of the samples (groups) comes from a different population than the others.

Example questions answered:

How do test scores differ between the different grade levels in elementary school?

Do job satisfaction scores differ by race?

The distribution of the Kruskal-Wallis test statistic approximates a chi-square distribution, with k-1 degrees of freedom, if the number of observations in each group is 5 or more. If the calculated value of the Kruskal-Wallis test is less than the critical chi-square value, then the null hypothesis cannot be rejected. If the calculated value of Kruskal-Wallis test is greater than the critical chi-square value, then we can reject the null hypothesis and say that at least one of the samples comes from a different population.

Assumptions

1. We assume that the samples drawn from the population are random.
2. We also assume that the observations are independent of each other.
3. The measurement scale for the dependent variable should be at least ordinal.

2.13. Chi-Square Test

Chi-square test is used to compare categorical variables. There are two type of chi-square test

1. Goodness of fit test, which determines if a sample matches the population.

2. A chi-square fit test for two independent variables is used to compare two variables in a contingency table to check if the data fits.

a. A small chi-square value means that data fits

b. A high chi-square value means that data doesn't fit.

The hypothesis being tested for chi-square is

Null: Variable A and Variable B are independent

Alternate: Variable A and Variable B are not independent.

The statistic used to measure significance, in this case, is called chi-square statistic. The formula used for calculating the statistic is

$X2 = \Sigma [(Or,c — Er,c)2 / Er,c]$ where

Or,c = observed frequency count at level *r* of Variable A and level *c* of Variable B

Er,c = expected frequency count at level *r* of Variable A and level *c* of Variable B

Note: As one can see from the above examples, in all the tests a statistic is being compared with a critical value to accept or reject a hypothesis. However, the statistic and way to calculate it differ depending on the type of variable, the number of samples being analyzed and if the population parameters are known. Thus depending upon such factors a suitable test and null hypothesis is chosen.

Illustration

Now let's say we observe drivers at an intersection with a stop sign. We record two pieces of information: whether the driver stops at the stop sign and whether the driver is talking on a cell phone. The question we want to answer is whether drivers are less likely to come to come to a complete stop at the stop sign when talking on a cell phone compared to when drivers are not talking on a cell phone.

Here are the data…

	Cell phone	No cell phone	
stop	5	15	20
don't stop	20	10	30
	25	25	N=50

The question is really whether the ***percentage*** of drivers who stopped while talking on a cell phone is significantly different from the ***percentage*** of drivers who stopped while not talking on a cell phone.

Obviously, 5/25 or 20% of drivers with no passengers came to a complete stop. 15/25, or 60% of drivers with one or more passengers came to a complete stop.

The chi-square test will still tell us if the observed frequencies are significantly different from the expected frequencies. We can use the numbers in the row and column margins to help us to compute the expected frequencies.

The expected frequency for each cell is computed using the following formula:

f(e) = [row total)(column total)] / N

For the top left cell, the expected frequency is:

(20)(25) / 50 = 500/50 = 10

top right:
(20)(25) / 50 = 500/50 = 10

bottom left:
(30)(25) / 50 = 750/50 = 15

bottom right:
(30)(25) / 50 = 750/50 = 15

- Once you have the expected frequencies, the calculation for chi-square is exactly the same except that you have four rows in the calculations instead of two.

	f(o) - f(e)	$[f(o) - f(e)]^2$	$[f(o) - f(e)]^2 / f(e)$
Stop – no pass	5 - 10 = -5	25	2.5
Stop -- pass	5 - 10 = 5	25	2.5
No stop – no pass	20 - 15 = 5	25	1.67
No Stop -- pass	10 - 15 = -5	25	1.67
			8.34

You compare this observed value for chi-square against a comparison value you look up in the chi-square table.

You still use the .05 column because we're using 5% as the odds we're using to make our decision. The number of degrees of freedom to use is found from the following equation:

df = (# of rows - 1)(# of columns - 1)

In this case we have (2-1)(2-1) = (1)(1) = 1, giving us a comparison value of 3.84.

We can say that the percentages found in the first column are significantly different from the percentages found in the second column.

Drivers on a cell phone are significantly less likely to stop at a stop sign than drivers not talking on a cell phone, X^2 (1, N = 50) = 8.34, $p < .05$.

