Handbook
on
Usage of Statistical Test and Software in Business Data Analysis – Beginners Module

Research Committee
The Institute of Chartered Accountants of India
(Set up by an Act of Parliament)
New Delhi

Handbook
on
Usage of Statistical Test and Software in Business Data Analysis – Beginners Module

Research Committee
The Institute of Chartered Accountants of India
(Set up by an Act of Parliament)
New Delhi

© THE INSTITUTE OF CHARTERED ACCOUNTANTS OF INDIA

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form, or by any means, electronic, mechanical, photocopying, recording, or otherwise without prior permission, in writing, from the publisher.

Edition	:	July, 2020
Committee/Department	:	Research Committee
E-mail	:	research@icai.in
Website	:	www.icai.org
Price	:	₹110/-
ISBN	:	978-81-8441-993-1
Published by	:	The Publication Directorate on behalf of The Institute of Chartered Accountants of India, ICAI Bhawan, Post Box No. 7100, Indraprastha Marg, New Delhi – 110 002 (India)
Printed by	:	Sahitya Bhawan Publications, Hospital Road, Agra – 282 003 July | 2020 | P2706 (New)

Foreword

Business Data Analysis includes the activities that help managers to take informed strategic decisions, achieve major goals and solve complex problems by collecting, analysing and reporting the most useful information relevant to requirements. These activities can include identifying and verifying potential strategies, solutions and testing the feasibility of the most favoured solutions. Analysis are based on relevant, accurate and reliable information, often involving interactive and automated statistical analysis or data analysis.

In the world of new age business intelligence ecosystem, statistical analysis is a smart way for organisations to collect, analyze and present data and make better decisions. In this direction, I am happy to note that the Research Committee of the Institute of Chartered Accountants of India (ICAI) is bringing out this handbook on '**Usage of Statistical Test and Software in Business Data Analysis – Beginners Module**'. This handbook is an effort to bring all the relevant topics of business research at one place for the ease of reference to the users of this handbook.

I compliment CA. Anuj Goyal, Chairperson, CA. Kemisha Soni, Vice-Chairperson, and members of the Research Committee for their invaluable contribution in preparation of this Beginners Module for the benefit of members and other stakeholders.

I hope that this handbook will be immensely helpful to the members as well as to the others concerned.

New Delhi
June 29, 2020

CA. Atul Kumar Gupta
President, ICAI

Preface

Statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to reject a conjecture or hypothesis about the process. Software are a set of tools used to retrieve, analyse and transform data into useful business insights. Use of appropriate statistical test and software leads to a good analysis of business data for making meaningful decisions.

This handbook on '**Usage of Statistical Test and Software in Business Data Analysis- Beginners Module**' is a compilation of user-friendly understanding of topics of research with easy language and presentation at one place for beginners. It comprises of classification of business research, steps of business research process, definition and types of hypothesis, data collection and entry, process of hypothesis testing.

I am thankful to CA. Atul K. Gupta, President, ICAI and CA. Nihar N. Jambusaria, Vice President, ICAI for inspiring me to conduct this study.

I would like to thank Dr. Vishwas Gupta, Resource Person and Dr. Amit Kumar Agrawal, Secretary, Research Committee for drafting the publication.

My sincere thanks to CA. Kemisha Soni, Vice-Chairperson, Research Committee, CA. Deepak Brij Gupta, Co-opted Member, Research Committee and members of the Committee for their invaluable contribution to bring the publication in present form.

I would also like to compliment the Secretariat of the Committee for their untiring efforts in bringing out this handbook.

I hope that this endeavour of the Research Committee will be immensely helpful to the members as well as to the others concerned.

June 21, 2020
New Delhi

CA. Anuj Goyal
Chairman, Research Committee

Contents

Foreword

Preface

How to use the book

Symbol

1. Introduction to Business Research / Data Analysis 1
2. Steps of Business Research Process 5
3. Hypothesis – Definition and Types 13
4. Process of Hypothesis Testing 26
5. Parametric Test 33
   - (a) Use of Z-test 33
   - (b) Use of T-test 36
   - (c) Use of ANOVA 38
6. Non-parametric Test 42
   - (a) Use of Chi-square test 42
7. Correlation 46
8. Regression 51
9. Factor Analysis 56
10. A Quick Guide to Identify the Use of A Statistical Test 64
11. Terminology 65
12. Statistical Tables 66

How to Use the Book

This book is an effort to bring all the relevant topics of business research at one place for ease of students. Many students have raised an issues that most of the books on research are written with too much details. Some of those details are not required for the beginners. Also, the language used in those books is not user-friendly. In this book, we have tried to overcome those shortcomings. All the relevant topics of research have been brought at one place. Ease of language and easy presentation have been taken care of. The data analysis has been explained in two ways manually as well as with the help of software applications. We have used MS-Excel and SPSS for data analysis.

Students must install the MS office. We recommend MS office 2019. In MS-Excel the student has to add data analysis tool under data tool. Students can click on file→ options → Add Ins → Select analysis tool pack → Click Go →Select analysis tool pack → Click OK.

Also, students must have installed the SPSS. They should prefer SPSS 20.0 or higher version. They can take the help online as to how to install the SPSS.

With both these applications, students can solve different data analytical problems in three ways. It is hoped that the book will guide students to understand the concept of Research easily.

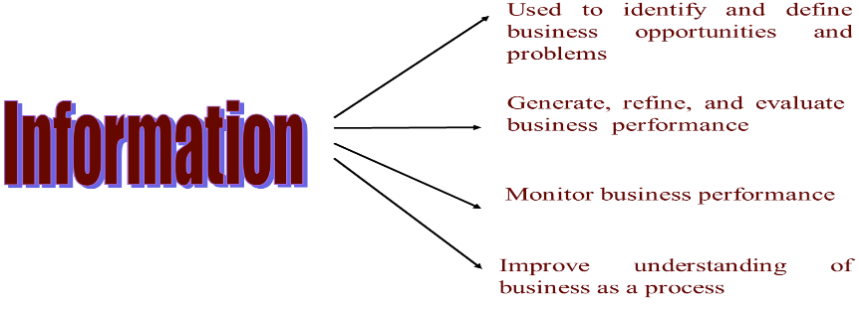
Symbols

χ (Chi)	Used in calculation of Chi-square test
$\sum$ (Summation)	Used to show the sum of different variables
μ (Mew)	Mean of population
df	Degree of freedom
e	Margin of error
H_0	Null hypothesis
H_1	Alternate hypothesis
r	Correlation
R^2	Coefficient of determination
$\bar{x}$	Mean
α (Alpha)	Level of significance & Type I error
β (Beta)	Type II error (1 – β: Power of test)
σ (Sigma)	Standard deviation of population

Chapter 1
Introduction to Business Research / Data Analysis

Research is an essential component for any field of study. The most basic necessity of research is data. In different scientific courses we have laboratories for data collection. However in case of social sciences, the data is collected from the environment with the help of various types of instruments. Environment here refers to Internet, market, consumers and much more. 'Instruments' here refer to the questionnaire, observation, focused group discussion and in-depth interviews etc. The data is collected in the form of detailed report, individual observation details, responses in the form of numerical data etc. The data is analyzed with the help of suitable statistical tools to approve or disapprove the theoretical assumption of research. Based on the findings, recommendations are suggested.

Business research is a systematic and scientific procedure of data collection, compilation, analysis, interpretation and finding the implications pertaining to any business problem. It is an important tool to collect information of business for decision making in any business activity. The information must have certain characteristics as under:



- Accurate
- Current
- Sufficient
- Availability
- Relevant

But now the scenario has changed and business research has become a necessity due to the following reasons:

- Physical separation
- Intermediaries
- Job specialization
- Size of the market
- Complexity of the market
- Need for more information

It aims at collecting the data from customers and market, analyzes it and provide relevant and required information. Therefore, its functions can be divided into description, explanation and prediction & evaluation of data to provide accurate information. It also supplies the necessary inputs for planning for future activity, control of business operations in the present and evaluation of business results.

It is widely practiced in American companies and elsewhere abroad. In India it is still not applied in all companies. In India companies use marketing research mainly for estimating the market potential, analyzing the market share and characteristics of sales. It is also used for forecasting the sales and understanding the business trends.

Classification of Business Research

Business research can be classified on the basis of various criteria. Essentially it is of two types only:

- ***Basic research:*** This type of business research verifies the acceptability of a given theory like performance behaviour analysis and verifying the Pavlov' classic conditioning.
- ***Applied research:*** This type of business research tries to answer the questions about some specific business problems like to know the reasons for product failure and to know the causes for higher attrition rate.

Another classification is based on problem-identification and problem-solving research. In 'problem-identification research' we try to identify the various reasons and causes responsible for the business problems while in case of problem-solving research we try to provide the solution for the business problem. Both types of problems can be analyzed with the help of various approaches (*Fig 2*). We will discuss these approaches in details in the later part of this Chapter.

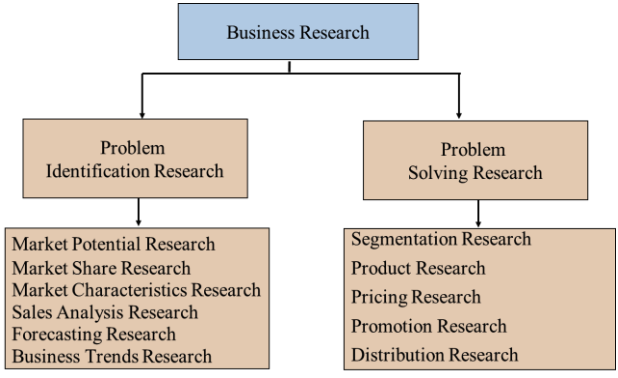


Figure 1: Examples of both types of researches

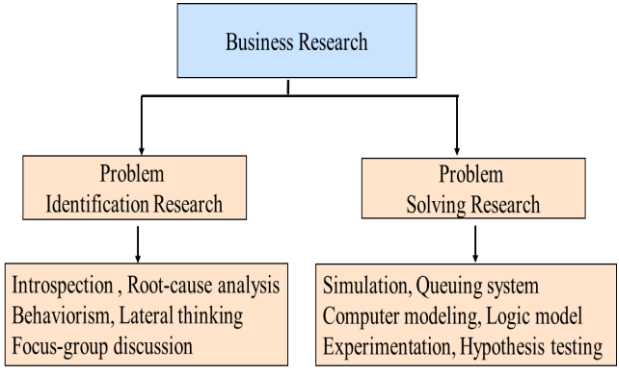


Figure 2: Approaches to both types of researches

Business research is also known as a 'Scientific research study'. It is a systematic, controlled empirical and critical investigation of hypothesis propositions about presumed relations. It provides information that may assist business managers to recognize and react to business opportunities and problems. The various approaches to problem solving are observation technique, define the variables and develop hypothesis, test the hypothesis empirically, and arrive at statistical conclusions.

An empirical research consists of developing of hypothesis based on induction method and accepting or rejecting the hypothesis statistically and then again use the hypothesis to conclude the result in a deductive method.

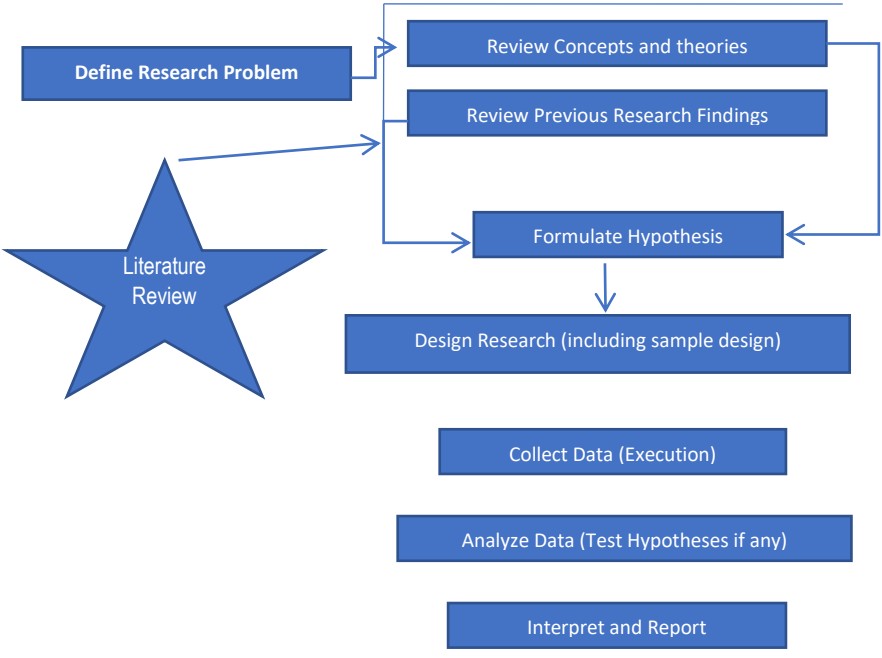
A good research is not necessarily based merely on statistical conclusions. It can be based on theoretical analysis of assumptions. We use various qualitative measures to analyze such business problems like case study methods, ethnography, phenomenology, ground theory and field research methods. Such type of business research is useful when the data studied is unstructured and non-numerical.

In the case of structured and numerical data, we prefer the quantitative measures to analyze the business problems. We use different types of statistical test to test the hypothesis. The tests can be divided into parametric and non-parametric. The selection of a test based on nature of data and research objectives.

Conducting a business research is an art. It requires the creativity and logical reasoning. A sincere and honest effort from a researcher is the prerequisite for a good research. A researcher should try to find the answers on the basis of data available only and his own thoughts and presumptions must be kept aside. The analysis must be done keeping in mind the interests of all the stakeholders. The outcome of the research must be useful for society, business managers, academicians and other researchers.

Chapter 2
Steps of Business Research Process

A proper chronological sequence of steps is required to conduct business research. The research process consists of a series of actions or steps necessary to effectively carry out the research and the desired sequencing of these steps is must.

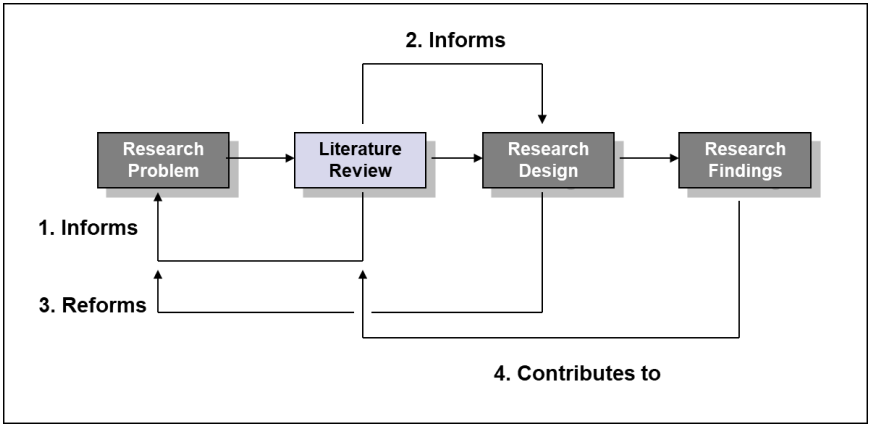


1. **Define Research Problem**: The tentative and general statements of problems obtained in the identification phase are converted into researchable questions and propositions. Clear and unambiguous statements of the problem are made and the information required for research is stated. For this we use deductive as well inductive theory.

2. **Review of Literature**: The previous research reports, research papers and projects are evaluated in this phase. The review of literature tries to answer the following questions:

What are the major issues and debates about the topic? What are the key sources? Who are the key authors? Who influenced them? What are the key

concepts, models, principles, theories and ideas? What are the main questions and problems that have been addressed to date? What are the origins of the topic?



How has it been defined? Official sources; academic definitions. What is the context to the topic? How is knowledge on the topic structured and organized?

The critical literature review is central point of a good research. There are different sources of literature review. They can be classified as under :

Primary Sources: Reports, theses, emails, conference Reports, company reports, some Govt. publications, and unpublished manuscript sources.

Secondary sources: Newspapers, books, journals, and internet.

Tertiary sources: Indexes, abstracts, catalogues, encyclopedia, dictionaries, bibliographies, and citations.

The outcome of literature reviews distinguishes what has been done from what needs to be done, discovering important variables relevant to the topic, synthesizing and gaining a new perspective, identifying relationships between ideas and practice, establishing the context of the topic or problem and rationalizing the significance of the problem.

3. **Research Objectives:** The literature review provides us the information about what work has been done regarding that business problem and what work has not been done so far. The later part is written in the form of literature gap. On the basis of literature gap, we develop the research objectives of our study. The research objectives must be minimum 3 in

numbers. Based on the research objectives we postulate the hypothetical model or research framework for our study. Based on the objectives and framework we develop the hypothesis for our study.

4. Formulate Hypothesis: The problem definitions/propositions are converted into hypothesis or models, which are in testable form to ascertain whether they can be verified statistically or are feasible for solution procedures. Hypothesis and theoretical framework are parts of the research problem formulation.

Variables required for solving the problem are identified, partly from literature and partly by the researcher for defining the problem. The problem is related to the existing research-theories, constructs, and hypothesis in a theoretical framework that will ensure step-by-step progress of knowledge (as in pure research) or a strong basis for the current problem solving (as in applied research).

5. **Research Design / Methodology**: A research design is a framework or blueprint for conducting the marketing research project. It details the procedures necessary for obtaining the information needed to structure or solve marketing research problems. It is the arrangement of conditions for collection and analysis of data in a manner that aims to be relevant to the research purpose with economy in procedure. The design helps a researcher to utilize the available resources efficiently to achieve the research objectives. This involves the following steps:

(i) Determining the type of research to be carried out for data collection: Secondary data, experiment, ex-post facto, or model building.

(ii) Selection of the measurement and scaling of the variables: Whether questionnaire, or observation or interview technique should be used.

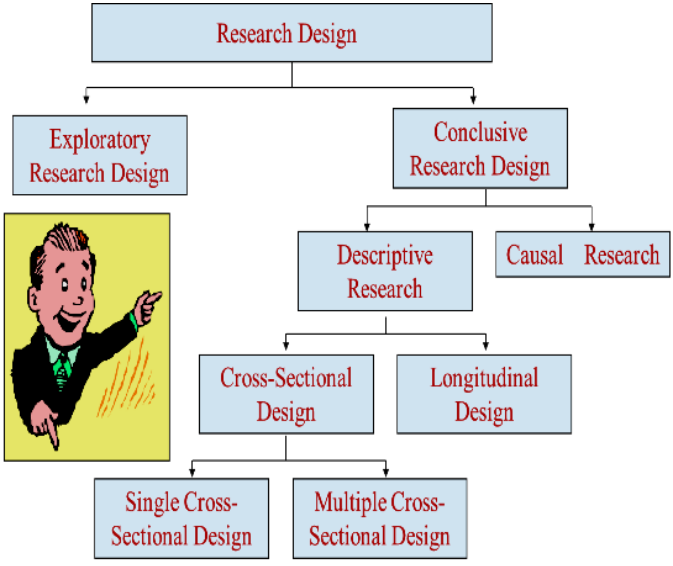
(iii) Selection of the representative sample: Specification of how many respondents, and what kind of respondent or objects to measure.

(iv) Selection of the appropriate method/techniques of analysis of data.

(v) Preparation of a research proposal.

(vi) In addition to the above steps in research design specification of the time schedule of research.

Classification of research design: The research design broadly can be classified into qualitative and conclusive research.



***Qualitative Research**:* The main objective of this research is to provide insights and understanding of the business problem. Information required is roughly defined only. Research process is flexible and unstructured. Sample is small and non-representative. Analysis of primary data is qualitative. Findings are tentative in nature. Generally, outcomes of this research are followed by further exploratory or conclusive research. This research is often used in the front end of total research design. The different methods of this type of research are expert surveys, pilot surveys, case studies, ethnography, grounded theories etc. Different usage of qualitative research is:

(i) Formulate a problem or define a problem more precisely

(ii) Identify alternative courses of action

(iii) Develop hypothesis

(iv) Isolate key variables and relationships for further examination

(v) Gain insights for developing an approach to the problem

(vi) Establish priorities for further research

***Conclusive Research**:* This type of research design is used to test specific hypothesis and examine relationships. The Information required is clearly defined. Research process is formal and structured. Sample is large and

representative. Data analysis is quantitative. Findings are conclusive in nature and used as input into decision making. This research can be further divided into descriptive and conclusive research.

Descriptive Research: The main objectives of this is to describe business characteristics or functions. It is marked by the prior formulation of specific hypothesis. This research is well preplanned and structured design. The different methods used for this research are secondary data, quantitative analysis, surveys, panels and observations. Different usage of descriptive research is:

(i) To describe the characteristics of relevant groups, such as consumers, salespeople, organizations, or market areas

(ii) To estimate the percentage of units in a specified population exhibiting a certain behavior

(iii) To determine the perceptions of product characteristics

(iv) To determine the degree to which marketing variables are associated

(v) To make specific predictions

Causal Research: The main objective of this research is to determine the cause and effect relationships. The manipulation of independent variables and effect on dependent variables are studied here. The controlling of mediating variables and experiments are conducted in this research.

Measurement and Scaling: Research data consists of *quantitative variables* like price, income, sales etc. and, *qualitative variables* like knowledge, power, performance etc. The qualitative measures must be converted into numerical form for further analysis. This is possible through *measurement and scaling techniques*. The assignment of numbers or other symbols to characteristics of objects according to their pre-specified rules is known as measurement while the generation of continuum upon which measured objects are located is known as scaling. Scaling is a tool used to convert qualitative data into quantitative data.

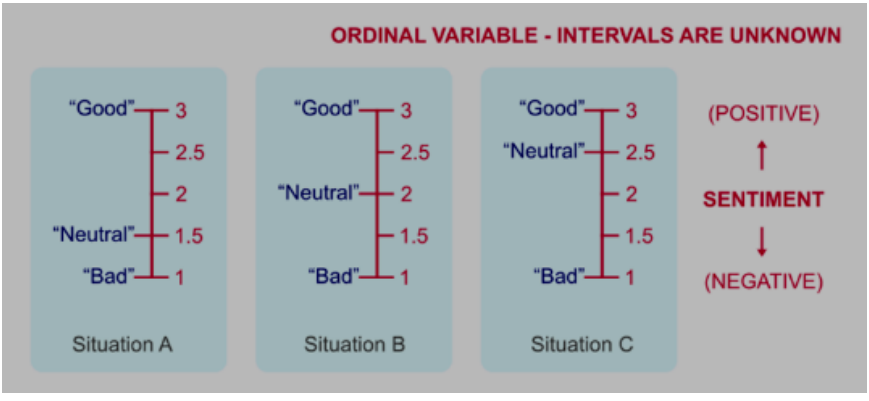
Four different types of measurement scales are widely used. Nominal, ordinal, interval and ratio scales.

Nominal scale: This is also known as categorical scale. Numbers serve as "tags" or "labels" only, to identify or classify an object. Measurements do not make any quantitative distinctions between the observations. "How would

you describe your behavior?" Select '1' for Extrovert and '2' for Introvert and '3' for Ambivert. A respondent selecting '3' on the scale does not imply that Ambiverts are three times better than extroverts.

What is your Gender?	What is your Political preference?	Where do you live?
• M- Male • F- Female	• 1- Independent • 2- Democrat • 3- Republican	• 1- Suburbs • 2- City • 3- Town

***Ordinal scale**:* "Ordinal" indicates "order". It reports the ranking and ordering of the data without actually establishing the degree of variation between them. A restaurant measuring the level of customer satisfaction for the 'new dish' introduced in their menu has asked the following questions while conducting research. The ratings so provided are as: 5 Star-Excellent, 4 Star-Good, 3 Star-Average, 2 Star-Poor, 1 Star-Worst.



***Interval scale**:* A scale in which the numbers are used to rank attributes such that numerically equal distances on the scale represent equal distance in the characteristic being measured. Contains all the information of an ordinal scale, but it also allows one to compare the difference/distance between attributes. It is a measurement scale where the difference between two variables is meaningful. Difference between two variables on a scale is an actual and equal distance. Scale has no true zero point. For example, the difference between '1' and '2' is equal to the difference between '3' and '4'. Further, the difference between '2' and '4' is twice the difference between '1' and '2'. Measuring temperature is an example of interval scale. But we cannot say 40°C is twice as hot as 20°C or Zero-degree temperature does not mean there is no temperature existing.

***Ratio scale**:* Ratio scale allows any researcher to compare the intervals or differences. It possesses a zero point or character of origin. Examples: Sales, profit, wages, units of production, weight, height, age etc. A researcher is undertaking a research project on "Problems faced by people living in Slums in Mumbai". He can ask a question on number of children living in each household.

> Please circle the number of children **under** 18 years of age currently living in your household.
>
> 0 1 2 3 4 5 6 7 (if more than 7, please specify ___.)

Sampling: Data collection is one of the important steps of business research. It requires proper planning and execution. It is very difficult to collect data from each and every member of the population and therefore, we select a small subset of this large data which truly represents the entire population. This dataset is known as sample and the process of collecting the data from this sample is known as sampling technique.

Element Selection Technique	Representation Basis	
	Probability Sampling	Non-Probability Sampling
Unrestricted Sampling	Simple Random Sampling	Convenience Sampling
Restricted Sampling	Complex Random Sampling Cluster, Systematic, Stratified, Multistage, Area	Purposive Sampling Panel, Master Quota, Judgment

Sampling Technique: How to select the sample is known as sampling technique. It has two basic parameters. First the representative basis, if each and every member of population has equal chances to be selected as sample (probability sampling) or not (Non-probability sampling). The second parameter is level of restrictions to the researcher, whether he is free to select any member as sample (unrestricted sampling) or he is not free (restricted sampling). On the basis of these two parameters few sampling techniques have been proposed.

Determining the sample size: Choose a sample that is large enough to approximate all the variation likely in the population and yet, small enough to provide significant savings in cost and time. The size of sample depends on

the importance or the value of information required, the cost involved, the time available and the extent of variation in the population. Two factors affect the sample size:

Confidence level - Degree of confidence you have in results. Most obvious value of it is 95%. It means that 95% research is sure that his results are true.

Confidence interval - Margin of error you are prepared to accept. It is denoted by *e*. Most obvious value of *e* is 10%. It means that there may be 10% errors in your findings.

Different methods are recommended to calculate the appropriate sample size: Ad-Hoc method - Rule of Thumb; 1% sample from the finite population (SS at least 25-30). If you are dividing your population into sub-groups then each sub group will consist of the same percentage. If you are going to apply factor analysis; the sample size must be at least 5 times the number of questions available in the questionnaire. There are few online calculators also available to determine the sample size http://www.surveysystem.com/sscalc.htm and http://www.opinioni.net/sample.php

Here you provide the basic requirement for your research like confidence level, confidence interval, population etc. It will give you the required number of sample size to proceed for data collection.

Chapter 3
Hypothesis – Definition and Types

Hypothesis are presumed answers for the research objectives. Minimum two variables are required to develop a hypothesis. It tests the relationships between two or more variables. Hypothesis tells us the Kind and magnitude of relationship. For example; 'Height (variable) depends on age (variable)', 'Rahul' (case) has a higher-than-average achievement motivation (variable, Students of research class (cases) have a higher-than-average achievement motivation (variable).

To predict a relationship between two variables, we formulate two hypothesis statements. One that describes your prediction and one that describes all the other possible outcomes with respect to the hypothesized relationship. Your prediction is that variable A and variable B will be related (you don't care whether it's a positive or negative relationship). The only other possible outcome would be that variable A and variable B are not related. Usually, we call the hypothesis that you support (your prediction) the alternative hypothesis, and we call the hypothesis that describes the remaining possible outcomes the null hypothesis.

The alternate hypothesis is written as H_1 or H_A, while, the null hypothesis is written as H_0 or H_O.

In case, you are investigating the effects of a new sales training program. You believe one of the outcomes of the program will be higher sales. You want to test the significance of training program on sales. Your two hypothesis might be stated something like this:

The null hypothesis for this study may be:

H_0: As a result of the company sales training program, there will be no significant difference/increase in sales.

This will be tested against the alternative hypothesis:

H_1: As a result of the company sales training program, there will be a significant difference/increase in sales.

(Generally null hypothesis is a negative statement. You prefer to frame null hypothesis in the form of that statement which is incorrect.)

Hypothesis Error: In testing of hypothesis*, a researcher generally may come across some errors. These errors may be in framing the hypothesis or in testing the hypothesis.

Decision	State of Nature	
	H_0 is True (S1)	H_0 is False (S2)
Accept H_0 (A1)	Correct Decision	Type II Error (β)
Reject H_0 (A2)	Type I Error (α)	Correct Decision

There can be four possible outcomes for hypothesis testing:

- The hypothesis is true but test leads to its rejection
- The hypothesis is false but test leads to its acceptance
- The hypothesis is true but test leads to its acceptance
- The hypothesis is false but test leads to its rejection

If we look at all four outcomes, two errors are possible during hypothesis testing. They are known as Type I and Type II error.

Type I error: The hypothesis is true but test leads to its rejection. In a technical term it is (α); level of significance. This is the probability a researcher is ready to accept in his work. (1-α) is called the confidence level.

Type II error: The hypothesis is false but the test leads to its acceptance. In a technical term, it is (β). 1−β and is called the power of the test. α is inversely proportional to β.

(Testing the hypothesis, it refers to null hypothesis always)*

Data Collection and Data Entry

Now when, the required variables have been extracted from literature review, the research objectives have been defined, the research framework has been

proposed, hypothesis have been framed and research methodology has been decided, the questionnaire should be prepared.

The questionnaire must be strictly based on research objectives. It must try to satisfy all and answer issues which are sought to solve that business problem. The close-ended questions should be preferred while dichotomous questions should be avoided. The questionnaire must be divided into few sections. Each section must try to satisfy each of the objectives.

A good questionnaire must be developed in proper format. It must start with declaration stating what you are expecting from the respondent, why are you doing this survey and state that information given by respondents will be kept confidential. It should also provide a little information about the topic or business problem for which you are collecting the data. At the end, a proper acknowledgement note must be added.

The Questionnaire should not be too long. A good questionnaire must not exceed two pages. The languages used for the questions must be grammatically correct and easy to understand. Questions must be arranged chronologically and questions related to demographic data should be at the end of questionnaire. Too much personal questions should be avoided unless they are part of your study.

A one to one survey is mostly recommended because it helps both the respondent as well as researcher to get the best outcome. But nowadays online questionnaire is also preferred, it saves the paper and also helps to collect the data in MS-Excel easily. There are various online websites which provide this type of services, where a researcher can develop the question online and collect the data in soft format in excel sheet. Few of those platforms are Google form, Survey Monkey, Qualtrics, Type form, and Survey Legend.

After collection of data, data management is required.