[i] **Summery**

After collection of data using appropriate research instruments, selection of appropriate method of analysis plays a vital role in addressing the research problem. Descriptive and inferential statistical methods are very often found to be suitable in data analysis. Descriptive statistics frequently use the statistical measures such as mean, median, mode, range, standard deviation and variance to describe groups. It describes a sample, i.e., straightforward. Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn. The most common methodologies in inferential statistics are hypothesis tests, confidence intervals, and regression analysis.

Two tailed hypotheses (rejection at both side of the distribution) is called non-directional and one-tailed hypothesis (rejection at either side of the distribution) are called directional

hypothesis. Parametric and non-parametric tests are very often used for hypothesis testing. When data is found to be non-normal and not homogenous, parametric tests are normally preferred, otherwise, non- parametric tests are used. Parametric test for two group samples are Independent T Test and Paired T Test. Mann Whitney Test and Wilcoxon Test are normally used as non- parametric test for two group samples. When Sample size is large, Z Test is also often used instead of independent T Test. For more than two group samples, ANOVA and Repeated Measures are used as parametric tests. Kruskal Wallis and Friedman ANOVA are used as non- parametric tests for more than two samples. Chi-square test is used to compare categorical variables.

Review Question.

Long Questions

1. Compare and contrast between parametric and non-parametric Test with suitable examples.

2. Describe the significance and types of hypothesis Testing with examples

Short type question

1. Explain null and alternative hypothesis with examples.
2. Distinguish between one tail and two tail test.
3. Distinguish between population and sample with examples.
4. Distinguish between descriptive statistics and inferential statistics.
5. Explain the relation between relationship between p-value, critical value and test statistic
6. Explain the assumption of Z Test.
7. Distinguish between T- Test and Paired T Test.
8. Explain the assumption of ANOVA.
9. Distinguish between one way and two way ANOVA.
10. Explain the assumptions of sign Test.
11. Explain the types of sign test.
12. Explain the assumptions of Kruskal Wallis Test.
13. Explain the assumptions of Chi Square Test.

Very short type question

1. Write the significance of critical value
2. What is T Score.
3. What do you mean by P value?
4. How is F value calculated?
5. Write two of the parametric test.
6. Write two of the non-parametric test.
7. What is alfa?
8. Why is chi square test used?
9. Write the equation of one tailed hypothesis testing?

10. Write the equation of two tailed hypothesis testing?
11. Write the formula of variance.
12. How s sampling error calculated?

MODULE III

REPORT WRITING AND PRESENTATION

3.0. Learning objectives

After the end of this unit, the students will be able to understand:

3.1. Dependence and Interdependence techniques

3.2. Factor Analysis

3.3. Discriminant Analysis

3.4. Research Report writing

3.1. Introduction

This section discussed about univariate, bivariate and multivariate research methods for data analysis and report writing. There are two kinds of multivariate techniques: (a) dependence techniques and; (b) interdependence techniques. In dependence techniques, we check the relationship between one or more dependent variables with one or more independent variables. But in interdependence technique, we don't see this dependence relationship. Rather, we look for the interdependence among the variables. However, In this section, we will discuss a few of the concepts of these techniques such as factor analysis, multiple regression analysis and discriminant analysis to cover the syllabus.

Multivariate Techniques

1. Dependence techniques
    a. Multiple regression analysis
    b. Discriminant analysis
    c. Logistic regression
    d. MANOVA
    e. Conjoint analysis
    f. Canonical correlation
    g. Structural equation modeling
2. Interdependence techniques
    I. Exploratory factor analysis
    II. Cluster analysis
    III. Confirmatory factor analysis
    IV. Multidimensional scaling
    V. Correspondence analysis

3.2. Factor Analysis

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. This technique extracts maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis. Factor analysis is part of general linear model (GLM) and this method also assumes several assumptions: there is linear relationship, there is no multicollinearity, it includes relevant variables into analysis, and there is true correlation between variables and factors. Several methods are available, but principal component analysis is used most commonly.

Types of factoring:

There are different types of methods used to extract the factor from the data set:

1. **Principal component analysis:** This is the most common method used by researchers. PCA starts extracting the maximum variance and puts them into the first factor. After that, it removes that variance explained by the first factors and then starts extracting maximum variance for the second factor. This process goes to the last factor.