Data Entry into MS-Excel

If the data has been collected online, the ready-to-work data sheet in MS-Excel format can be downloaded from that online software where the questionnaire has been created. If the data is collected physically, each individual response must be entered in MS-Excel cautiously. The variables or questions must be entered in first row column wise, while the responses of each respondent must be filled from 2nd row onwards. Look at the following

example:

Gender	Age	Education	Preference	Choice	Distance	Price	Discount	Quality
Male	16-22	Graduation	Any one	Any one	Somewhat important	Most Important	Most Important	Doesnt Matter
Female	22-30	Post Graduation	Company Retail Outlet	Any one	Somewhat Important	Somewhat important	Doesnt Matter	Most Important
Male	16-22	Graduation	Company Retail Outlet	Non Branded Only	Least Important	Doesnt Matter	Somewhat Important	Doesnt Matter
Male	16-22	Graduation	Shopping Mall	Branded Only	Somewhat Important	Least Important	Somewhat Important	Least Important
Female	22-30	Post Graduation	Traditional Market	Non Branded Only	Doesnt Matter	Somewhat Important	Least Important	Least Important
Male	30-45	Post Graduation	Company Retail Outlet	Branded Only	Doesnt Matter	Doesnt Matter	Least Important	Doesnt Matter
Male	22-30	Post Graduation	Any one	Branded Only	Doesnt Matter	Least Important	Least Important	Somewhat Important
Male	16-22	10-12	Shopping Mall	Branded Only	Least Important	Somewhat Important	Somewhat Important	Least Important
Male	<16	10-12	Traditional Market	Branded Only	Least Important	Doesnt Matter	Somewhat Important	Least Important
Female	22-30	Post Graduation	Company Retail Outlet	Non Branded Only	Doesnt Matter	Somewhat important	Somewhat Important	Least Important
Female	30-45	Post Graduation	Any one	Non Branded Only	Doesnt Matter	Somewhat Important	Least Important	Doesnt Matter
Female	22-30	Graduation	Any one	Non Branded Only	Somewhat Important	Somewhat important	Least Important	Somewhat Important
Male	16-22	Graduation	Company Retail Outlet	Branded Only	Somewhat Important	Somewhat important	Somewhat Important	Least Important

Now, this text data must be converted into numeric format. We can assign unique number to each option of a question like in case of gender, the male can be assigned 1 and female can be assigned 2 and in case of price we can assign 1 to 'not at all important', 2 to 'least important', 3 to 'neutral', 4 to 'important' and 5 to 'highly important'. The data now will look like this.

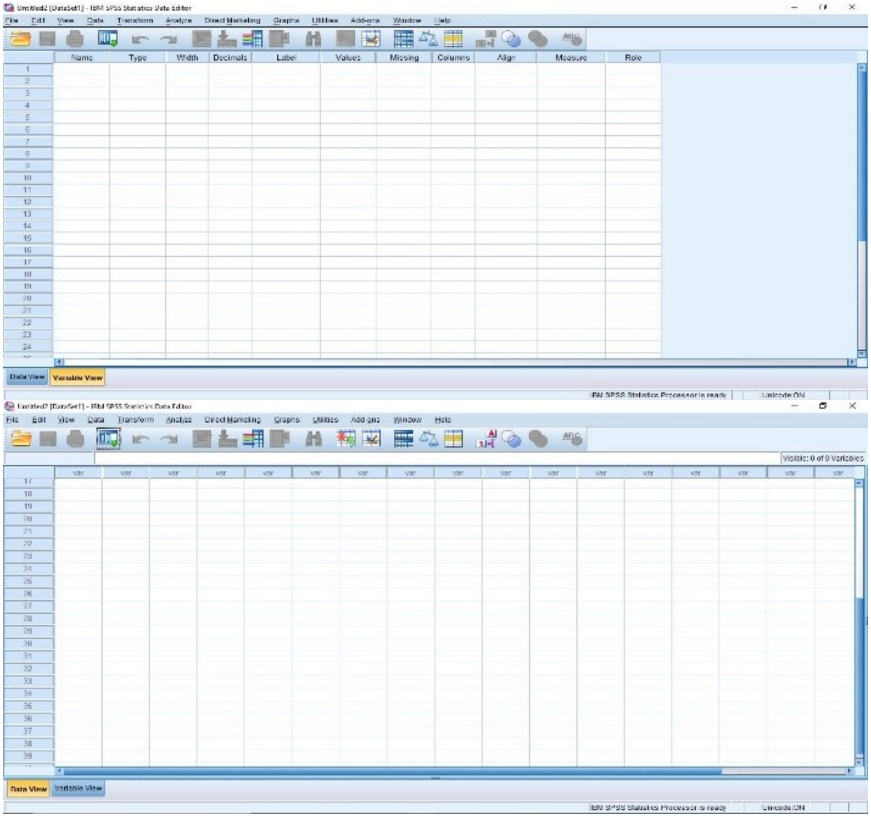
Gender	Age	Education	Prefereence	Choice	Distance	Price	Discount	Quality
1	4	3	5	3	4	5	5	3
1	4	4	1	3	4	4	3	5
1	4	3	1	2	2	3	4	3
2	3	3	2	1	4	2	4	2
1	3	4	3	2	3	4	2	2
1	3	4	4	1	3	3	2	3
1	3	3	1	1	3	2	2	4
1	5	3	5	1	2	4	4	2
1	3	2	2	1	2	3	4	2
2	2	3	3	2	3	4	4	2
2	3	4	1	2	3	4	2	3
2	4	4	5	2	4	4	2	4
1	3	4	5	1	4	4	4	2

Data Entry into SPSS

Now when the Excel data sheet is ready, we have to enter the data into SPSS. We will discuss SPSS 22.0 version here. When you start SPSS in your computer, it opens two different files. One is called data editor and the second data viewer. In data editor we enter or modify the data, while in viewer, we can find the output of statistical analysis.

In data editor, we find two different sheets or tabs. They are data view and variable view. In variable view, the details of variables is entered while in

data view, we can enter the data of individual responses for those variables.

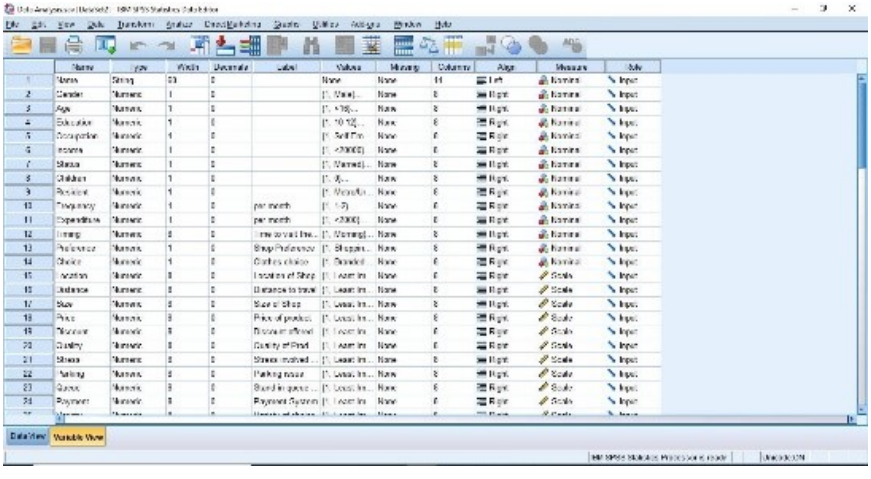


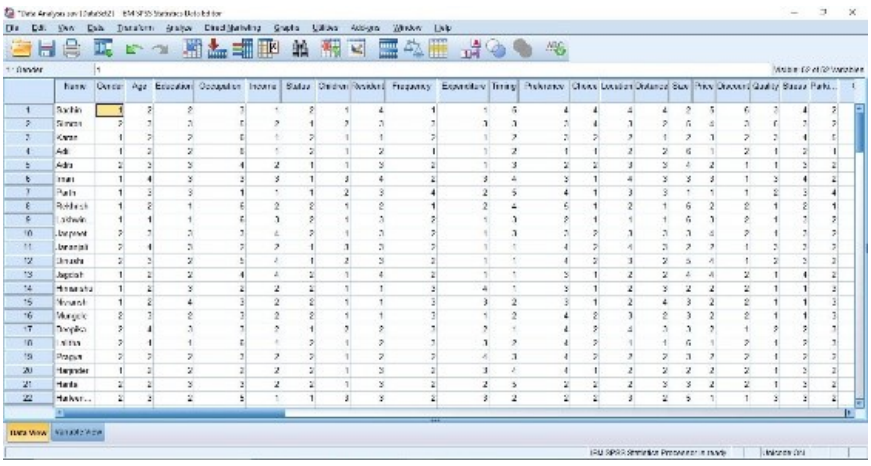
First, we have to enter the variables. This sheet contains the following parameters. A researcher should first provide the details of variables as under :

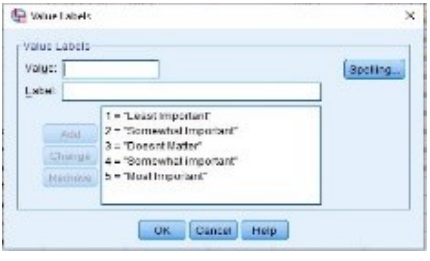
Parameters	Details
Name	Name of variable must be unique and a single word. SPSS does not accept variable name in two words. Name of variable must be clear and short too.
Type	9 different types of variables can be entered into SPSS. Numeric, comma, dots, scientific notation, date, dollar, custom currency, string and restricted numeric.
Width	It shows the number of digits or characters allowed to enter for that variable

Decimals	It gives the number of decimals required in the numeric data
Label	Since the name of variable must be in one word, we can give the proper details of variable here.
Values	This is an important step in data entry. We have to assign the values for each of the options. For example: In case of gender, 1 can be assigned to male and 2 to female.
Missing	Any missing value in that variable is counted here.
Columns	This provides the column number, can be arranged manually.
Align	Alignment of values in the cell can be adjusted here.
Measure	This is also an important step in data entry. SPSS understand three different types of data for different statistical calculations. It accepts nominal and ordinal and ratio data. For interval and ratio data both we can assign measure value as ratio only.
Role	The role of the variable is automatically selected. When you feed the data first time it is considered as input or independent variable, while target is for output or dependent variable

After entering the data, the data editor looks like this:







Before we go for further analysis of data, we have to check the following characteristics of data.

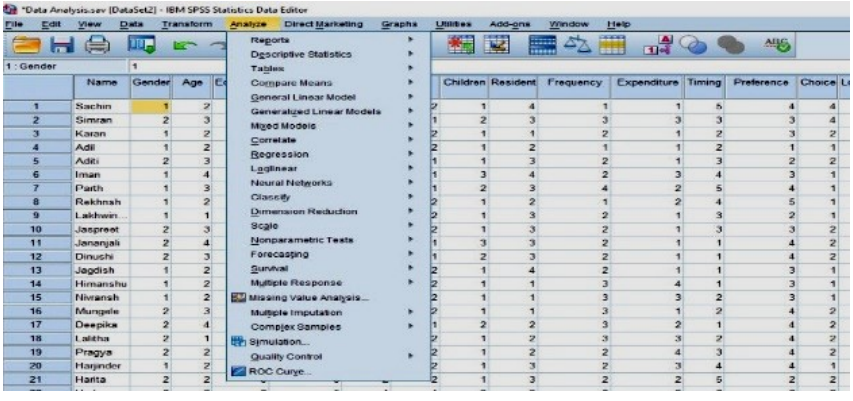
1. We must avoid uniformity (Similar responses for all questions) in the individual responses. To check this, we calculate the standard deviation of all the responses given by a respondent. If the value of standard deviation is near to zero, it means the respondent has given almost same answers to all the questions. We must delete the data of this respondent. (This must be done in MS-Excel only before importing the data into SPSS).

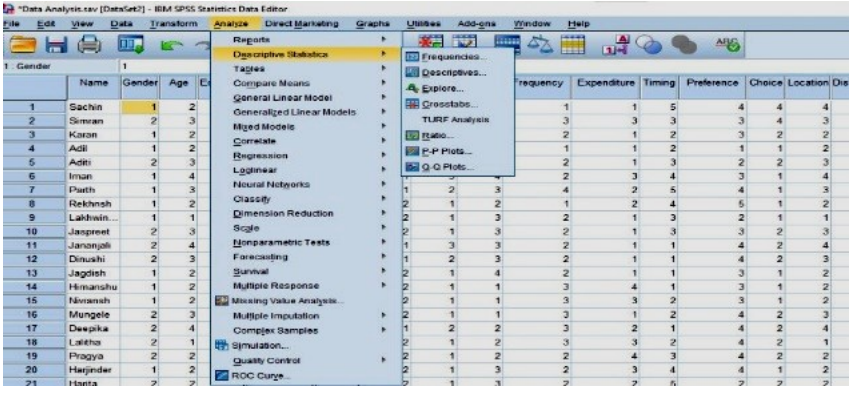
2. We should check the N/A values in each of the variables. The

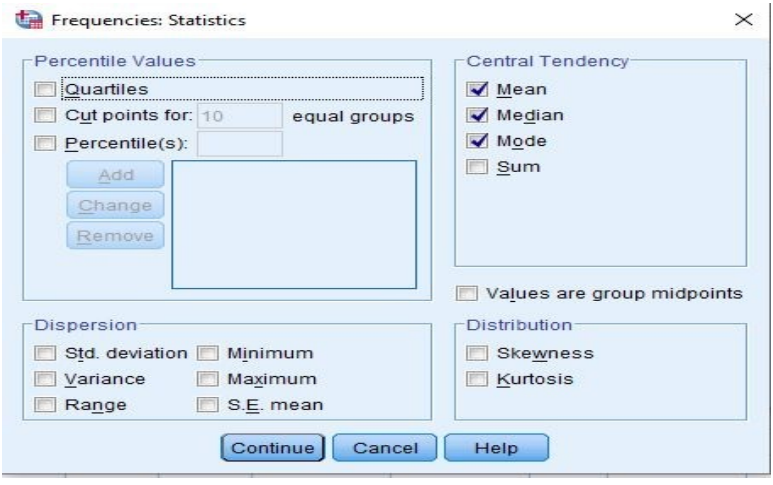
presence of N/A may disturb the statistical calculations. We should replace all N/A with the mean of the rest of responses given by respondents for that variable. (This must be done in MS-Excel only before importing the data into SPSS).

3. That data should be normally distributive. We generally don't get the perfectly normally distributive data; actually it must be slightly normally distributive. Various methods are available to check the normality of data. The easiest way to check the normality is to test the mean, median and mode values of the data. If there is not much difference among the three (remember they will never be exactly equal), the data is considered approximately normally distributed.

Step by step process in SPSS to check the normality of data is as under :



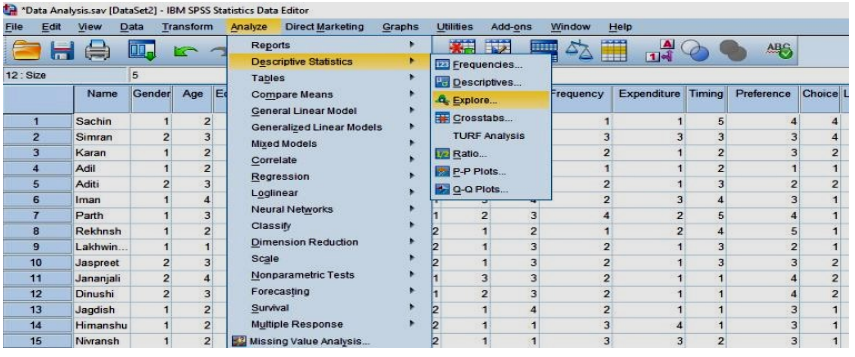


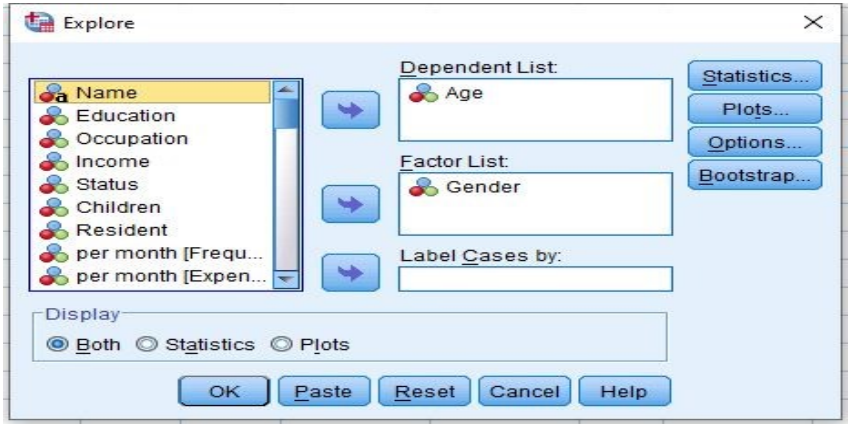


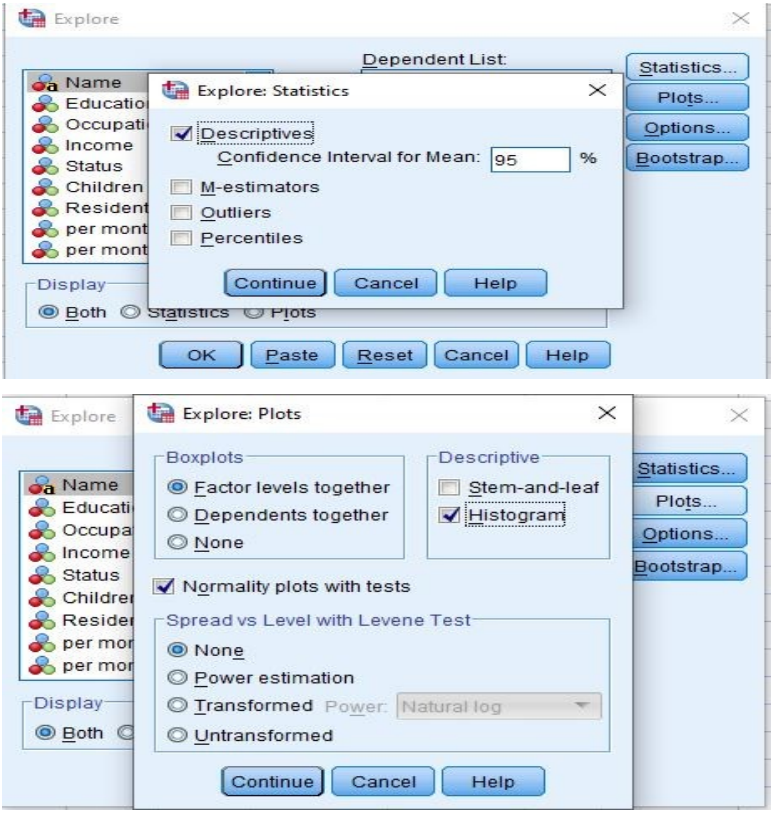
We get the output from the given table. If you look into this table you will find that the values of mean, median and modes do not vary too much for both the variables. In case of gender the mean is 1.4, median is 1.0 while the mode is 1.0. Therefore, we can say that data has approximately normally distributed. In case of age too we can conclude the same output.

Statistics			
		Gender	Age
N	Valid	120	120
	Missing	0	0
Mean		1.4	2.74
Median		1	3
Mode		1	2

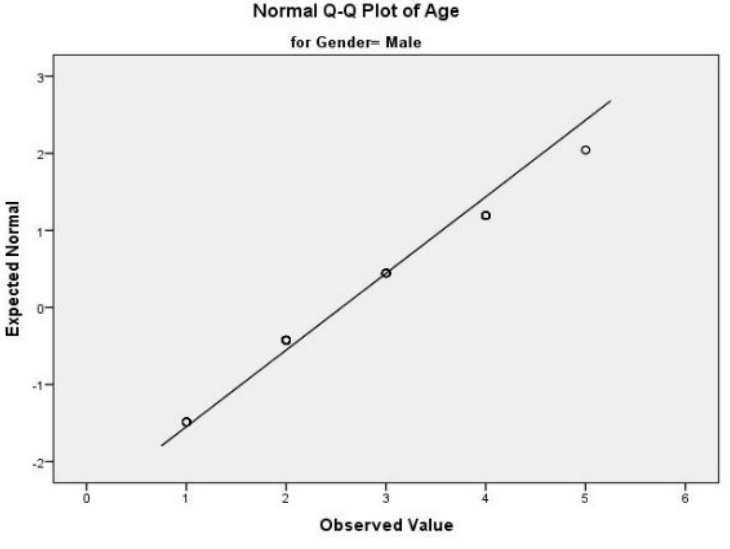
Another method is graphical method. Here we check the Q-Q plot for the variables. We follow the below given steps for this:

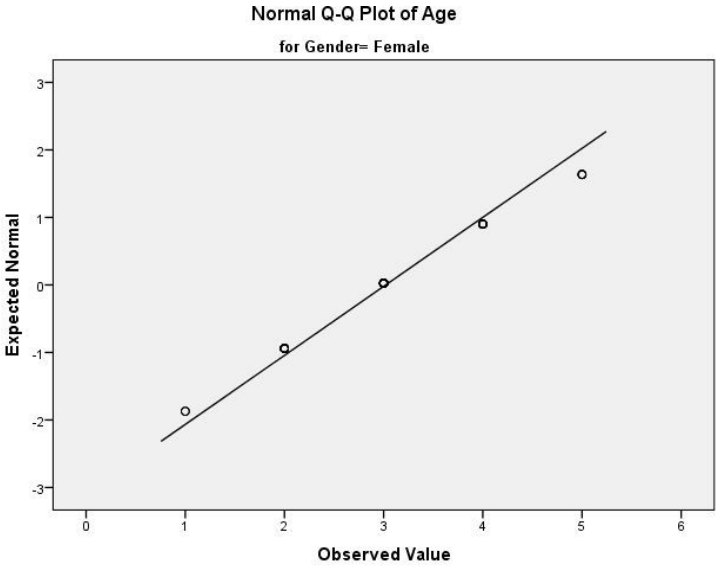






Click on 'continue' after each step. The final output looks like the following:





The observations of both the variables lie almost on the same plane. Therefore, the data is approximately normally distributed.

Data Cleaning: If the data is normally distributed, we can go ahead with further analysis. If we find too much deviation from normal distribution then either we can

1. clean the data (step 1 and 2 given above), or
2. decrease and increase the value of confidence level and confidence interval respectively, or
3. apply non-parametric test

Parametric vs. Non-Parametric Test

Parameters	Parametric Test	Non – Parametric Test
Distribution of Data	Normal Distribution	Data is not normally Distributed
Spread of Data	Different spread	Data has same spread
Info about Population	Completely Known	Unknown
Measurement Level	Interval or Ratio	Nominal or Ordinal
Central Tendency	Mean	Median

Correlation	Pearson	Spearman
Sample Size	Small to Large	Very small
Statistical Power	More Statistical Power	More Robust
Example	t-test, z-test, ANOVA, Pearson correlation	Chi square test, Friedman test, Kruskal-Wallis test, Wilcoxon signed rank test, Spearman correlation

Chapter 4
Process of Hypothesis Testing

Hypothesis are hypothetical statement developed to confirm the answer of research questions. These hypothesis are based on proposed research model. The entire hypothesis testing process is completed through the following steps:

1. Formulate a hypothesis
2. Set up a suitable significance level
3. Choose a test criterion
4. Determine critical value and set the decision rule
5. Collect the sample data
6. Analysis and Comparison
7. Make decisions

1. Hypothesis Formulation

We develop two different hypothesis for each of the problems. The obvious answer of the problem is written as alternative hypothesis while the opposite of it should be preferred as null hypothesis.

1. MRF Tyres has launched a new brand of tyres for tractors and claims that under normal circumstances the average life of the tyres is 80,000 km with standard deviation of 5000 km. A retailer wants to test this claim and has taken a random sample of 30 tyres and found average life as 95000 km.

For this, hypotheses are:

$H_0 : \mu = 80000$

$H_1 : \mu \neq 80000$

2. Set up a suitable significance level

Once the hypothesis are decided, we set up a suitable significance level to test the hypothesis. It is denoted by α.

Significance Level (α) = 1 - Confidence Level

Generally, the confidence level is 95%; therefore, the most obvious value of significance level is 5%. It means that the possibility of rejecting the true null hypothesis (Type I error) is 5 per cent.

2. A company manufacturing automobile tyres finds that tyre-life is normally distributed with a mean of 80,000 km and standard deviation of 5,000 km. It is believed that a change in the production process will result in a better product and the company has developed a new tyre. A sample of 64 new tyres has been selected. The company has found that the mean life of these new tyres is 91,200 km. Can it be concluded that the new tyre is significantly better than old one?

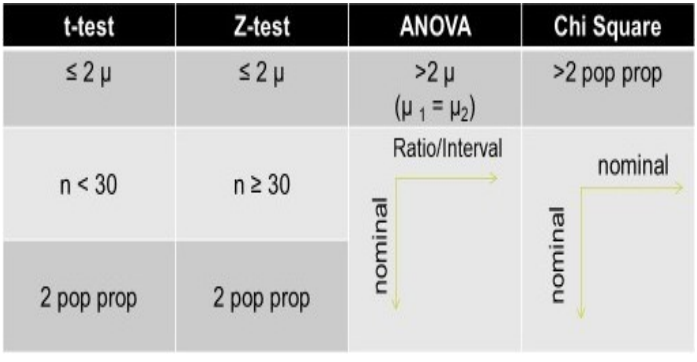
For this, hypotheses are:

$H_0 : \mu = 80000$

$H_1 : \mu > 80000$

3. Choose a test criterion

This is an important step of hypothesis testing process. The selection of suitable statistical test depends on few conditions:



(i) Type of variable (dependent or independent)

(ii) Data type (Nominal, ordinal, interval or ratio)

(iii) Population size (One, two or more)

(iv) Sample size (less than 30 or more)

We can also select the suitable test based on the following scale given here. If the data fulfills all the conditions as shown in the given table, the particular test should be used to check the hypothesis.

t-test - If population or sample proportion is/are one or two only and sample size is less than 30 then data can be collected on nominal, interval or ratio scale. To compare the mean values of one or two population, we prefer this test.

z-test - If population or sample proportion is/are one or two only and sample size is equal to or more than 30 then data can be collected on nominal, interval or ratio scale. To compare the mean values of one or two population, we prefer this test.

ANOVA - If population or sample proportion is more than two we compare the group mean values. In this case one data must be on nominal basis while other data can be based on interval or ratio.

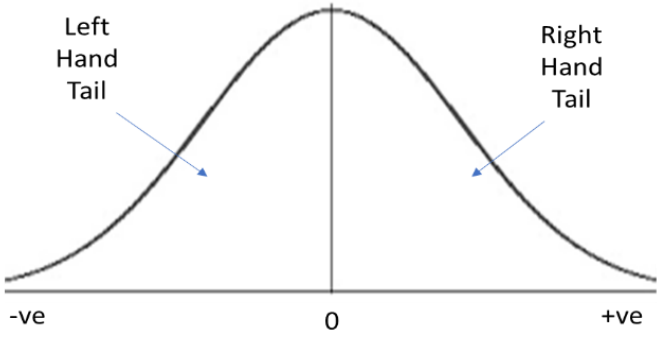
χ^2-test - If population or sample proportion is more than two and we check the relationship among them, both the datasets must be based on nominal scale only.

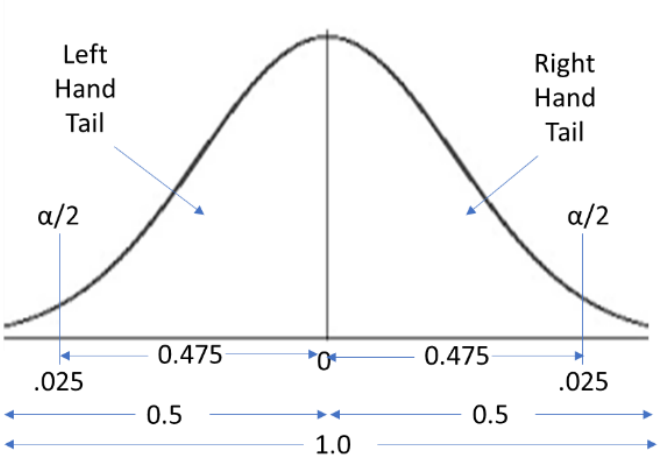
4. Determine critical value and set the decision rule

There are two approaches to determine the critical value and setting the decision rule; critical value and p-value approach.

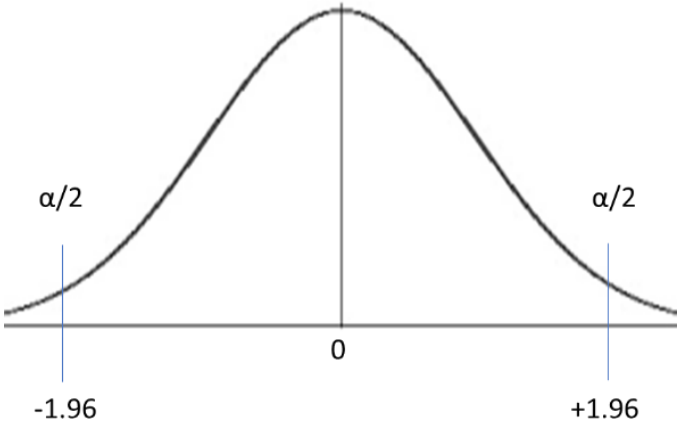
Critical value approach

z-test: There are further two types of decision rules in case of z-test. One-tail test and two-tail test. To understand it, we have to discuss the normal distribution curve. This curve looks like this-





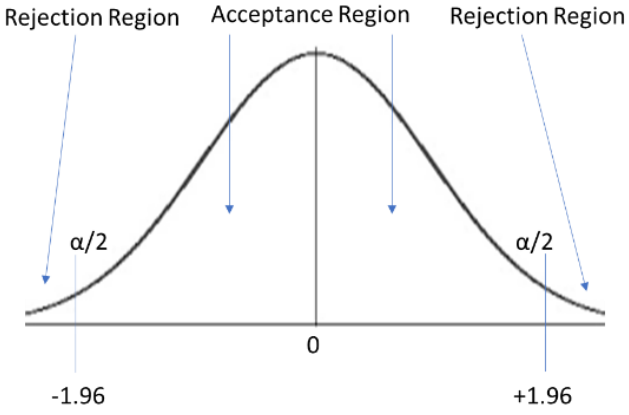
The normal distribution curve can be divided into two equal portions, right-hand tail and left-hand tail. It makes the case as ***two-tailed test.*** In such a case-



(i) The area below the entire curve can be represented as 1.0 (means 100% data lie in this area)

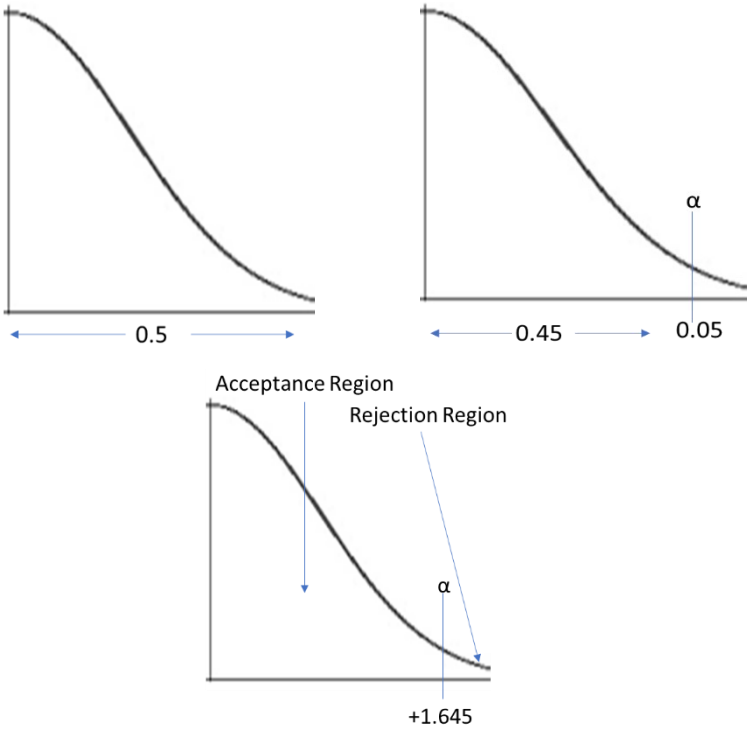
(ii) If data is divided into two equal portions, the area is 0.5 (means 50% data lie in both right-hand and left-hand side).