2. **Common factor analysis:** The second most preferred method by researchers, it extracts the common variance and puts them into factors. This method does not include the unique variance of all variables. This method is used in SEM.

3. **Image factoring:** This method is based on correlation matrix. OLS Regression method is used to predict the factor in image factoring.

4. **Maximum likelihood method:** This method also works on correlation metric, but it uses maximum likelihood method to factor.

5. **Other methods of factor analysis:** Alfa factoring outweighs least squares. Weight square is another regression based method which is used for factoring.

Factor loading:

Factor loading is basically the correlation coefficient for the variable and factor. Factor loading shows the variance explained by the variable on that particular factor. In the SEM approach, as a rule of thumb, 0.7 or higher factor loading represents that the factor extracts sufficient variance from that variable.

Eigenvalues: Eigenvalues is also called characteristic roots. Eigenvalues shows variance explained by that particular factor out of the total variance. From the commonality column, we can know how much variance is explained by the first factor out of the total variance. For

example, if our first factor explains 68% variance out of the total, this means that 32% variance will be explained by the other factor.

Factor score: The factor score is also called the component score. This score is of all row and columns, which can be used as an index of all variables and can be used for further analysis. We can standardize this score by multiplying a common term. With this factor score, whatever analysis we will do, we will assume that all variables will behave as factor scores and will move.

Criteria for determining the number of factors: According to the Kaiser Criterion, Eigenvalues is a good criteria for determining a factor. If Eigenvalues is greater than one, we should consider that a factor and if Eigenvalues is less than one, then we should not consider that a factor. According to the variance extraction rule, it should be more than 0.7. If variance is less than 0.7, then we should not consider that a factor.

Rotation method: Rotation method makes it more reliable to understand the output. Eigenvalues do not affect the rotation method, but the rotation method affects the Eigenvalues or percentage of variance extracted. There are a number of rotation methods available: (1) No rotation method, (2) Varimax rotation method, (3) Quartimax rotation method, (4) Direct oblimin rotation method, and (5) Promax rotation method. Each of these can be easily selected in SPSS, and we can compare our variance explained by those particular methods.

Assumptions:

1. **No outlier:** Assume that there are no outliers in data.
2. **Adequate sample size:** The case must be greater than the factor.
3. **No perfect multicollinearity:** Factor analysis is an interdependency technique. There should not be perfect multicollinearity between the variables.
4. **Homoscedasticity:** Since factor analysis is a linear function of measured variables, it does not require homoscedasticity between the variables.
5. **Linearity:** Factor analysis is also based on linearity assumption. Non-linear variables can also be used. After transfer, however, it changes into linear variable.
6. **Interval Data:** Interval data are assumed.

Key concepts and terms:

Exploratory factor analysis: Assumes that any indicator or variable may be associated with any factor. This is the most common factor analysis used by researchers and it is not based on any prior theory.

Confirmatory factor analysis (CFA): Used to determine the factor and factor loading of measured variables, and to confirm what is expected on the basic or pre-established theory. CFA assumes that each factor is associated with a specified subset of measured variables. It commonly uses two approaches:

1. **The traditional method:** Traditional factor method is based on principal factor analysis method rather than common factor analysis. Traditional method allows the researcher to know more about insight factor loading.
2. **The SEM approach:** CFA is an alternative approach of factor analysis which can be done in SEM. In SEM, we will remove all straight arrows from the latent variable, and add only that arrow which has to observe the variable representing the covariance between every pair of latents. We will also leave the straight arrows error free and disturbance terms to their respective variables. If standardized error term in SEM is less than the absolute value two, then it is assumed good for that factor, and if it is more than two, it means that there is still some unexplained variance which can be explained by factor. Chi-square and a number of other goodness-of-fit indexes are used to test how well the model fits.