(iii) Now if the level of significance is 5; in case of two-tailed test it will divided into two equal potions too like α/2. It means both side .025 area will be known as level of significant area.



(iv) Now both side 0.475 area (0.5 -0.025 = 0.475) which will be known as region of acceptance. If we look at the value of 0.4750 in z-table the corresponding tabular value is 1.96. Therefore, the area from +1.96 to -1.96 will be termed as acceptance region while outside area at both sides is known as rejection region. It means if calculated value of z lies in between +/-1.96, it will fall in acceptance region, the null hypothesis will be accepted. If the calculated value is less than -1.96 or more than +1.96, it will fall in area of rejection, the null hypothesis will be rejected. This +/-1.96 is also known as critical value of z.

One tail-test: In case of one-tail test (either left-hand or right-hand tail), we take the data on either side of normal distribution. Here we are taking right-hand tail example for demonstration to get the critical value of z. In such a case



(i) The area below the given curve can be represented as 0.5 (means 50% data lie in this area).

(ii) Now, if the level of significance is 5%, it means .05 area will be known as level of significant area.

(iii) Now, the remaining area of acceptance is left to 0.45. If we look at this value in z-table, the corresponding tabular value is +1.645. Therefore, the area from 0 to +1.645 will be termed as acceptance region while outside area at right side will be known as rejection region. It means if the calculated value of z lies in between 0 to +1.645, it will fall in acceptance region and the null hypothesis will be accepted. If the calculated value is more than + 1.645, it will fall in area of rejection and the null hypothesis will be rejected. This +1.645 is also known as critical value of z in this case.

If we summarize the decision rule for hypothesis, they will look like the following:

If $Z_{calculated} < Z_{critical}$ *: Accept* H_0 & *If* $Z_{calculated} > Z_{critical}$ *: Reject* H_0

If $t_{calculated} < t_{critical}$ *: Accept* H_0 & *If* $t_{calculated} < t_{critical}$ *: Accept* H_0

If $\chi^2_{calculated} < \chi^2_{critical}$ *: Accept* H_0 & *If* $\chi^2_{calculated} < \chi^2_{critical}$ *: Accept* H_0

If $F_{calculated} < F_{critical}$ *: Accept* H_0 & *If* $F_{calculated} < F_{critical}$ *: Accept* H_0

P-value Approach: Another approach to test the hypothesis is P value approach. This is called as probability distribution approach. In this approach we compare the P value with the level of significance.

The decision rules are if

P value < level of significance (α): Reject H_0 & *P value > level of significance (α): Accept* H_0

5. **Collect the sample data**: We collect the data online or physically with the help of questionnaire and do the data entry either in MS-Excel or SPSS.

6. **Analysis and Compare**: We do analyze the hypothesis with the help of suitable tests and compare the calculated values with critical values.

7. **Make decisions**: We accept or reject the null hypothesis and make suitable decisions.

Chapter 5
Parametric Test

(a) Use of Z-Test

A company claims that the mean life of its product is higher than that of a competitor's product. To check the claim, a researcher randomly chooses some products of that company and test the mean life of sample products. If the researcher wants to say that there is no difference in the mean life of products as claimed by the company, he should prefer to apply z or t test.

The basic assumption of z and t test are almost similar. The only difference is sample size. We will use both MS-Excel and SPSS to understand the use of z-test.

Example 1: *A researcher wants to analyze if there any difference in the product image and image of celebrity endorsing the product. The data has been collected on the basis of 5-point Likert scale. The data is collected from 64 respondents. The data is given below where 1 means not at all important, 4 means somewhat not important, 3 means neutral, 4 means somewhat important and 5 means most important.*

Test the difference.

There is a match between the product image and celebrity image	4	3	4	3	1	4	5	3	5	2	2	2	4	4	4	3
	3	3	4	5	4	4	3	3	5	4	4	3	4	5	4	5
	2	5	2	4	4	4	4	4	5	3	4	4	4	4	3	4
	4	4	4	4	4	4	4	3	2	4	3	4	4	4	5	3

We will solve this problem with the help of 7 steps of hypothesis testing:

1. H_0: There is no match between product image and celebrity image.

   H_1: There is match between product image and celebrity image.

2. Leve of significance (α) = 5%

3. Since the population is single here, we are going to analyze the mean value of opinion, and sample size being 64, we will go for z-test

4. The decision rule is if $z_{cal} > z_{critical}$; *Reject H_0 or if P value<.05; Reject H_0*

5. To analyze it, we can use either excel or SPSS

Solution in MS-Excel Data Tool:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		4	3	4	3	1	4	5	3	5	2	2	2	4	4	4	3	
2	There is a match between the product image and celebrity image	3	3	4	5	4	4	3	3	5	4	4	3	4	5	4	5	
3		2	5	2	4	4	4	4	4	5	3	4	4	4	4	3	4	
4		4	4	4	4	4	4	4	3	2	4	3	4	4	4	5	3	
5																		
6	z =	=ZTEST(B1:Q4,3,0.05)																
7																		

6. The P-value = 0

   Therefore, the H_0 is rejected in this case

7. We can conclude that there is a match between product image and brand image. There is no difference among them .

Solution with the help of critical value: First 4 steps will be same as above. Since

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Here

$\bar{x}$ = mean value of observations

μ = mean of population

σ = standard deviation of observation

n = number of observations

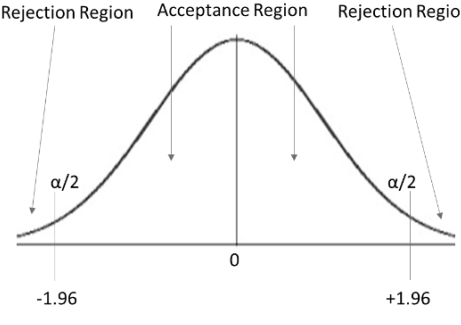
In this example

$\bar{x}$ = 3.6875, μ= 3.0, σ= 0.89, n=64

finally, Z = (3.6875 - 3.0)/0.89/$\sqrt{64}$

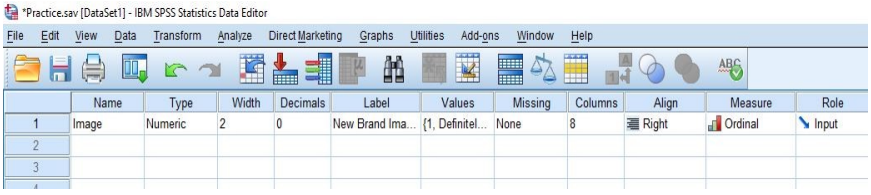
Z_{cal} = 6.189

Since α = .05 the $Z_{critical}$ = +/-1.96

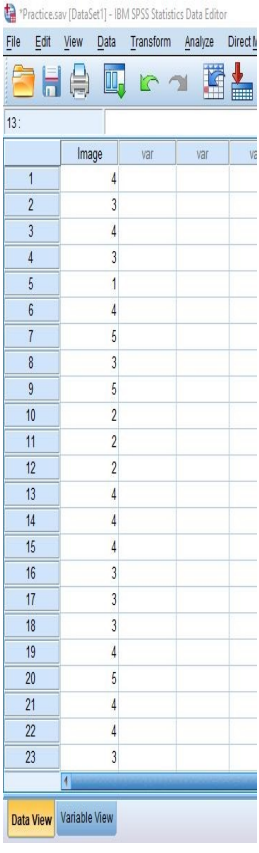


Since $Z_{cal} > Z_{crit}$ and falling in rejection region, the H0 is rejected.

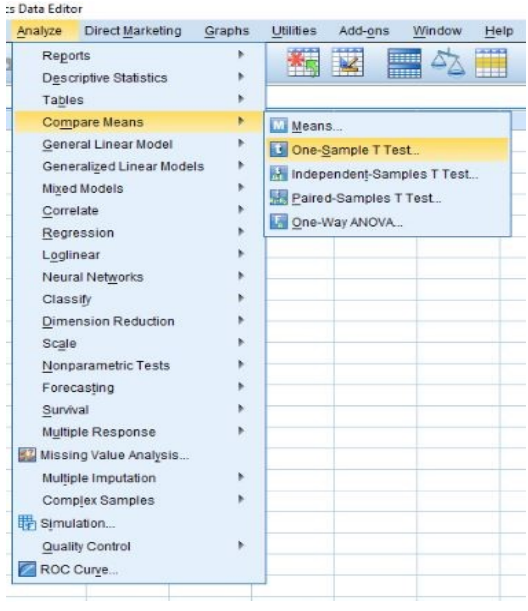
Solution with SPSS: In SPSS we enter the data first and then apply the following steps:



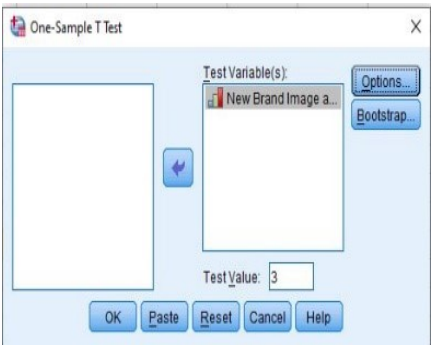
Step 1: Create Variable



Step 2: Insert Data



Step 3: Pick one sample t-test



Step 4: Select the data to analyze, click OK

One-Sample Test

	Test Value = 3					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
New Brand Image and Product Image	6.189	63	.000	.688	.47	.91

Step 5: Final Output

In this case too, the significance value (p value) is coming to 0.00 which is quite significant and we can reject the H_0. Since SPSS does not have z calculation, we can adopt t calculation here too and it does not impact the calculation in case of p-value approach.

Similar steps can be adopted in case of one tail z-test too.

(b) Use of T-Test

In this test too, we often compare the mean value of the one or two population. The only difference from z-test is the sample size.

Example 2: If the number of samples in example 1 is 18 only, we have to apply t-test to check the hypothesis.

There is a match between the product image and celebrity image	4	3	4	3	1	4	5	3	5
	3	3	4	5	4	4	3	3	5

We will solve this problem with the help of 7 steps of hypothesis testing

1. H_0: There is no match between product image and celebrity image.

H_1: There is match between product image and celebrity image.

2. Leve of significance (α) = 5%
3. Since the population is single here, we are going to analyze the mean value of opinion, and as sample size is 18, we will go for t-test.
4. The decision rule is if $t_{cal} > c_{critical}$; *Reject H_0 or if p value<.05; Reject H_0*
5. To analyze it, we can use either excel or SPSS.

Since in Excel, there is no short cut for one sample-test, we can use critical value method and SPSS in this case.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}} \quad \text{where} \quad s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Solution with the help of critical value: To calculate the t-test we use the following formula

Here

$\bar{x}$ = mean value of observations

μ = mean of population

s = standard deviation of sample

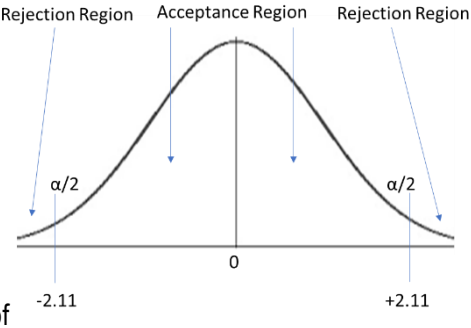
N = number of observations

In this example

To check the critical value of t, we go into t-table

where α/2 = .025 and n-1 (degree of freedom) = 17

Since α = .05 the $t_{critical}$ = +/- 2.11



Therefore, $t_{.025,\ 17}$ = +/- 2.11

If we look into the values, $\bar{x}$ = 3.888889, μ= 3.0, s= 0.832352, n=18, finally, t_{cal} = + 4.531

Since $t_{cal} > t_{crit}$ and fall in rejection region, the H_0 is rejected

Solution with SPSS: In SPSS we enter the data first and then apply the same steps similar to z-test. The final output we get will be as under:

One-Sample Test

	Test Value = 3					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
New Brand Image and Product Image	4.531	17	.000	.889	.47	1.30

In this case too, the significance value (p value) is coming to 0.00 which is quite significant and we can reject the H_0.

(c) Use of ANOVA

This test is used to check whether three or more sample means computed from three different population are equal or not. A researcher may compare more than two sample means in this case. Here we can assume the null hypothesis as – "Three or more population means are equal" or alternatively – "These population means are not equal".

Assumptions:

- Samples are drawn from normally distributed population
- Samples are randomly drawn and are independent of each other
- Population from which samples are drawn have equal variances
- Data must be nominal vs ratio or interval

Example 3*:* One Biscuits company wants to check the retail price per pack in 4 different metro cities. The company wanted to sell at uniform prices and different prices can tarnish the image of the company. To check the prices, the company has collected the data from six different retail stores in each city. The price per pack in each city were

Delhi	Mumbai	Kolkata	Chennai
22	19	18	21
22.5	19.5	17	20
21.5	19	18.5	21.5
22	20	17	20
22.5	19	18.5	21
21.5	21	17	20

Is there any significant difference in prices? Use level of significance as 5%.

We will solve this question too with the help of the above steps only.

1. H_0: There is no significant difference in prices of biscuits among 4 different metro cities

   Or

   H_0: μ1 = μ2 = μ3 = μ4

   H1: There is significant difference in prices of biscuits among 4 different metro cities

   Or

   H1: μ1 ≠ μ2 ≠ μ3 ≠ μ4 or (All the means are not equal)

2. Level of Significance (α) = 0.05

3. Appropriate Test is *F* test in one-way ANOVA (Because populations>2, data is based on nominal and ratio scale).

4. If $F_{cal} > F_{critical}$: Reject H_0

   Degree of freedom: df1 = c - 1 & df2 = n - c (Columns here are nominal data)

   $F_{critical} = F$ (α, df1, df2) = F (0.05, 3, 20) = 3.10

Solution with the help of critical value:

5. To calculate the *F* value we take the help of following table:

Source of Variation	Sum of squares	Degrees of freedom	Mean squares	Test Statistics
Between Samples	SSC	Df_1 = c - 1	$MSC = \frac{SSC}{Df1}$	$F = \frac{MSC}{MSE}$
Within samples	SSE	Df_2 = n - c	$MSE = \frac{SSE}{Df2}$	

	Delhi	Mumbai	Kolkata	Chennai
	22	19	18	21
	22.5	19.5	17	20
	21.5	19	18.5	21.5
	22	20	17	20
	22.5	19	18.5	21
	21.5	21	17	20
Total	132	117.5	106	123.5
Mean ($\bar{x}$)	22	19.58333	17.6667	20.58333
Average of Mean ($\bar{\bar{X}}$)	19.96			
n1 = n2 = n3 = n4 =	6			

SSC (Sum of squares between column) = $\sum_{j=1}^{k} n_j(\bar{x} - \bar{\bar{X}})^2$

$= 6(22 - 19.958)^2 + 6(19 - 19.958)^2 + 6(18 - 19.958)^2 + 6(21 - 19.958)^2 + \ldots\ldots$

$= 25.0104 + 0.847 + 31.5104 + 2.34 = 59.7115$

SSE (Sum of squares within samples) = $\sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \bar{X})2$

$= (22 - 22)^2 + (22.5 - 22)^2 + (21.5 - 22)^2 + (22 - 22)^2 + (22.5 - 22)^2 + (21.5 - 22)^2 + \ldots\ldots$

$+ (21.0 - 20.5833)^2 + (20 - 20.583333)^2 + (21.5 - 20.5833)^2 + (20 - 20.583333)^2 + (21.0 - 20.5833)^2 + (20 - 20.583333)^2$

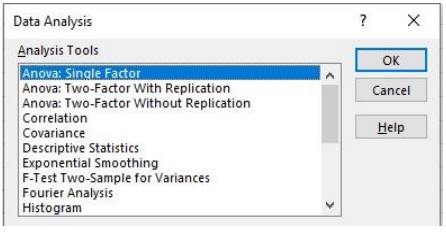
$= 9.25$

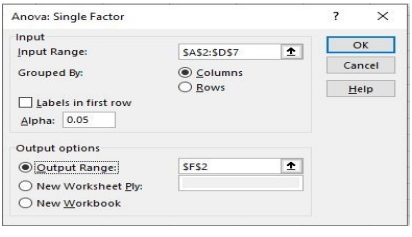
Source of Variation	Sum of squares	Degrees of freedom	Mean squares	Test Statistics
Between Samples	SSC	$Df_1 = 4 - 1 = 3$	$MSC = \frac{59.71}{3} = 19.90$	$F = \frac{19.90}{0.4625} = 43.03$
Within samples	SSE	$Df_2 = 24 - 4 = 20$	$MSE = \frac{9.25}{20} = 0.4625$	

6. $F_{Cal} = 43.03$, $F_{Critical} = 3.10$

$F_{Cal} > F_{Critical}$: Reject H_o.....

There is a significant difference in the price per pack in different cities. The company has to rethink over its pricing strategies.





Solution with Excel Data Tool: Click on data, then click data analysis tool. Select ANOVA: single factor, select the required data set grouped by columns (because it is nominal data), click OK. We get the following output:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	6	132	22	0.2
Column 2	6	117.5	19.58333	0.641667
Column 3	6	106	17.66667	0.566667
Column 4	6	123.5	20.58333	0.441667

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	59.70833	3	19.90278	43.03303	0.00	3.098391
Within Groups	9.25	20	0.4625			
Total	68.95833	23				

p value is 0.00, which is quite significant. The H_0 can be rejected. *There is a significant difference in the price per pack in different cities. The company has to rethink over its pricing strategies.*

***Solution with SPSS**:*

After entering the data in SPSS, we click on 'Analyze', then compare the means and select the one-way ANOVA. In dependent cell feed sales and in the independent cell feed cities. Click OK. We get the following output:

Oneway

ANOVA

Sales

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	59.708	3	19.903	43.033	.000
Within Groups	9.250	20	.463		
Total	68.958	23			

p value is 0.00, which is quite significant. The H_0 can be rejected. *There is a significant difference in the price per pack in different cities. The company has to rethink over its pricing strategies.*

Chapter 6
Non-parametric Test

(a) Use of Chi-Square Test

Chi-square is also known as a test of goodness of fit because it is commonly used to compare observed data with expected data according to a specific hypothesis. With the help of this test we check whether the deviations (differences between observed and expected) are due to the result of chance, or they are due to other factors.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O = *the frequencies observed*

E = *the frequencies expected*

$\sum$ = *the 'sum of'*

The chi-square test always tests the null hypothesis, which states that there is no significant difference between the expected and observed result. It is represented by χ^2.

The formula to calculate the χ^2 is:

Example 4.: 1000 students at college level are graded according to their IQ and their economic conditions. Find out whether there is any association between economic conditions and level of IQ.

Economic Conditions	IQ		
	High	Medium	Low
Rich	160	300	140
Poor	140	100	160

Solution: We will solve this problem with similar 7 steps:

Economic Conditions	IQ (Observed Value)			
	High	Medium	Low	Total
Rich	160	300	140	600
Poor	140	100	160	400
Total	300	400	300	1000

1. H_0: There is no association between economic conditions and level of IQ.

H_1: There is association between economic conditions and level of IQ.

Economic Conditions	IQ (Expected Value)		
	High	Medium	Low
Rich	180	240	180
Poor	120	160	120

2. Level of significance (α) = 5% = 0.05
3. Since population>2, both data (economic conditions and IQ) are taken on nominal scale.
4. If $\chi^2_{(calculated)} > \chi^2_{(critical)}$; Reject the H_0

??5. Degree of freedom (df) = (n-1) * (r-1) = (3-1) * (2-1) = 2 *1 = 2

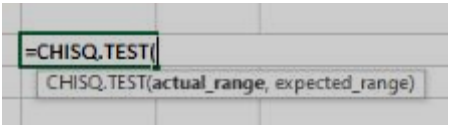
Critical value of $\chi^2 = \chi^2_{(\alpha, df)} = \chi^2_{(.05, 2)} = \chi^2_{Critical} = 5.991$ &

From shown calculation: $\chi^2_{(calculated)} = 65.27778$

6. It is clear that $\chi^2_{(calculated)}$: 65.27778 > $\chi^2_{Critical}$: 5.991, Therefore H_0 is rejected.
7. There is an association between economic conditions and level of IQ

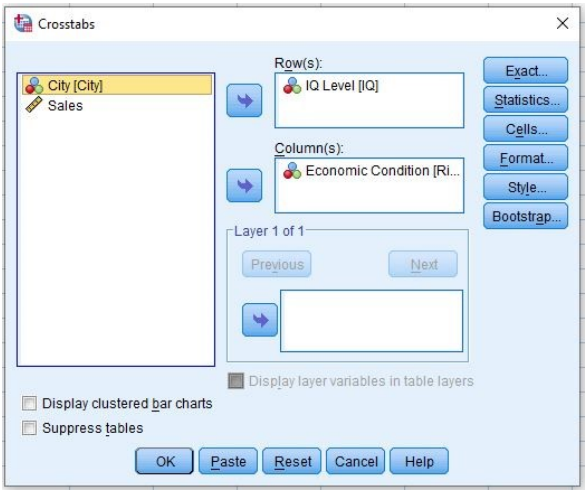
O	E	$(O-E)^2$	$(O-E)^2/E$
160	180	400	2.222222
140	120	400	3.333333
300	240	3600	15
100	160	3600	22.5
140	180	1600	8.888889
160	120	1600	13.33333
		$\chi 2$ =	65.27778

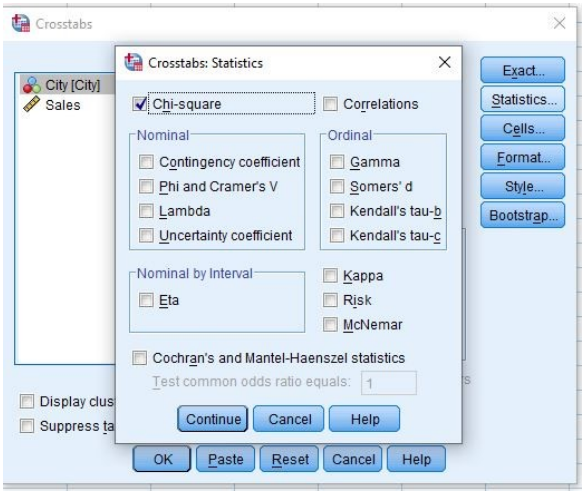
Solution with Data Analysis Tool in MS-Excel: The first 4 steps will be similar as above method. We fit the data into MS-Excel. Make two different tables of observed and expected values. Now we type the χ^2 formula in a cell like the following:

χ^2 = 

We, select the actual or observed values and expected values in the formula and click on 'Enter'. We find the p value of χ^2. In this case, the p value comes as 0.00, which is quite significant. *We can reject the H_0.*

Solution with SPSS: We enter the required data into SPSS. Select 'Analyze', then crosstab.





We continue and click OK. We get the following output:

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	65.278[a]	2	**.000**
Likelihood Ratio	67.047	2	.000
Linear-by-Linear Association	2.775	1	.096
N of Valid Cases	1000		
a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 120.00.			

The calculated value of χ^2 is 65.278 while the p value is quite significant (0.000).

We can reject the H_0.

Chapter 7
Correlation

Correlation is used to measure the strength of relationship between two or more variables. We can measure the relationship between age and height and can conclude that if age increases the height of that person also increases. We can quantify this relationship and identify the direction too. Correlation measures the strength and direction of relationship between two or more variables. It is denoted by r.

Characteristics of Correlation

1. Measured value of correlation is either **positive or negative**, like if

   r_{xy} is +ve: It means that if x increases, y will also increase or if x decreases, y will also decrease.

   r_{xy} is -ve: It means that if x increases, y will decrease or if x decreases, y will increase.

2. **Strength of Correlation**:

   If r_{xy} = +1.0 - Perfect Positive Correlation

   If r_{xy} = -1.0 - Perfect Negative Correlation and

   r_{xy} = 0.0 - No Correlation

   remaining strength are given in this diagram:

Correlation Coefficient Value	Direction and Strength of Correlation
−1.0	Perfectly negative
−0.8	Strongly negative
−0.5	Moderately negative
−0.2	Weakly negative
0.0	No association
+0.2	Weakly positive
+0.5	Moderately positive
+0.8	Strongly positive
+1.0	Perfectly positive

3. **Simple, Partial and Multiple Correlation**: We know height(x) and age(y) are related but few other variables like nutrition(n), climate(c),

heredity(h), also may impact height indirectly. On the basis of these variables, correlation may be as under :

Simple correlation: Relationship between two variables only (r_x)

Multiple correlation: Relationship among more than two variables (r_{xynch})

Partial correlation: Relationship between two variables whilst taking away the effects of another variable, or several other variables, on this relationship ($r_{xy.nch}$)

Measurement of Correlation: There are various methods to measure the correlation.

(i) Scatter Diagram Method

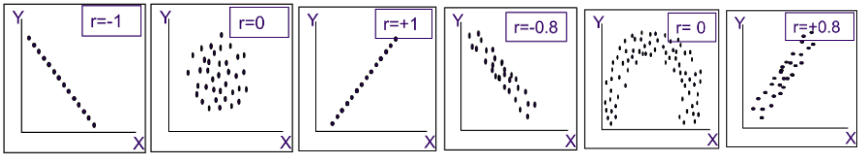
(ii) Graphic Method

(iii) Rank Method

(iv) Concurrent Method

(v) Karl Pearson's Coefficient of Correlation

***Scatter Diagram Method**:* With the help of scatter diagram we can plot each of the individual values of two variables on the same graph. The position of each corresponding dots gives the idea of degree of correlation. With the help of scatter plot we also can identify the linearity and non-linearity of correlation. From the following diagrams we can infer both the above characteristics of correlation.



***Karl Pearson's Coefficient of Correlation**:* This is one of the mathematical methods to compute the relationship between two variables. We compute the correlation with the help of the following formula:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- N = number of pairs of scores
- Σxy = sum of the products of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

Height (cm)	Age (Years)	Weight(Kg)	Cholesterol(mmoL/Lit)
159	16	108	200
167	18	80	100
178	19.5	90	110
178.5	18.6	98	159
178.4	17	70	98
157	17	60	90
159.5	16.5	89	120
168	15	67	110
173	16.3	70	100
160	15.9	60	98
190	18	92	105
160	17	102	178
170	18	67	89
195	19	95	123
193	19	90	105
191.5	16	92	111
167	16	60	100
178	17	70	101
189	18	78	125
191	18	93	131
173	16	67	98
182	17	78	107
192	18	90	110
179	18	88	106
181	17	99	150
169	18	90	128
172	16	78	102

Example 5: The height, age, weight and the cholesterol level of 28 persons have been given in the table. Compute the correlation between weight and cholesterol level.

Solution: We can solve this problem with the help of three different methods:

(i) With the help of formula

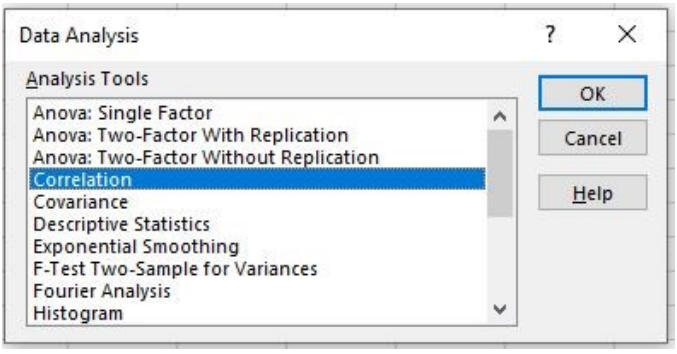
(ii) With data analysis tool in MS-Excel

(iii) SPSS

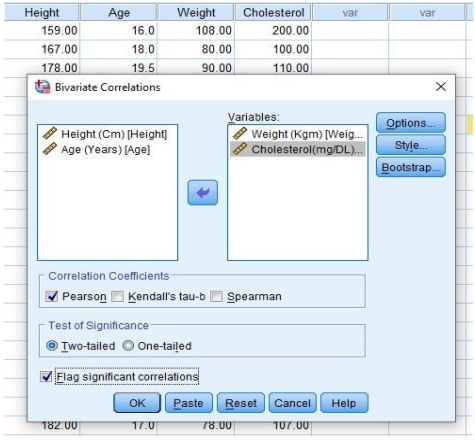
Solution with data analysis tool in MS-Excel: We enter the data into Excel. Click the data analysis tool and select the correlation. We get the following output:

	Weight(Kg)	*Cholesterol(mmoL/Lit)*
Weight(Kg)	1	0.766452778
Cholesterol(mmoL/Lit)	0.766452778	1

r_{wc} =0.77 ; There is a strong positive correlation between weight and cholesterol level. That means if the weight of a person increases the cholesterol level too may increase.



Solution with SPSS: After entering the data into SPSS we click 'Analyze', select correlation and then click on 'Bivariate'. We select the weight and cholesterol data into variables and click on 'Pearson coefficient' and click OK. We get the following output:



Correlations

		Weight (Kgm)	Cholesterol(mg/DL)
Weight (Kgm)	Pearson Correlation	1	.766**
	Sig. (2-tailed)		.000
	N	27	27
Cholesterol(mg/DL)	Pearson Correlation	.766**	1
	Sig. (2-tailed)	.000	
	N	27	27

**. Correlation is significant at the 0.01 level (2-tailed).

The Pearson coefficient of correlation between weight and cholesterol level is 0.766. Also, the p value is 0.00. It means that correlation between weight and cholesterol is significant. It means correlation between weight and cholesterol level is real and not by chance.

***Coefficient of Determination**:* It is denoted by r^2. Mathematically, it can be calculated by square of correlation coefficient.

$$r^2 = r * r$$

It means how much per cent variation will be in other variable in case of change in one variable. r^2 gives you the percentage variation in y explained by x-variables. The range is 0 to 1. It means 0% to 100% of the variation in y can be explained by the x-variables.

Chapter 8
Regression

A statistical measure that attempts to determine the strength of the relationship between one dependent variable and a series of other changing variables (known as independent variables). We can estimate the value of the dependent variable from the independent variable. This measurement is expressed in the form of a mathematical equation.

Type of Regression: It depends on a number of variables.