Illustration

Say you have a list questions and you don't know exactly which responses will move together and which will move differently; for example, purchase barriers of potential customers. The following are possible barriers to purchase:

1. Price is prohibitive
2. Overall implementation costs
3. We can't reach a consensus in our organization
4. Product is not consistent with our business strategy
5. I need to develop an ROI, but cannot or have not
6. We are locked into a contract with another product
7. The product benefits don't outweigh the cost
8. We have no reason to switch
9. Our IT department cannot support your product
10. We do not have sufficient technical resources
11. Your product does not have a feature we require

12. Other (please specify)

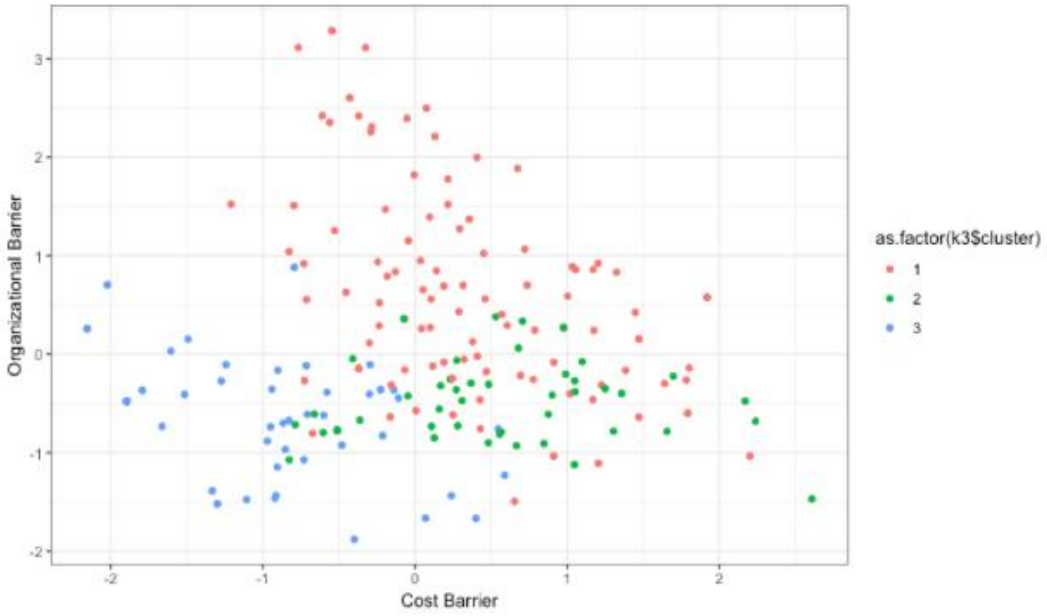
Factor analysis can uncover the trends of how these questions will move together. The following are loadings for 3 factors for each of the variables.

	PC1	PC2	PC3
Price is prohibitive	0.49	-0.16	0.13
Overall implementation costs	0.41	0.38	0.14
We can't reach a consensus in our organization	-0.03	-0.07	0.57
Product is not consistent with our business strategy	0.01	-0.03	0.45
I need to develop an ROI, but cannot or have not	-0.06	0.17	0.62
We are locked into a contact with another product	-0.21	0.08	-0.27
The product benefits don't outweigh the cost	0.68	0.06	-0.27
We have no reason to switch	-0.07	-0.77	-0.25
Our IT department cannot support your product	0.00	0.38	-0.25
We do not have sufficient technical resources	-0.21	0.58	-0.12
Your product does not have a feature we require	-0.40	0.05	0.08
Other (please specify)	-0.50	-0.06	-0.10

Notice how each of the principal components have high weights for a subset of the variables. The first component heavily weights variables related to cost, the second weights variables related to IT, and the third weights variables related to organizational factors. We can give our new super variables clever names.

	Cost	IT	Org
The product benefits don't outweigh the cost	0.68	0.06	-0.27
Price is prohibitive	0.49	-0.16	0.13
Overall implementation costs	0.41	0.38	0.14
We do not have sufficient technical resources	-0.21	0.58	-0.12
Our IT department cannot support your product	0	0.38	-0.25
Product is not consistent with our business strategy	0.01	-0.03	0.45
We can't reach a consensus in our organization	-0.03	-0.07	0.57
I need to develop an ROI, but cannot or have not	-0.06	0.17	0.62
Your product does not have a feature we require	-0.4	0.05	0.08
We are locked into a contact with another product	-0.21	0.08	-0.27
Other (please specify)	-0.5	-0.06	-0.1
We have no reason to switch	-0.07	-0.77	-0.25

If we were to cluster the customers based on these three components, we can see some trends. Customers tend to be high in Cost barriers or Org barriers, but not both.