Linear Regression: Uses one independent variable to explain and/or predict the outcome of dependent variable. Linear regression finds the straight line, called the least squares regression line. Two regression equations can be framed in this case:

Regression equation of Y on X	*Regression equation of X on Y*
$Y = a + bX$	$X = a + bY$
Y: A dependent variable	X: A dependent variable
X: An independent variable	Y: An independent variable
a: Intercept	a: Intercept
b: Slope of line (here it will be b_{yx})	b: Slope of line (here it will be b_{xy})

***Multiple Regression**:* Uses more than two variables to predict the outcome:

Characteristics of Regression

(i) This is a cause and effect relationship.

(ii) The dependent variable Y has a linear relationship to the independent variable X.

(iii) For each value of X, the probability distribution of Y has the same standard deviation σ.

(iv) With the help of regression coefficient, we can calculate the correlation coefficient too.

(v) The slope and Y intercept are known as regression coefficients.

Regression equation of Y on X_1 and X_2

$$Y = a + b_1X_1 + b_2X_2$$

Y: A dependent variable

X_1, X_2: Independent variables

a: Intercept

b_1, b_2: Slope of line

Regression equation of X on Y1 and Y2

$$X = a + b_1Y_1 + b_2Y_2$$

X: A dependent variable

Y_1, Y_2: Independent variables

a: Intercept

b_1, b_2: Slope of line

Computation of regression line: We use Least Square Method to calculate the regression equation. Here we have to calculate the regression coefficient only. Different calculations are there for y on x and x on y.

Regression equation of Y on X

$$Y = a + bX$$

Where

$$a = \frac{\sum y - b\sum x}{n}$$

$$b = \frac{n\sum(xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

x = **The current time period**

n = **The number of time periods**

Regression equation of X on Y

$$X = a + bY$$

Where

$$a = \frac{\sum x - b\sum y}{n}$$

$$b = \frac{n\sum(xy) - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2}$$

x = **The current time period**

n = **The number of time periods**

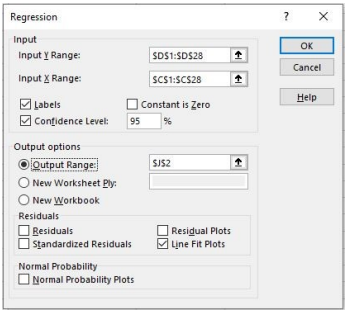
Example 6. : In example5, the age, height, weight and cholesterol level of 28 persons have been given.

(i) Calculate the linear regression between weight and cholesterol level.

(ii) If weight is 105 Kg, what will be cholesterol level in that person?

(iii) If cholesterol level is 160 mmoL/Lit, what will be approximate weight of that person?

Solution: We will solve this problem with the help of data analysis tool in MS-Excel and SPSS.



Solution with data analysis tool in MS-Excel: Click on 'Data', 'Data Analysis Tool', then select 'Regression'. We get the given tab te. Select the appropriate data into

required box as shown in the pic. Click OK.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.766453							
R Square	0.58745							
Adjusted R Square	0.570948							
Standard Error	17.48383							
Observations	27							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	10881.96474	10881.96474	35.59869487	0.00			
Residual	25	7642.109336	305.6843734					
Total	26	18524.07407						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-3.80324	20.49410725	-0.185577478	0.85	-46.01164873	38.40515926	-46.01164873	38.40515926
Weight(Kg)	1.466316	0.245759608	5.966464185	0.00	0.960164511	1.972467284	0.960164511	1.972467284

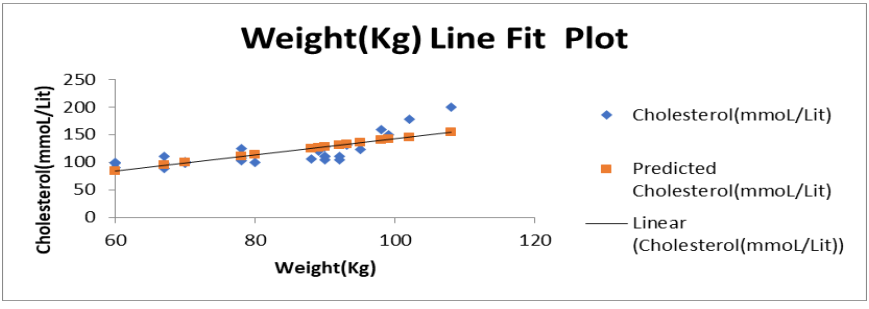
We get the following output: If we take cholesterol level as dependent variable (X), and the weight as independent variable (Y) the required regression equation of Y on X will be:

$Y = a + bX$

As per the output the required regression equation between cholesterol level and weight will be-

$$Y = -3.80324 + 1.466316\ X \text{.....(1)}$$

If we make a plot of predicted dependent values as per the regression coefficient (Line of Best Fit), we get the following deviation from the original values. This deviation occurs due to some random errors.



If we take cholesterol level as dependent variable (X), and the weight as independent variable (Y) then the required regression equation of X on Y will be:

$X = a + bY$

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.766453							
R Square	0.58745							
Adjusted R Square	0.570948							
Standard Error	9.138912							
Observations	27							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	2973.192533	2973.192533	35.59869487	0.00			
Residual	25	2087.992652	83.51970609					
Total	26	5061.185185						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	35.45976	8.038522968	4.411228557	0.000171411	18.90411411	52.01541003	18.90411411	52.01541003
Cholesterol(mmol/l	0.40063	0.06714694	5.966464185	0.00	0.2623381	0.538921522	0.2623381	0.538921522

We will get the following output :

The required regression equation of X on Y will be

$$\boxed{X = 35.45976 + 0.40063\ Y}.....(2)$$

i. The required equation which we finally get will be

Y = -3.80324 + 1.466316 X.....(1)

X = 35.45976 + 0.40063 Y.....(2)

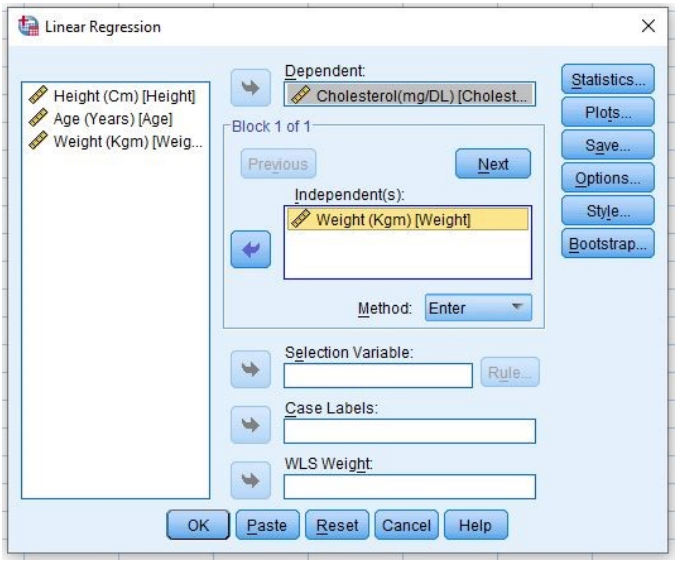
ii. If weight is 105 Kg the cholesterol can be counted on the basis of Y on X equation

Y = -3.80324 + 1.466316 * 105 = ***150.1599 mmoL/Lit***

iii. If cholesterol is 160 mmoL/Lit the weight can be counted on the basis of X on Y equation

X = 35.45976 + 0.40063 * 160 = ***99.5605 Kg***

Solution with SPSS: We click 'Analyze', select 'Regression' and click 'Linear'. To get the regression equation of Y on X, fill the information in given table as given instruction.



From the output we get similar equation of Y on X and X on Y respectively:

Y = -3.803 + 1.466 X &

X = 35.460 + 0.401 Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3.803	20.494		-.186	.854
	Weight (Kgm)	1.466	.246	.766	5.966	.000

a. Dependent Variable: Cholesterol(mg/DL)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	35.460	8.039		4.411	.000
	Cholesterol(mg/DL)	.401	.067	.766	5.966	.000

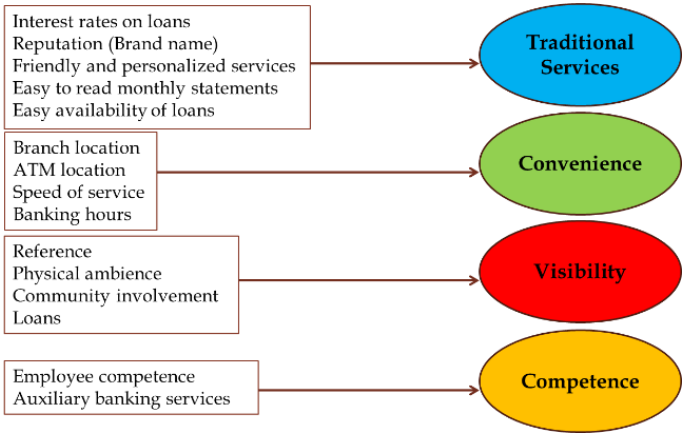
a. Dependent Variable: Weight (Kgm)

Similarly we can develop multiple regression too, keeping cholesterol level as dependent element and age, height and weight as independent elements.

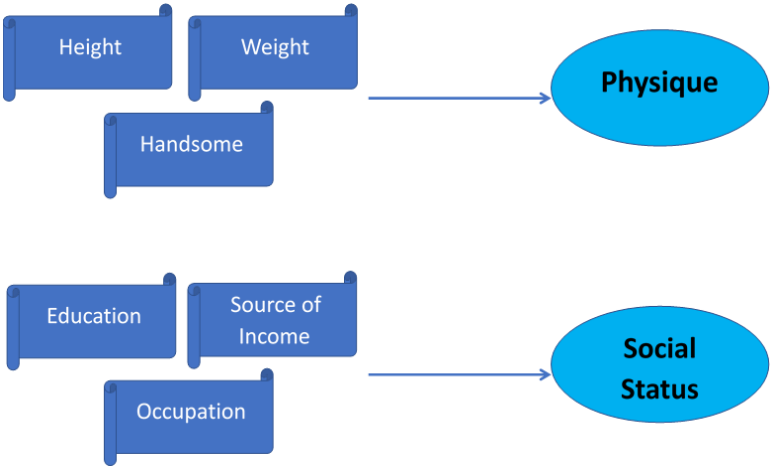
Chapter 9
Factor Analysis

This is one of the most widely adopted research tools in business research. The general objective related to factor analysis is to explore the factors responsible. Factor analysis is actually used for data reduction and summarization technique. From the literature review, we get so many variables related to the research topic. Factor analysis helps to reduce the vast data up to manageable level. Sometimes it is very difficult to deal with so many numbers of variables and therefore we reduce the data by grouping them into different factors.

The data is collected on interval or ratio scale for each of the variables. We use inter-dependence technique and make no difference between independent & dependent variable. The statistical purpose of factor analysis is to determine the linear combinations of variables to find the inter-relationships.



Factor Analysis of Banking Services



Factor Analysis of Personality

For example, if we analyze the different services offered by a bank, we can group them into some factors by applying factor analysis. Similarly, we can group individual personality traits into groups.

Assumptions for Factor Analysis

(i) ***Normality:*** Data must be normally distributive. Factor analysis is actually a part of multivariate analysis. Therefore, before running the factor analysis we much check the normality of data as discussed earlier in this book.

(ii) ***Linearity:*** Factor analysis is a part of general linear model. There should be linear relationship among variables

(iii) ***Factorability of correlation matrix***

(a) There should be sufficient correlations in the data matrix of variables to justify the application of factor analysis.

(b) Number of correlations at 0.01 level of significance can range from 30% to 50% according to the requirement.

(iv) ***No outlier:*** There should not be any outlier in data.

(v) ***Suitable sample size:*** To run the factor analysis, the required sample size must be 5 times the number of variables.

When to apply Factor Analysis?

(i) To identify the underlying dimensions or factors.

(ii) To identify a new, smaller set of uncorrelated variables to replace the original set of correlated variables in subsequent multivariate analysis.

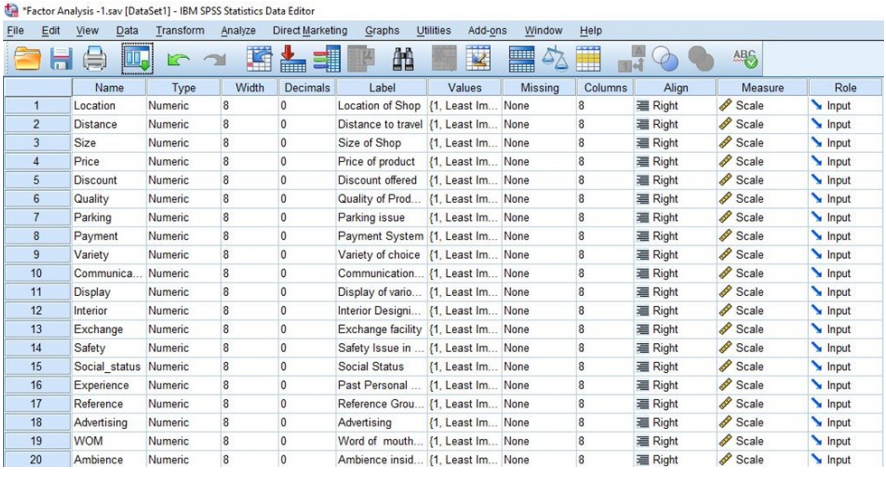
(iii) To identify a smaller set of salient variables from a larger set for use in subsequent multivariate analysis.

How to Run Factor Analysis in SPSS

To run the factor analysis manually is a tough task. It can be done with the help of Excel or SPSS only. SPSS runs it smoothly as compared to Excel. The following steps should be taken to run the factor analysis in SPSS:

Location	Quality	Display	Experience
Distance	Parking	Interior	Reference
Size	Payment	Exchange	Advertising
Price	Variety	Safety	WOM
Discount	Communication	Social_status	Ambience

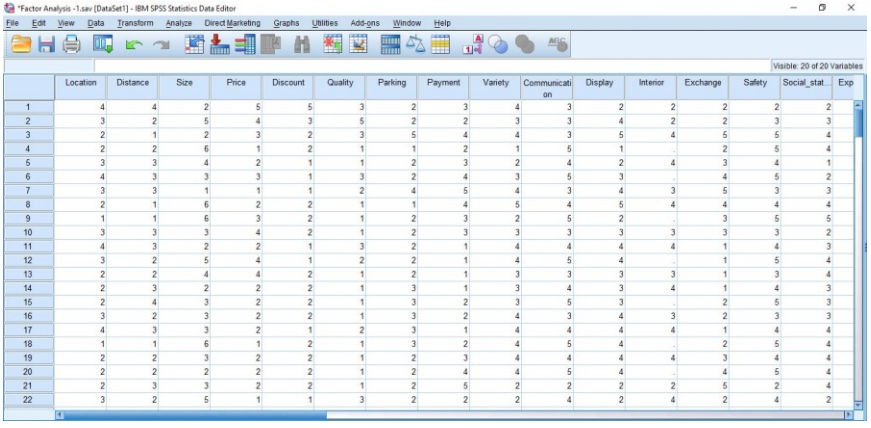
1. Data entry: The data is generally collected on interval scale. We prefer 5-point Likert scale to collect the data. Each variable has been put as separate variable. Here, we will analyze the factors responsible for the retail store selection by consumers. From literature review we got a list 20 variables which impact the store selection decision. The variables are given here. The variables are now entered into SPSS and it looks like this :


*Factor Analysis -1.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Location	Numeric	8	0	Location of Shop	{1, Least Im...	None	8	Right	Scale	Input
2	Distance	Numeric	8	0	Distance to travel	{1, Least Im...	None	8	Right	Scale	Input
3	Size	Numeric	8	0	Size of Shop	{1, Least Im...	None	8	Right	Scale	Input
4	Price	Numeric	8	0	Price of product	{1, Least Im...	None	8	Right	Scale	Input
5	Discount	Numeric	8	0	Discount offered	{1, Least Im...	None	8	Right	Scale	Input
6	Quality	Numeric	8	0	Quality of Prod...	{1, Least Im...	None	8	Right	Scale	Input
7	Parking	Numeric	8	0	Parking issue	{1, Least Im...	None	8	Right	Scale	Input
8	Payment	Numeric	8	0	Payment System	{1, Least Im...	None	8	Right	Scale	Input
9	Variety	Numeric	8	0	Variety of choice	{1, Least Im...	None	8	Right	Scale	Input
10	Communica...	Numeric	8	0	Communication...	{1, Least Im...	None	8	Right	Scale	Input
11	Display	Numeric	8	0	Display of vario...	{1, Least Im...	None	8	Right	Scale	Input
12	Interior	Numeric	8	0	Interior Designi...	{1, Least Im...	None	8	Right	Scale	Input
13	Exchange	Numeric	8	0	Exchange facility	{1, Least Im...	None	8	Right	Scale	Input
14	Safety	Numeric	8	0	Safety Issue in ...	{1, Least Im...	None	8	Right	Scale	Input
15	Social_status	Numeric	8	0	Social Status	{1, Least Im...	None	8	Right	Scale	Input
16	Experience	Numeric	8	0	Past Personal ...	{1, Least Im...	None	8	Right	Scale	Input
17	Reference	Numeric	8	0	Reference Grou...	{1, Least Im...	None	8	Right	Scale	Input
18	Advertising	Numeric	8	0	Advertising	{1, Least Im...	None	8	Right	Scale	Input
19	WOM	Numeric	8	0	Word of mouth...	{1, Least Im...	None	8	Right	Scale	Input
20	Ambience	Numeric	8	0	Ambience insid...	{1, Least Im...	None	8	Right	Scale	Input

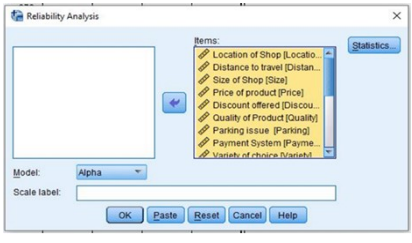
After variable entry, we have to enter the data in the form of responses given by respondents. Each row in data represents the responses given by one respondent for all 20 questions. Remember we must ask at least one question per variable in the questionnaire.