3.2. Multiple Regression Analysis

Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables- also called the predictors.

For example, the yield of rice per acre depends upon quality of seed, fertility of soil, fertilizer used, temperature, rainfall. If one is interested to study the joint affect of all these variables on rice yield, one can use this technique.

An additional advantage of this technique is it also enables us to study the individual influence of these variables on yield.

Dependent and Independent Variables

By multiple regression, we mean models with just one dependent and two or more independent (exploratory) variables. The variable whose value is to be predicted is known as the dependent variable and the ones whose known values are used for prediction are known independent (exploratory) variables.

The Multiple Regression Model

In general, the multiple regression equation of Y on X_1, X_2, …, X_k is given by:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots\ldots\ldots\ldots\ldots\ldots + b_k X_k$$

Interpreting Regression Coefficients

Here b_0 is the intercept and b_1, b_2, b_3, …, b_k are analogous to the slope in linear regression equation and are also called regression coefficients. They can be interpreted the same way as slope. Thus if $b_i = 2.5$, it would indicates that Y will increase by 2.5 units if X_i increased by 1 unit.

The appropriateness of the multiple regression model as a whole can be tested by the F-test in the ANOVA table. A significant F indicates a linear relationship between Y and at least one of the X's.

How Good Is the Regression?

Once a multiple regression equation has been constructed, one can check how good it is (in terms of predictive ability) by examining the coefficient of determination (R2). R2 always lies between 0 and 1.

R_2 - coefficient of determination

All software provides it whenever regression procedure is run. The closer R_2 is to 1, the better is the model and its prediction.

A related question is whether the independent variables individually influence the dependent variable significantly. Statistically, it is equivalent to testing the null hypothesis that the relevant regression coefficient is zero.

This can be done using t-test. If the t-test of a regression coefficient is significant, it indicates that the variable is in question influences Y significantly while controlling for other independent explanatory variables.

Assumptions

Multiple regression technique does not test whether data are linear. On the contrary, it proceeds by assuming that the relationship between the Y and each of X_i's is linear. Hence as a rule, it is prudent to always look at the scatter plots of (Y, X_i), i= 1, 2,…,k. If any plot suggests non linearity, one may use a suitable transformation to attain linearity.

Another important assumption is non- existence of multicollinearity- the independent variables are not related among themselves. At a very basic level, this can be tested by computing the correlation coefficient between each pair of independent variables.

Other assumptions include those of homoscedasticity and normality.

Multiple regression analysis is used when one is interested in predicting a continuous dependent variable from a number of independent variables. If dependent variable is dichotomous, then logistic regression should be used.

3.3. Discriminant Analysis

Discriminant analysis is a technique that is used by the researcher to analyze the research data when the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature. The term categorical variable means that the dependent variable is divided into a number of categories. For example, three brands of computers, Computer A, Computer B and Computer C can be the categorical dependent variable.

The objective of discriminant analysis is to develop discriminant functions that are nothing but the linear combination of independent variables that will discriminate between the categories of the dependent variable in a perfect manner. It enables the researcher to examine whether significant differences exist among the groups, in terms of the predictor variables. It also evaluates the accuracy of the classification.

Discriminant analysis is described by the number of categories that is possessed by the dependent variable.

As in statistics, everything is assumed up until infinity, so in this case, when the dependent variable has two categories, then the type used is two-group discriminant analysis. If the dependent variable has three or more than three categories, then the type used is multiple discriminant analysis. The major distinction to the types of discriminant analysis is that for a two group, it is possible to derive only one discriminant function. On the other hand, in the case of multiple discriminant analysis, more than one discriminant function can be computed.

There are many examples that can explain when discriminant analysis fits. It can be used to know whether heavy, medium and light users of soft drinks are different in terms of their consumption of frozen foods. In the field of psychology, it can be used to differentiate between the price sensitive and non price sensitive buyers of groceries in terms of their psychological attributes or characteristics. In the field of business, it can be used to understand the characteristics or the attributes of a customer possessing store loyalty and a customer who does not have store loyalty.