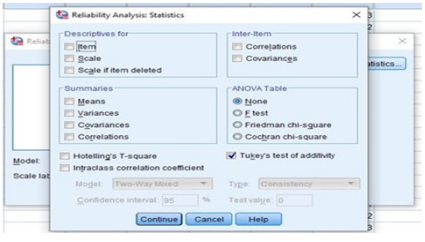


	Location	Distance	Size	Price	Discount	Quality	Parking	Payment	Variety	Communication	Display	Interior	Exchange	Safety	Social_stat	Exp
1	4	4	2	5	5	3	2	3	4	3	2	2	2	2	2	
2	3	2	5	4	3	5	2	2	3	3	4	2	2	3	3	
3	2	1	2	3	2	3	5	4	4	3	5	4	5	5	4	
4	2	2	6	1	2	1	1	2	1	5	1	.	2	5	4	
5	3	3	4	2	1	1	2	3	2	4	2	4	3	4	1	
6	4	3	3	3	1	3	2	4	3	5	3	.	4	5	2	
7	3	3	1	1	1	2	4	5	4	3	4	3	5	3	3	
8	2	1	6	2	2	1	1	4	5	4	5	4	4	4	4	
9	1	1	6	3	2	1	2	3	2	5	2	.	3	5	5	
10	3	3	3	4	2	1	2	3	3	3	3	3	3	3	2	
11	4	3	2	2	1	3	2	1	4	4	4	4	1	4	3	
12	3	2	5	4	1	2	2	1	4	5	4	.	1	5	4	
13	2	2	4	4	2	1	2	1	3	3	3	3	1	3	4	
14	2	3	2	2	2	1	3	1	3	4	3	4	1	4	3	
15	2	4	3	2	2	1	3	2	3	5	3	.	2	5	3	
16	3	2	3	2	2	1	3	2	4	3	4	3	2	3	3	
17	4	3	3	2	1	2	3	1	4	4	4	4	1	4	4	
18	1	1	6	1	2	1	3	2	4	5	4	.	2	5	4	
19	2	2	3	2	2	1	2	3	4	4	4	4	3	4	4	
20	2	2	2	2	2	1	2	4	4	5	4	.	4	5	4	
21	2	3	3	2	2	1	2	5	2	2	2	2	5	2	4	
22	3	2	5	1	1	3	2	2	2	4	2	4	2	4	2	

(Remember to check the normality and N/A values in the data before going for further analysis)

2. Test of Reliability: Now we test the reliability of data. Reliability has been defined earlier too. In simpler terms it means if two researchers collect data on two different occasions with the help of similar questionnaire, they both must get almost similar results. To test the reliability, we check the Cronbach's Alpha value. If the data is reliable the value of Cronbach's Alpha must be more than 0.80. When we run the reliability test, we can also run the Tukey's test of additivity. The *F* value of this test must be significant to confirm the reliability of data. To run this, we click 'Analyze' in SPSS, click on 'Scale' and then select the 'Reliability Analysis':





When we click 'Continue' in the second picture and then OK, we get the following output:

Reliability Statistics	
Cronbach's Alpha	N of Items
.853	20

ANOVA with Tukey's Test for Nonadditivity

			Sum of Squares	df	Mean Square	F	Sig
Between People			270.238	128	2.111		
Within People	Between Items		1655.697	19	87.142	92.281	.000
	Residual	Nonadditivity	40.316[a]	1	40.316	43.438	.000
		Balance	2256.237	2431	.928		
		Total	2296.553	2432	.944		
	Total		3952.250	2451	1.613		
Total			4222.488	2579	1.637		

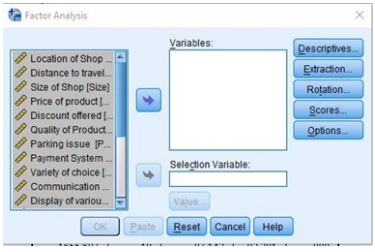
Grand Mean = 3.19

a. Tukey's estimate of power to which observations must be raised to achieve additivity = -.540.

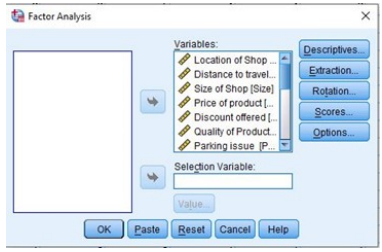
The Cronbach's alpha value is 0.853. Also, the *F* value of Tukey's test of non-additivity is 0.00 which is quite significant. It means the data is quite reliable and we can go for further analysis.

3. Factor Analysis: To reduce these 20 variables into some factors we now run Factor Analysis. For this, we click on 'Analyze', click 'Dimension Reduction' and then

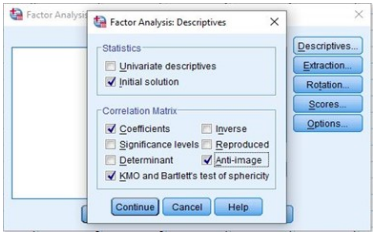
1. Select factor analysis.



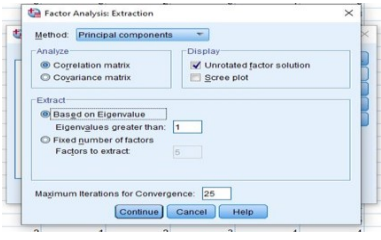
2. Select all variables and put into right column



3. Click on descriptive



4. Click on extraction



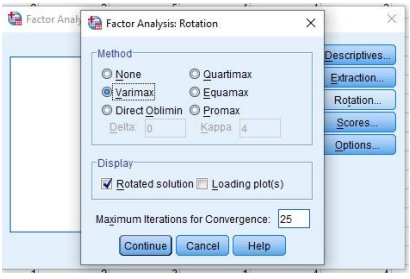
In the descriptive part, we check the initial solution, coefficient, anti-image and KMO and Bartlett's test of sphericity. The Kaiser-Meyer-Olkin Measure of sampling adequacy states that variance in the variables is caused by proposed factors. Its maximum value can be 1.0 which is very rare. Generally, KMO higher than 0.80 is considered as best. If its value is less than 0.50, then Factor Analysis is not possible.

In case of extraction – the method of factor analysis is selected. Here we have picked Principal Component method. This is the default selection of

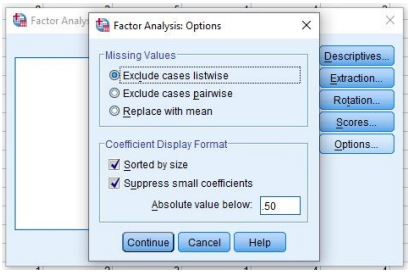
factor method. Also, the unrotated factor solution and Eigen value should be equal to 1. We click on 'Continue'.

After this we click 'Rotation' and select the Varimax Rotation Techniques. In this method the factors are rotated to better fit the data. We click on 'Continue' and then go to 'Options'. Here we suppress the variables whole coefficient value is less than 0.50

5. Click rotation



6. Click on options



Rotated Component Matrix[a]

	Component					Name of Factor
	1	2	3	4	5	
Advertising	.892					**Promotion**
Social Status	.872					
Word of mouth communication	.771					
Reference Group Experience	.748					
Interior Designing of Shops		.925				**Ambience**
Ambience inside shop		.909				
Size of Shop		.850				
Communication with sales person		.808				
Display of various items			.763			**Product**
Variety of choice			.737			
Price of product			.649			
Quality of Product			.529			
Exchange facility				.927		**Exchange**
Payment System				.924		
Discount offered				.687		
Location of Shop					.879	**Location**
Parking					.808	
Distance to travel					.679	
Safety Issue in trail room and Parking					.613	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.

Finally, we get the following output:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.790
Bartlett's Test of Sphericity	Approx. Chi-Square	2258.127
	df	190
	Sig.	.000

The KMO value is 0.790 which is sufficient to move with factor analysis.

We check the rotated component matrix. Out of 20 variables one has been omitted and the remaining 19 variables have been retailed and rotated into 5 different factors. We have to give appropriate name to the group of variables.

Such type of factor analysis is known as exploratory factor analysis where the new factors are explored from various variables. Another technique is known as Confirmatory Factor Analysis where we confirm the existing or already proposed factors again with the help of new data.

Report Writing

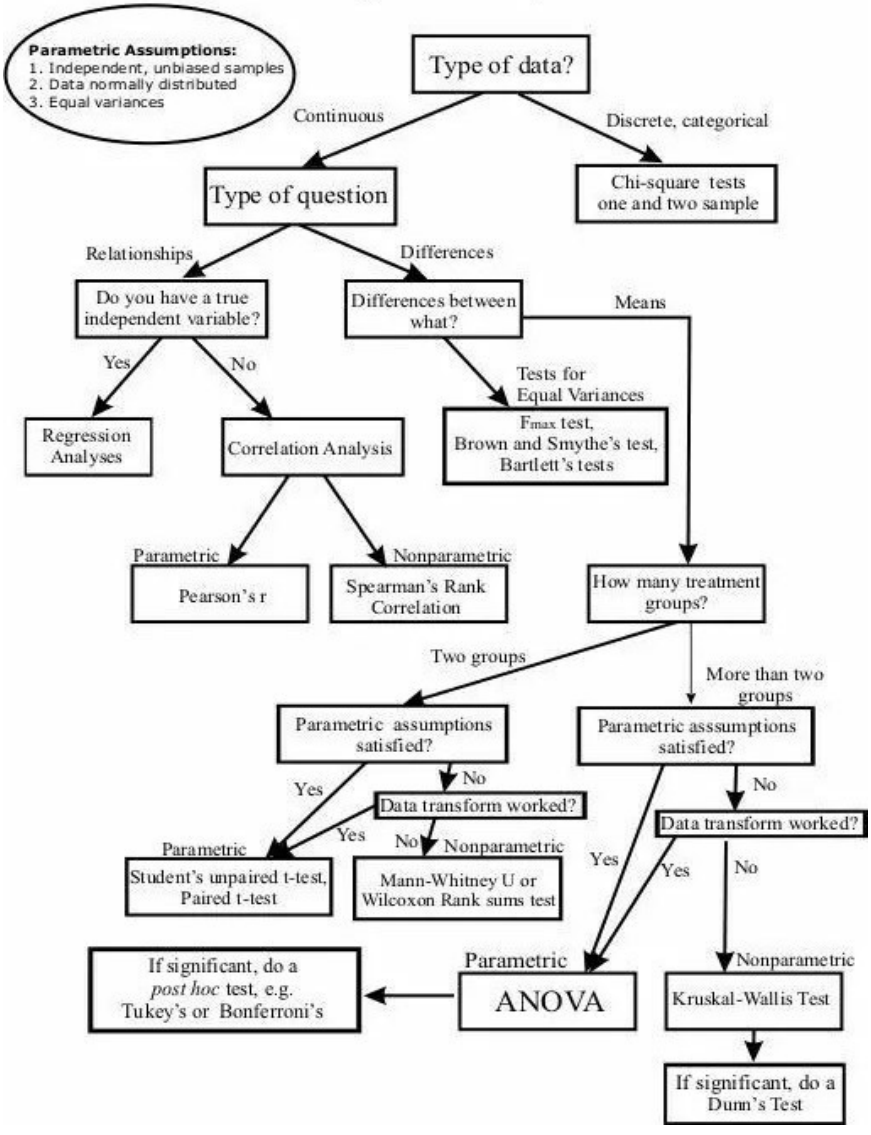
On completion of data analysis, we summarize the entire write up in the form of structured and organized report. The contents are arranged in well synchronized order. The contents may conform to the given order but the researchers can change them as per their preferences.

1. Title of Business Research Report
2. Abstract
3. Introduction
4. Literature Review
5. Research Gap
6. Research Objectives
7. Proposed Research Framework
8. Research Hypothesis
9. Research Methodology
10. Data analysis
11. Findings

12. Recommendations
13. Conclusion
14. References
15. Appendix

A researcher must be careful about the novelty and uniqueness of thoughts, concepts and ideas. There should not be any fake or wrong data in the report. A researcher should also be careful about plagiarism. If a researcher wants to include any work which is done by others, proper citation and credence must be in given in the report. Each diagram in the report must have proper reference. The language used must be professional as per the guidelines of that university, institute or journal.

Chapter 10
A Quick Guide to Identify the Use of A Statistical Test



Chapter 11
Terminology

Applied Research	Research conducted with specific objective like to develop a new product
Coding	The data is given unique code while entering the data into software to analyze
Dependent Variable	Variables which depends on independent variables
Dichotomous Question	Closed end question with two answers only
Extraneous Variable	Which affect the dependent variables indirectly
Focus Group	A qualitative study, where opinion of group of experts is solicited
Independent Variable	Variables which can be controlled
Primary Data	The data collected from the market by researcher himself
Secondary Data	The data which already existed in the market.
Variables	A mathematical unit where any value can be assigned

Chapter 12
Statistical Tables

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

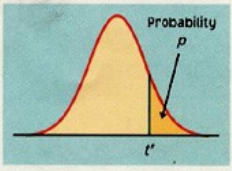
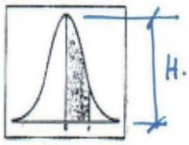


TABLE C t distribution critical values

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Table 5 Areas of a Standard Normal Distribution

The table entries represent the area under the standard normal curve from 0 to the specified value of z.



H →

z ↓	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

F - Distribution ($\alpha = 0.05$ in the Right Tail)

df_2 \ df_1	Numerator Degrees of Freedom								
Denominator Degrees of Freedom	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

Notes

Notes