For a researcher, it is important to understand the relationship of discriminant analysis with Regression and Analysis of Variance (ANOVA) which has many similarities and differences. Often we can find similarities and differences with the people we come across. Similarly, there are some similarities and differences with discriminant analysis along with two other procedures. The similarity is that the number of dependent variables is one in discriminant analysis and in the other two procedures, the number of independent variables are multiple in discriminant analysis. The difference is categorical or binary in discriminant analysis, but metric in the other two procedures. The nature of the independent variables is categorical in Analysis of Variance (ANOVA), but metric in regression and discriminant analysis.

The steps involved in conducting discriminant analysis are as follows:

• The problem is formulated before conducting.

• The discriminant function coefficients are estimated.

• The next step is the determination of the significance of these discriminant functions.

• One must interpret the results obtained.

• The last and the most important step is to assess the validity.

3.4. Research Report:

"Research report is a research document that contains basic aspects of the research project".

Research report is a medium to communicate research work with relevant people. It is also a

good source of preservation of research work for the future reference. In simple words, Research report is the systematic, articulate, and orderly presentation of research work in a written form. Many times, research findings are not followed because of improper presentation. Preparation of research report is not an easy task. It is an art. It requires a good deal of knowledge, imagination, experience, and expertise. It demands a considerable time and money.

3.4.1. Definition

Research report is the final stage of every research in which research procedure, analysis, findings and so forth aspects of research endeavors are presented in organized and systematic way. It is the process of scientific and professional communication regarding research findings. The general purpose of research report is to convey the sufficient details of research works. It not only convinces the readers but let them known about the findings of already carried out research or project work or the purpose of the work have been done. According to Krishna Swami "research report is a formal statement of a research process and its result." Writing a report is both an art as well as science so that it pertain certain skills, rules and format suited for proper delivery in orderly and scientific manner. Effective report deserves:

1. Uniformity,
2. Consistency, and
3. Regularity

Neuman (2006) states that a research report is a written document (or oral presentation based on a written document) that communicates the methods and findings of a research project to others. It is more than a summary of findings; it is a record of the research process. In addition to findings, the report includes the reasons for initiating the project, a description of the project steps, a presentation of data, and a discussion of how the data relate to the research question or topic.

3.4.2. Purpose

Research report is an indispensable task of every research work in which findings of a research make known to others. Needs or purposes of research report can be outlined as follow:

- To provide the information regarding the findings of research work i.e. methods, data analysis, conclusion and so on in the systematic, scientific and accepted way.
- To elicit crucial facts for solution derived and decision making.
- To prove the worth and legitimacy of assigned research job.

- To provide the judgement tools for the judgement of quality and talent of researcher within and outside the academia.
- To communicate the research findings professionally.
- To pertain the credibility of the research.
- To develop appreciation of standards, consolidate arguments and identify the knowledge gaps.

3.4.3. Types

"Research report can vary differently in its length, type and purpose. Kerlinger (2004) states that the results of a research investigation can be presented in number of ways via a technical report, a popular report, a monograph or at times even in the form of oral presentation." Some typology of research reports are more popular for business purposes can be as:

1. Formal and Informal report
2. Written and Oral report
3. Internal and external report
4. long and short report
5. Descriptive and Analytical report
6. Technical and popular report

But, for the academic report like Thesis, GRP or Project reports, only either descriptive or analytical report is prepared. A short description of each type of description and analytical report is given below:

1. Descriptive Report

In descriptive report, researcher describes the facts, trends or opinions experienced or gathered during the research work. In such reports, data presentation and analysis are more importantly presented. Such reports are more suitable in case of describing current situations, etc. It is more popular method of report writing.

2. Analytical report

As name given analytical, such reports are prepared with analyzing and interpretation of the facts or trends or situations. This means analytical report is one step ahead than descriptive reports. Such reports follow the scientific investigation and reporting. Analytical reports also recommend some measures to improve the situation with stating different problems on the

situation. Policy research and managerial research which are normally funded by any agencies seeking solution of prevailing problems demand analytical report.

3.4.4. Structure/ Components

Scientific research articles provide a method for scientists to communicate with other scientists about the results of their research. A standard format is used for these articles, in which the author presents the research in an orderly, logical manner. This doesn't necessarily reflect the order in which you did or thought about the work. The following is a general outline for a research report.

Beginning Material:	i.e. title page, abstract, key word list, table of contents, list of figures and tables, acknowledgements
Chapter 1:	Introduction – statement of the problem, hypotheses, why it is important, objectives of the work, scope of the work
Chapter 2:	Background and Literature Review – discuss related work and indicate how it relates to report
Chapter 3:	Procedure – describe the procedure used in project, data used, and how it was obtained
Chapter 4:	Results – indicate what happened and interpret what it means
Chapter 5:	Conclusions and Recommendations – summarize conclusions and what they mean (i.e., answer the question, "So what?"). What changes and further work do you recommend?

A research report should also have a relevant title, researchers names and affiliation details, an abstract or executive summery for giving the readers a 'preview' of the report. It should also acknowledge the ladies and gentlemen, institutes etc before the beginning of the chapters. It should also contain 'list of tables, 'list of figures', definition of abbreviation used in the research report. The first chapter should also start with a beautiful introduction to answer the following questions, although not in depth:

a) What is the research about?

b) Why is it relevant or important?

c) What are the issues or problems?

d) What is the proposed solution or approach?

e) What can one expect in the rest of the research?

Foot notes and end notes have to be given wherever necessary. The page set up alignment, font type and size have to be followed in accordance with the provisions of the institute/ publishers

where the report is targeted for submission or publication. Throughout, report writing, research ethics such as plagiarism, acknowledgement of sources of citation following the appropriate style (such as APA, Harvard etc.) must be taken care of. At the end of the research report, questionnaire, policy documents used, interview schedule etc., used in the study should be attached in forms of appendices.

Summery

This unit discussion some of the multivariate techniques used for analysis of research data and research report writing. Some of these techniques are factor analysis, multiple regression analysis and discriminant analysis. Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables- also called the predictors. Discriminant analysis is a technique that is used by the researcher to analyze the research data when the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature. Research report is the systematic, articulate, and orderly presentation of research work in a written form. Many times, research findings are not followed because of improper presentation. Preparation of research report is not an easy task. It is an art. It requires a good deal of knowledge, imagination, experience, and expertise. It demands a considerable time and money. In academic, mainly two types of research reports are used: Descriptive and Analytical. In descriptive report, researcher describes the facts, trends or opinions experienced or gathered during the research work. In such reports, data presentation and analysis are more importantly presented. As name given analytical, such reports are prepared with analyzing and interpretation of the facts or trends or situations. This means analytical report is one step ahead than descriptive reports. Such reports follow the scientific investigation and reporting. The presentation of research report covers, an introduction, review of literature, research methodology, findings and analysis, and summery of findings, recommendations and conclusion.

Long Type Questions

1. Describe the concept of factor analysis. When analysing data through factor analysis. What are the various aspects looked into?
2. Describe the uses of multiple regression analysis in business decision making with a suitable business example.

3. Describe the uses of multiple regression analysis in business decision making with a suitable business example.
4. Describe the uses of discriminant analysis in business decision making with a suitable business example.
5. Describe the concept, types and significance of Research Report.
6. Describe the purpose and structure of a good research report.

Short type Questions

1. Explain the significance of research report.
2. What are the purposes of research report?
3. Explain the characteristics of a good research report.
4. Distinguish between interdependence and dependence techniques.
5. Explain the uses of factor analysis in business research.
6. Explain the uses of multiple regression analysis in business research.
7. Explain the uses of discriminant analysis in business research.
8. Distinguish between R Square and R and their uses in regression.
9. Explain the uses of varimax matrix in principal component analysis.
10. Explain the assumptions of factor analysis.
11. Explain the assumptions of multiple regression analysis.
12. Distinguish between descriptive and analytical research report.
13. Explain the structure of a good research report.

Very short types Questions.

1. Why is research report important?
2. Write two criteria of a good research report.
3. What is coefficient of determination?
4. Write the equation of multiple regression analysis.
5. What is a discriminant function?
6. What does beta stand for in multiple regression analysis?
7. What does factor- loading in factor analysis state?
8. What does eigen value in factor analysis state?
9. What do you mean by multi collinearity?